Incentives for Data Producers to Create "Archive-Ready" Data: Implications for Archives and Records Management

MARGARET HEDSTROM and JINFANG NIU
School of Information, University of Michigan

Abstract: This paper presents results from a survey of researchers' behavior and attitudes about depositing data from sponsored research. The survey was conducted as part of a larger project called "Incentives for Data Producers to Create 'Archive-Ready' Data.[1] "Archive-ready data" are data that meet the minimum requirements for data quality, metadata, and documentation of the repository or archives where they will be preserved. We found that researchers often are unaware of data deposit requirements, but that they would be more willing to deposit "archive-ready" data if they knew it would be used and have a broader public benefit. Our research has identified barriers that data producers face when preparing data for deposit in an archives. We discuss broader implications of the findings for data producers and archivists in a variety of different settings and suggest ways the archivists and records managers might use incentives to encourage data producers to create "archive-ready" data.

Introduction

Digital repositories are predicated on an assumption of some degree of cooperation between the individuals and organizations that produce digital content and the archivists and repositories that preserve and provide access to that content. Typically, repositories set standards for data quality and metadata that must accompany submissions and expect data producers to comply with their requirements. Repositories may require producers to prepare digital content for submission, supply documentation, and assign some of their rights as content owners to the repository. Compliance with these requirements enables preparation of data and metadata for long-term preservation and facilitates dissemination and use. Preparing content for submission to a repository and providing the requisite metadata and content, however, requires time and effort on the producers' part. Research shows that even when requirements are kept to a minimum, data producers may be unwilling to invest the time and effort necessary to meet submission guidelines (Chan, 2007; Foster & Gibbons, 2005; Davis & Connolly, 2007; Hedstrom & Niu, 2008). Archival systems will break down if content providers are unwilling or unable to meet the requirements and expectations of repositories. It is important to understand the barriers that producers face when submitting data to a repository for long-term preservation and provide the right combination of incentives that will motivate producers to create and deposit 'archive-ready data.'

The precise criteria for 'archive-ready' data vary depending on the nature of the data being preserved, the capabilities and resources available to a repository to further process the data, and the ability of the user community to understand and interpret the data. We define data as 'archive-

[1] This research was supported by the National Science Foundation (NSF Award Number IIS-0456022).

ready' when the content and associated metadata submitted to a repository is sufficient to prepare the data for long-term preservation and dissemination without recourse to the original data producer. In the parlance of the Open Archival Information System (OAIS) Reference Model, 'archive-ready' means that the Submission Information Package (SIP) is accurate and complete enough for the archive to create Archival Information Packages (AIPs) and Dissemination Information Packages (DIPs) without further negotiations or discussions with the original data producer (Open Archival Information Systems Reference Model, January 2002). 'Archive-ready' is not equivalent to 'dissemination-ready' because archivists need to verify the completeness and accuracy of submissions, add descriptive and preservation metadata, and integrate new accessions into archival preservation and access systems.

Data Producer Survey

We have analyzed the behavior, barriers, and incentives of one group of data producers who are required to deposit their research data in the National Archive of Criminal Justice Data (NACJD). In many respects, our study represents a "best case" scenario for producer/archive cooperation. The producer community consists of a finite and known set of social science researchers who are funded by the National Institute of Justice (NIJ), which mandates deposit of data as a condition for receiving federal government funding. The Inter-university Consortium for Political and Social Research (ICPSR) operates the NACJD under contract with NIJ. Both NIJ and ICPSR provide guidelines, training, and technical assistance to help data producers comply with data archiving requirements. In this "ideal" situation, data producers should be aware of the data archiving requirement, know where they are supposed to deposit the data, and understand how to prepare the data, metadata, and additional documentation for submission. Guidelines and training even encourage data producers to incorporate data archiving requirements into their data collection and analysis plans in order to minimize extra effort for data preparation at the end of the project.

