Streamlining the "Producer/Archive" Interface: Mechanisms to Reduce Delays in Ingest and Release of Social Science Data.

Jinfang Niu
Margaret Hedstrom
University of Michigan

Abstract

Sharing research data has become an important concern around the world. Nevertheless, the situation with depositing and sharing social science data is far from ideal. In this paper, we identify some of the barriers to depositing social science data and discuss some mechanisms to improve the acquisition, processing, and release of data. We use deposit and processing records from one social science data archive for 184 data sets acquired over a six and one-half year period (December 1999 to April 2006) to analyze the time lag between the completion of a research project and release of data to the public. There are two types of delays before data and documentation are released and accessible to the public. The deposit delay is the elapsed time between the completion of a grant-funded project and the receipt of data by the data archive. The processing delay is the elapsed time between the archive's receipt of the data and the public release. On average, the total delay is three years between completion of a grant and release of data to the public.

In this paper, we identify causes for the deposit delay and the processing delay. We also propose several process improvements and incentive mechanisms that could be tested to reduce deposit and processing delays and expedite the release of social science data. These mechanisms may also be useful for data archives in other disciplines.

Introduction

Sharing research data has become an important concern around the world. In the United States, according to the Freedom of Information Act (FOIA), in response to requests for data relating to published research findings produced under a federal government award, the federal agency shall, within a reasonable time, obtain the requested data so that they can be made available to the public. Many funding agencies require their grantees to share data, such as the National Institute of Justice (NIJ), the National Institutes of Health (NIH), and the National Science Foundation (NSF) in the United States, and the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC) in the United Kingdom. A central principle of the Organization of Economic Cooperation and Development (OECD) is that publicly-funded research data should be openly available to the maximum extent possible. Publicly-funded research data are a public good produced in the public interest. As such they should remain in the public realm (Arzberger, et. al., 2004). Based on a web survey, data sharing is expected to become a major policy issue in the next few years in many European countries (Wouters, 2002).

To share data, data collectors could respond directly to data requests. The benefit of this is that investigators may

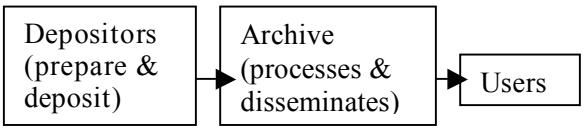
form collaborations with data requesters to pursue research of mutual interest. The downside for secondary users is that it may be difficult to locate and interpret relevant data. If there are many requests, the original investigator may find that disseminating data and responding to questions is very time-consuming. Grantees could also put data on their own websites, but this practice is problematic for long-term preservation.

Depositing data into a data archive is a better way of sharing and preserving data for the following reasons. Generally speaking, the data archive can preserve the data over a longer term than individuals who preserve the data on their own. Data producers collect data to address specific research questions. For the most part, their concerns about data management are short term and informal, whereas data archivists are concerned about long-term preservation (Zimmerman, 2003). Data archives have policies, procedures, and special expertise available to ensure that data are accessible in the future, such as saving multiple copies in multiple places and migrating the data to new technology environments. Data archivists help to compile and maintain associated documentation, and they manage, preserve and distribute the data. They often provide technical assistance, even training, in using the data sets. This save researchers' time and effort in sharing and preserving data and also lowers the cost to users in using data sets (NIH, 2003). Data archives collect many data sets, making it easier for researchers to access and use more than one data set or to select variables from many data sets.

Finally, the websites of data archives tend to be more visible than personal websites, and data archives already have many users. Dissemination through a data archive usually reaches more interested users. Studies have already shown that the placement of research papers in open access repositories can increase citation rates by anywhere from 50% to 500% (Hajjem, Harnard and Gingras, 2005). We should expect similar effects for the citation of data.

The model of data sharing through data archives is illustrated in Figure 1. Depositors prepare the deposit the data in a data archives and the data archive processes the data and disseminates it to users.

Fig. 1: Data Archiving Model



Good data sharing practices should satisfy the following conditions:

1. Data producers deposit their data in the appropriate data archive.
2. Data producers provide well-prepared data and documentation.
3. Data producers deposit in a timely manner.
4. The data archive processes and releases data for public use as soon as possible after receipt.
5. Users gain access to the deposited data shortly after the completion of a research project or the publication of an article based on the data.
6. Users receive sufficient documentation to interpret and use the data without consulting the

original data collector or the data archive.