The NACJD and ICPSR have explicit requirements for submission of 'archive-ready' data that specify acceptable file formats, essential metadata, adequate documentation, and necessary measures to prevent inappropriate disclosure of personal identities and confidential information (Marz & Dunn, 2000; Inter-university Consortium for Political and Social Research, 2005). Compliance with these requirements enables staff at the NACJD to check the quality, accuracy and completeness of the data, transform the data as needed for long-term preservation, organize the documentation, and integrate the content with other archive holdings so that it is searchable and discoverable by users. The NIJ requires grantees to submit electronic data and supporting documentation, such as a codebook or dictionary, capable of being re-analyzed and used by other researchers by the end date of their grants (U.S. National Institute of Justice, 2005). Several indicators show, however, that even in this "best case" scenario, data producers do not consistently comply with the archiving requirement or invest enough effort to produce archive-ready data.

Lack of compliance with data deposit guidelines causes delays in processing and release of data. In a separate study, we analyzed NACJD processing records covering the period from December 1999 to 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 grant and for processing the data after its receipt by NACJD. We found that the average delay between

completion of a grant and deposit 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 some grantees have not yet deposited their data in spite of the deposit requirement (Niu & Hedstrom, 2007).

Table 1: Deposit and Processing Delays (in days)

	Mean	Median	Min	Max
Deposit delay	767	664	-27*	2630
Processing delay	355	276	20	1187
Total delay	1160	1122	263	2846

* The delay is negative because the data was deposited before the grant was closed.

Incomplete data and inadequate documentation can cause inordinate delays in processing data for long-term storage and dissemination because data archivists have to spend time and effort tracking down additional information, rectifying inconsistencies, correcting errors, transforming data into acceptable formats, and screening data for disclosure risks, often in consultation with the original data producer. For the NACJD data, 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). Much of the processing delay (the number of days between receipt of the submission and release of the data for dissemination by the archive) was the result of time that lapsed when the archivists requested additional information from the data producer or submitted revised data or documentation to the producer for his or her review and approval. The average amount of time spent by the data archivists actually processing the data was only 79 hours or about 10 working days (See Table 2).

Table 2. Processing time (in hours)

	Mean	**Median**	**Min**	**Max**
Processing time	79	60	8.5	359

We conducted a survey of 55 NIJ grantees to learn more about the obstacles they face when preparing data and documentation for deposit and to investigate possible incentives that might motivate producers to submit 'archive-ready' data. (Hedstrom & Niu, August 2008). Data producers have few incentives besides the NIJ archiving requirement to deposit their data, let alone expend extra effort to prepare their data for deposit, and several disincentives are in place. Attitudes toward depositing data, however, were generally positive (See Table 3)

Table 3: Data Producers' Attitudes Toward the NIJ Deposit Requirement (N=55)

Very Favorable	**Favorable**	**Neutral**	**Unfavorable**
34%	31%	24%	11%

In fact, 57% of respondents said that they would deposit their data even if it were not required by NIJ. Almost 95 % of respondents were aware of the data deposit requirement, but 30% of the respondents never received the NIJ Handbook on data preparation and 44% did not receive the ICPSR Guide. Only a minority of grantees received the NIJ Handbook (40%) and ICPSR Guide (30%) by the start of their projects so that they could incorporate data archiving concerns into their overall data collection and research plans.

The survey identified several factors that may discourage timely deposit of data by researchers. When asked which factors were detrimental to depositing their data, the most common response was a desire to publish more papers from the data before releasing them to a public archive (44%), followed by concerns over confidentiality (35%), loss of control over the data (31%), loss of exclusive use of the data (28%), and the costs of preparing the data for release (20%). Conversely, when asked about incentives, respondents indicated that they would be more inclined to deposit data if they thought the data would really benefit many other people (65.4%), if deposit were mandatory to receive new funding from NIJ (50%), if the data and documentation counted as a publication (36.5%), if citations to the data counted like citations to published papers (32.7%), if depositing data was a requirement for publishing a paper based on the data (26.9%), and if they could get monetary compensation for the data (15.4%).

Incentives

Our research suggests a number of mechanisms that might be explored to increase the quantity, quality, and rate of data submitted to this particular archive. Archivists have long contended that intervention early in the data creation process is an essential underpinning of the effective transfer and preservation of digital information. We found that even though grantees were aware of the data deposit requirement and both the funding agency and the repository offered guidance and training for life cycle management of digital data, only a minority of grantees received the guidelines early enough to incorporate them into their project design and management. If archivists expect data producers to include provisions for archiving their data as an integral part of the research process, then it is imperative that researchers get the information they need at the beginning of their research projects. Better tools, such as software that automatically captures metadata and documentation, and more training, would also facilitate transfer of data if these resources were made available early enough in project planning and the research process.

Even with better information dissemination and better tools, additional incentive mechanisms may be necessary to motivate data producers to document, deposit, and share their data. The literature suggests that people are motivated by intrinsic incentives, or extrinsic incentives, or both. *Extrinsic incentives* consist of material rewards such as payments and promotions; moral rewards such as praise, public commendation, and a good reputation; and coercive laws and policies, the violation of which would bring punishment. Intrinsic incentives motivate people to engage in some activity without an obvious external incentive present. *Intrinsic incentives* may include the enjoyment of participating in an activity, the satisfaction of contributing to a common good, or a sense of fulfilling a responsibility or obligation to a group or society at large (Bénabou, R. & Tirole,2003; Deci, 1975; Lepper, et al, 1996; Sansone, C. and Harackiewicz, 2000).

Our survey suggests that the right combination of extrinsic and intrinsic incentives could be used to encourage federally-funded social science researchers to expend more effort preparing data for deposit in an archive. The respondents were most favorable toward indirect rewards, such as counting data and documentation as publications and including references to these works in citation counts, with 37% and 33% respectively indicating that these factors would motivate them

to deposit data. More effective enforcement mechanisms with sanctions for failure to comply with data deposit requirements are potentially the most powerful extrinsic mechanisms. More than two-thirds of the respondents believed that their chances of being funded by NIJ in the future would be damaged if they did not deposit their data, even though NIJ does not currently consider previous compliance with data deposit requirements in funding decisions. Half of the respondents said that they would be more inclined to deposit their data if data deposit was mandatory for receiving new funding from NIJ. Only about 15% of respondents indicated that direct monetary compensation for depositing data would make them more likely to invest effort in preparing data and documentation for deposit.

Intrinsic incentives are more likely than any of the extrinsic incentives to motivate this group of social science researchers to do a better job preparing their data for deposit. The factor that the most respondents (65.4%) said would encourage them to deposit data was "if I thought the data that I deposited would really benefit many other people." Similarly, when asked what they considered as incentives for depositing their data in a public archive, 76% said because it saves other people the effort of collecting the same data again. Fewer respondents indicated that they were motivated by extrinsic rewards, such as increasing the chance that their data would be cited by others (52%) or saving their own effort in answering questions about the data (38%). These responses are consistent with the overall positive attitude of the respondents toward depositing data from grant-funded research in a public archive.

Although intrinsic rewards may be more powerful incentives for this group of data producers, designing incentive mechanisms that utilize intrinsic rewards is challenging because intrinsic rewards leverage social and behavioral norms such as satisfaction, enjoyment, group loyalty, and a sense of a larger public good. The respondents to our survey would be more inclined to deposit data if they thought it was a great benefit to others, but depositors are generally unaware whether and how their data are used. Of the 42 respondents who had deposited data in a public archive, 62% reported that they had no idea how often their data were used; 17% said the data were used rarely, 10% said they were used sometimes, and 10% reported that they were used heavily or often. Providing data producers with information about the general benefits of archived data and specific instances of reuse of their data is one incentive mechanism that would take advantage of intrinsic rewards. Making data deposit the norm among data producers would provide an additional incentive and reinforce the intrinsic reward of fulfilling a social obligation. Among our respondents, only 36% reported that "as a social scientist, I routinely deposit my data in a public archive." Additional incentives may be needed until such time that depositing data becomes a routine part of professional practice.

Finally, archivists should consider ways to balance the goals of timely and universal deposit of data with researchers' expressed concerns over release of data before they have fully exploited the data for their own research. There are distinct advantages for data archives to receiving data and documentation in a timely manner. Researchers who worked on a project are easier to track down. Problems with the data and documentation are easier to resolve when the work is still fresh in the researchers' minds and before data files, codebooks, survey instruments and the like go astray. Although the NIJ grantees are expected to deposit their data with the NACJD at the completion of their grant, we found that the average delay between grant completion and deposit of data is two years. From the researchers' perspective, however, immediate release of the data may be a disincentive deposit. When asked what they considered the best time frame for depositing data, 48% of respondents said when the grant is completed, 22% said when they no longer planned to use the data for additional analysis, 19% said when the results were published, and 11% said some other time frame. When asked to identify barriers to depositing data, 45.3%

of respondents said that they wanted to publish more from their data before depositing it. To balance the benefits of timely deposit with researchers' concern over premature release of data, data archives should consider offering data producers the option of an embargo on public release of data for a limited period of time after grant funding is exhausted.