What is the reality of data sharing practices through data archives? How is each of the conditions listed above approached? If there are problems, what measures might resolve them and improve data sharing practice? To explore those questions, we conducted a case study of one social science data archive. This data archive is operated under a contract with a federal funding agency. Grantees are required to deposit data in that data archive at some point after they complete their research projects. This paper reports findings related to above conditions 3, 4 and 5. It also proposes solutions to problems found and suggests areas where data archivists need additional knowledge.

Methodology

We analyzed the processing records of the data archive covering the period between December 1999 and April 2006. The processing records track each step in the ingest process from initial receipt of the data to public release. Complete records were available for 184 data sets over the six-and-a-half year period. From the processing records, we calculated the mean, median, minimum and maximum delay (in days) for deposit after the completion of a research project and for processing after receipt by the data archive. We interviewed several data archivists at the data archive to gain insights into the causes of the delays and to propose some solutions to reduce them. We used access records that track use of each data set to determine how frequently data sets in the archive are used. We also used selected results from a 2006 survey of data producers to validate our recommendations (Hedstrom & Niu, 2007).

Findings and Discussion

The deposit delay is the number of days between the date a grant was closed and the date that the data archive received the data associated with the project. Typically, the "closing date" is the date when the sponsor approves a project's final report. In our study, the average deposit delay was 767 days, with a median delay of 664 days, and a maximum delay of more than seven years after one project close out (See Table 1). It is also worth noting that, in spite of the deposit requirement, some grantees have not yet deposited their data.

Table 1: Deposit and Processing Delays (in days)

	Mean	Median	Min	Max
Deposit delay	767	664	-27[1]	2630
Process delay	355	276	20	1187
Total delay	1160	1122	263	2846

The processing delay is the number of days between the date when the data sets arrive at the archive and the date when the data is released to the public

[1] The delay is negative because the data was deposited before the grant was closed.

(release dates). The average processing delay was 355 days, with a median delay of 276 days and a maximum delay of a little more than three years (See Table 1). On average, the archive receives the data about two years after the close of a grant, and there is an additional year delay for processing the data. In total, secondary users have to wait more than three years before the data is shared with the public, but in one extreme case secondary users had to wait almost eight years.

We interviewed archivists at the data archive in order to identify factors that cause the deposit delay. One cause is the multiple steps and multiple hand-offs between different parties during the data submission procedure. For the 184 data sets under study, the data submission procedure was as follows: data producers submitted data to the funding agency; the funding agency then transfered the data to the data archive. Data sets may remain in the funding agency for a while before they are transferred to the data archive, and the transfer between the funding agency and the data archive also takes some time. A second reason is that there was no specific deadline for data deposit. In addition, even though the funding agency requires its grantees to deposit data, during the period covered by this study, the funding agency did not provide any incentives to encourage or expedite deposit of clean data with complete documentation. The funding agency closed out grants when they approved the final report, even if the researchers had not deposited their data. There were no punishments or rewards for depositing data, possibly leaving grantees the impression that depositing data was optional. Recognizing problems with the deposit process, the funding agency has changed the data submission procedure. Under the new procedure, grantees submit data directly to the data archive, and the archive signs off on the grant if the data and documentation submitted satisfy minimum criteria. This will eliminate the delay caused by the transfer between the funding agency and the data archive.

Another way to reduce the delay is to stipulate a clear timeline for deposit. The NIH, for example, expects data to be released and shared no later than the date an article based on the main findings from the final data set is accepted for publication (NIH, 2003). The ESRC requires data to be deposited "within three months of the end of an award (ESRC, 2000)." In the life sciences, the recommended timeline for transmitting data for sharing is 60 days after a paper is accepted by a journal (National Research Council, 2003).

We do not attribute all of the delay to the deposit procedures. In a 2006 survey of the same set of grantees, we found that the most common reason that researchers do not want to deposit data promptly is that they would like to publish more from the data before sharing it with others. Other reasons often cited were fear of compromising respondents' confidentiality, losing control over the data, losing rights of exclusive use, and the costs of preparing the data for deposit (Hedstrom & Niu, 2007).