Broader Implications for Archives

Considerable progress has been made in the last decade defining and developing architectures, software, processes, and standards for digital repositories. In addition to the OAIS reference model, repository management packages such as D-Space and Fedora are widely available, guidelines for storage and file formats are proliferating, and metadata and documentation standards are becoming mature. The Producer-Archive Interface Methodology Abstract Standard supplements the OAIS Reference Model with a methodology for structuring agreements and data flows between data producers and archives (Producer-Archive Interface Methodology Abstract Standard, May 2004). Nevertheless, our understanding of the relationships between producers and archives lags behind these technical developments for information created in a wide variety of formats and environments.

Archivists and records managers in government and institutional environments use numerous "incentive mechanisms" to change recordkeeping behavior and practice and encourage compliance with recordkeeping and archival requirements. Many of the mechanisms rely on extrinsic incentives, particularly laws, regulations, and policies that stipulate when and how to create, organize, store, destroy and transfer records. For electronic records, regulations and guidelines recommend or require use of specific storage media, file formats, data structures, and documentation standards. Archives and records management programs often provide additional training and technical assistance to help data producers satisfy these requirements. Ideally, information technology managers understand recordkeeping requirements and incorporate them into system design and acquisition decisions so that systems will generate 'archive-ready' records more or less automatically.

In spite of large investments in developing requirements and guidelines for 'archive-ready' data and records, studies of records management practices have shown consistent problems with compliance and enforcement. An assessment of recordkeeping practices in the federal government, conducted for the National Archives and Records Administration by SRA International, found that many significant records and most electronic records did not have disposition schedules, and as a consequence any permanent records were not being transferred to NARA (SRA International, 2001). The most recent biannual survey of records management by ARMA, AIIM, and Cohasset Associates found increased awareness of records management issues, but ongoing deficiencies in electronic records management, especially new forms and genres of electronic communications and long-term preservation (Williams & Ashley, 2007). Researchers at the University of Maryland asked explicitly about accountability and enforcement mechanisms for records management in federal agencies and discovered that majority of agencies investigated did not have accountability or enforcement mechanisms in place and only a few were considering compliance audits and rewards for good performance (Center for Information Policy, 2005).

Cooperation between producers and archives is an essential ingredient of long-term preservation. Archivists depend on producers to care for their records and to produce data for archival management in acceptable formats with sufficient metadata and documentation. If data producers are unwilling and unable to comply with submission requirements, digital repositories will not be

able to acquire, process, preserve, or disseminate data in an efficient and cost-effective manner. Yet archivists utilize a limited repertoire of incentive mechanisms to foster cooperation between records producers and archives. Most mechanisms use extrinsic incentives built almost exclusively on presumed compliance with regulations and policies. Assessment data on compliance consistently shows that regulations and guidelines are often ignored, sanctions for non-compliance are weak, and punishment is rare.

There are no simple answers or heuristics to follow when designing incentive mechanisms. Our research investigated a particular producer community of federally-funded social scientists working in an environment that imposes very specific deposit requirements and an archive dedicated to dissemination and preservation of criminal justice data. We found that the single most influential factor to encourage these researchers to deposit data in a public archive was their sense that other people would really benefit from access to their data. The researchers also indicated that more support from their sponsor for data management, better software tools, and some control on their part over the release of data would enhance cooperation. These and other insights about the data creation environment provide the basis for designing and testing alternative incentive mechanisms for these data producers.

There are many individual, organizational, and technological variables to explore in designing mechanisms that increase cooperation between data producers and archives. Archivists will not be able to investigate all of these variables for every data creation environment, but further attention to the obstacles that data producers face when creating 'archive-ready' data and a wider array of incentives could improve the dismal rate of compliance with many recordkeeping requirements. In designing incentive mechanisms, it is important to bear in mind that archiving data is at best a secondary concern for most producers. Unless archiving processes mesh seamlessly with the methods and tools that people use to do their work, satisfying recordkeeping requirements will require extra effort on their part with few, if any, direct benefits to them. Archivists should consider a wider variety of extrinsic incentives, such as including data quality, documentation or recordkeeping in performance evaluations, monetary rewards, and public commendations for exemplary contributions.

Our research suggests that intrinsic incentives are an especially fertile area for further investigation. This would open up new avenues for development and offer alternatives to coercive policies and punitive sanctions that have proved ineffective in the past. Appeals to larger social goals and the public good may resonate with the social and cultural values in government, academia, and the not-for-profit sector where incentives based on risk avoidance, efficiency, and personal gain have not taken hold. Moreover, new technologies are emerging, such social networking and community tagging, which leverage intrinsic incentives and facilitate individual contributions to larger communities of practice. Mechanisms based on intrinsic incentives are challenging to put into operation, but if these mechanisms catch on they tend to spread and work effectively without additional intervention because they appeal to individuals' sense of wellbeing and shape social norms. Although it is hard to imagine how to make it fun to create 'archive-ready' data, exposing the larger social benefits could create a powerful incentive for data producers.

References

Bénabou, R., and Tirole, J. (2003). *Intrinsic and Extrinsic Motivation*, Review of Economic Studies, 70: 489-520.

Center for Information Policy, College of Information Studies, University of Maryland. (December 19, 2005). *Best Practices in Electronic Records Management.*, http://www.archives.gov/records-mgmt/initiatives/umd-survey-main.pdf (accessed August 20, 2008).

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: the role of open-access institutional repositories. *Canadian Journal of Communication, 29*, 277-300.

Davis, P.M., and M.J.L. Connolly. (2007). Institutional Repositories: Evaluating the Reasons for Non-Use of Cornell University's Installation of DSpace. *D-Lib Magazine*: 13(3/4) http://www.dlib.org/dlib/march07/davis/03davis.html.

Deci, E. (1975). *Intrinsic Motivation*. New York: Plenum Press, 1975.

Foster, N.F., and S. Gibbons. (2005). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine, 11*(1). http://www.dlib.org/dlib/january05/foster/01foster.html.

Hedstrom, M., and Niu, J. (August 2008). "Incentives for Data Producers to Create 'Archive-Ready' Data: A Conceptual Model and Empirical Investigation" working paper.

Inter-university Consortium for Political and Social Research (ICPSR). (2005). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*. http://www.icpsr.umich.edu/access/dataprep.pdf (accessed August 20, 2008).

Lepper, M. R., Keavney, M., and Drake, M. (1996). *Intrinsic Motivation and Extrinsic Rewards: A Commentary on Cameron and Pierce's Meta-Analysis. Review of Educational Research*, 66: 5-32.

Marz, K., and Dunn, C.S. (August 2000). *Depositing Data With the Data Resources Program of the National Institute of Justice: A Handbook*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Niu, J., and Hedstrom, M. (April 2007). "Streamlining the "Producer/Archive" Interface: Mechanisms to Reduce Delays in Ingest and Release of Social Science Data," *DigCCurr* 07.

Open Archival Information Systems Reference Model (ISO 14721:2003) (January 2002). Available as Consultative Committee on Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book, http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed August 20, 2008).

Producer-Archive Interface Methodology Abstract Standard. (May 2004). Available as Consultative Committee on Space Data Systems (CCSDS). CCSDS 651.0-B-1, Blue Book, http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed August 20, 2008).

Sansone, C., and Harackiewicz, J.M., *Intrinsic and Extrinsic Motivation: The Search for Optimal Motivation and Performance*. San Diego, Calif.: Academic Press, 2000.

SRA International, Inc. (December 10, 2001). *Report on Current Recordkeeping Practices with the Federal Government.* http://www.archives.gov/records-mgmt/pdf/report-on-recordkeeping-practices.pdf. (accessed August 20, 2008).

U.S. National Institute of Justice (2005). *Data Resources Program 2005: Funding for the Analysis of Existing Data*. Washington, D.C.: National Institute of Justice. http://www.ncjrs.gov/pdffiles1/nij/sl000712.pdf (accessed August 20, 2008).

Williams, R.F., and Ashley, L.J. (2007). Cohasset Associates Inc., *2007 Electronic Records Management Survey – A Call for Collaboration*. Cohasset/ARMA/AIIM White Paper. http://www.cohasset.com/pdf/survey2007.pdf (accessed August 20, 2008).