We also interviewed the data archivists about causes of the one-year processing delay. Processing time is the total number of hours that data archivists

spend reviewing a data set for completeness, accuracy and interpretability, and compiling and editing documentation before the data is released to the public. From Table 2, we can see that the mean processing time is 79 hours (or about 10 eight-hour days). The most time spent processing a deposit was 359 hours (or about 45 of eight-hour days). The actually processing time accounts for only a very small portion (2.2 percent) of the one-year processing delay.

Table 2. Processing time (in hours)

	Mean	Median	Min	Max
Process time	79	60	8.5	359

There are several reasons for the processing delay. First, when data arrive at the data archive, data archivists check the completeness of data and documentation. If the deposit is not complete, archivists contact the data depositor to ask questions about the data or to request additional documentation. Waiting for responses from the data depositors can add significant amounts of time to the processing delay. Second, when the data archivists finish processing the deposit, they return all of the processed materials to data depositors for review prior to release. Normally they need to wait for at least two weeks for get feedback from the data depositors. Third, the data archive often waits between eight months and one year for the funding agency to provide the final grant report. The archive cannot release the data without the final report, even if they have received and fully processed the data. Finally, extremely large or complicated data sets require more time to process, causing a delay in processing other data sets in the queue. Under the new submission procedure, depositors will submit data and final report to the data archive. This will eliminate the waiting time for the final report coming from the funding agency.

To reduce or eliminate the waiting time for missing pieces of data and documentation, several things needs to be done. The data archive's guidelines include detailed instructions about how to document and deposit data, including a checklist of required items. If the grantees followed the guidelines the submissions would be complete. However, according to our survey data, almost half of the grantees were unaware of the data archive's guidelines, and almost one-third did not know about deposit guidelines issued by the funding agency (Hedstrom & Niu, 2007). We propose that the funding agency distribute the guidelines together with other award conditions to ensure the grantees receive them. The data archive should train grantees about the deposit requirements and submission process. To reduce the processing delay, we also propose additional research about the processing of data sets, and an exploration of using new technology to speed up the process.

In addition to the solutions proposed above, to reduce the delays and speed up data sharing, we should also consider the incentives for grantees. Data producers spend effort to prepare data for deposit, but most of the benefits of their efforts go to the secondary users. Some data depositors even worry that they or their reputations might be harmed if their data are used by unqualified researchers or are used in an inappropriate way. During

the period of our study, there was no punishment for failing to deposit data even though data deposit is a condition of the award. Likewise, there is no stipulated reward for depositing data, preparing good documentation, or for secondary use of the data. Some secondary users cite the data set or acknowledge the original data collector as a courtesy, but there is no policy requiring citation, acknowledgement, or co-authorship. As a result, many data depositors do not have strong incentives to deposit data or spend effort on data preparation.

We believe that either strong punishment for non-compliance or attractive rewards for compliance with the data deposit requirements would give data depositors incentives to be more cooperative in data preparation. Strong punishment, such as holding back a portion of grants until data are deposited, would force data collectors to prepare and deposit data, making all data collected with federal funding accessible to the public. On a cautionary note, however, strict punishment could have an adverse effect on the socially optimal expenditure of effort on preparing data for deposit and reuse, especially if researchers expend additional effort to prepare data that is not used heavily.

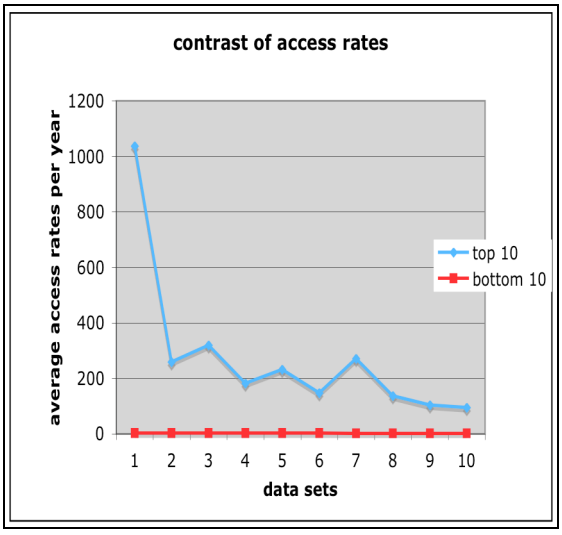
We analyzed the access logs for the data archive in order to compare the access rates for the most heavily used and least used data sets. Access to a study is counted whenever at least one file from a study is viewed on screen or downloaded. Monthly counts provide the number of unique users who have accessed the particular dataset. Reporting only unique users prevents over counting if a user accesses the same study multiple times during the month. Table 3 and Figure 2 show the access rates for the 10 most heavily used and the 10 least frequently used data sets.

Table 3: Cumulative and Average Monthly Access Rates for the 10 Most Frequently (Top 10) and 10 Least Frequently (Bottom 10) Accessed Data Sets.

Cumulative		Average (per month)	
Top 10	Bottom 10	Top 10	Bottom 10
5185	27	1037	3
2345	25	260	2.8
2234	24	319	2.7
1651	24	183	2.7
1623	23	232	2.6
1328	21	148	2.3
1267	20	271	2.2
1229	16	137	1.8
932	14	104	1.6
851	11	95	1.2

Enforcing uniform strong punishment on all data producers could result in a waste of resources because some data producers would spend effort on data

Figure 2: Comparison of Average Monthly Access Rates for the Top 10 and Bottom 10 Data Sets by Frequency of Access.



preparation to avoid the punishment even if his or her data sets are very unlikely to be used in the future. In addition, data preparation requires the sacrifice on the part of data collectors to benefit the society in general. This is not socially optimal when a researcher faces a more important and urgent research project that might produce more significant or higher quality data but still has to comply with the data deposit requirement.

Unlike the coercive and uniform nature of punishment, rewards induce rather than force researchers to expend more effort on data preparation. As an alternative to punishment, we recommend further exploration of several reward mechanisms to encourage compliance with data sharing policies. Requiring secondary users to include citations to data sets or acknowledgement of the data producer, similar to citations of published papers, would make the data producer's contributions explicit. Treating the citation of a data set similar to citations of published papers in performance evaluation of researchers would provide an incentive to depositors that is compatible with existing academic norms. Providing data depositors with feedback on use of their data or rewarding them when their data is reused would make data producers more aware of their contributions to a larger pubic good. We plan to test and evaluate the efficacy of these reward mechanisms in future work.

Implications for Data Curators

Data preservation practices are predicated on an assumption of some degree of cooperation between the data producer and the data archive. Data archives often expect data producers to expend some effort to prepare their data and documentation for deposit and reuse. Our research shows that data producers do not submit data promptly and do not deposit complete and accurate data and documentation even when deposit is a condition of grant funding. In the absence of punishment for non-compliance with data deposit requirements or rewards for the effort expended on data preparation, many data producers view data deposit as optional and expend as little effort as possible preparing their data for deposit.

In order to improve the quality of data and documentation deposited in an archive, data curators should make sure that data producers are aware of the deposit requirements and any related guidelines. Each archive should develop a combination of sanctions for non-compliance with data deposit requirements and rewards for high quality data and documentation tailored to the policy environment and norms of the data producer community.

Summary

In this paper, we analyzed the length of the delay between the end of research grants and the release of data for public use. We identified reasons of the delay in the communications between the archive and data producers, the submission process itself, and the incentive structures that would encourage data producers to expend more effort preparing their data for deposit. We propose process improvements and incentives to reduce the delay and streamline the ingest process.

Acknowledgment
This research as supported by the National Science Foundation (NSF Award Number IIS-0456022) as part of a larger project entitled "Incentives for data producers to created archive-ready data sets."

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., and Wouters, P. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal* (3:29): 135-52.

Economic and Social Research Council (2000). *Data Policy*. Available: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

Hajjem, C., Harnard, Y., and Gingras, Y. (2005). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *IEEE Data Engineering Bulletin* (25:4): 39-46.

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

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

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

Wouters, P. (2002) *Policies on Digital Research Data: an International Survey*. Amsterdam: Netherlands Institute for Scientific Information Services, NIWI-KNAW.

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.