The LEADS Database at ICPSR: Identifying Important "At Risk" Social Science Data[1]

Amy M. Pienta, Myron Gutmann, Lynette Hoelter, Jared Lyle

Inter-university Consortium for Political and Social Research, Institute for Social Research, University of Michigan

Darrell Donakowski

American National Election Studies, Institute for Social Research
University of Michigan

ICPSR has created a database to document information about the thousands of social science studies that have been conducted over the last 40 years. Included in the database are descriptions of social science data collections funded by the National Science Foundation and the National Institutes of Health. These records are supplemented with additional information gathered through correspondence with principal investigators of those awards with the goal of gathering information about the public availability of any research data collected with grant support. The goal of this paper is to describe the LEADS database and provide results regarding the scope of social science research data that are "at risk" of being lost. In the social science research community there have been longstanding expectations and mechanisms for archiving and sharing data. Even with this expectation, analysis of the LEADS database shows that the majority – nearly 75% -- of researcher-initiated social science research data is not archived publicly. Further, we find that a substantial minority have been lost.

[1] We would like to acknowledge the National Digital Information Infrastructure and Preservation Partnership program at the Library of Congress for supporting this work (NDIIPP Cooperative Agreement 8/04). We also thank Felicia LeClere, JoAnne O'Rourke, James McNally, Russell Hathaway, Kristine Witkowski, Kelly Zidar, Tannaz Sabet, Lisa Quist, and Robert Melendez for their contributions to the LEADS database project at ICPSR. The creation of the database was also supported by the following research projects at ICPSR: P01 HD045753, U24 HD048404, P30 AG004590. Email Address of Corresponding Author: apienta@umich.edu

Introduction

Despite the efforts of ICPSR and several other social science data archives in the United States (e.g. Odum Institute, Roper Center, Murray Archive), many social science studies do not reside in a permanent archive. Thus, the future availability of many legacy social science studies for secondary analysis is uncertain. Through the Data Preservation Alliance for the Social Sciences (Data-PASS), ICPSR set an ambitious goal of identifying the universe of quantitative social science data that has been collected with research grant support from the National Institutes of Health and the National Science Foundation. We also wanted to determine how much important social science data had been lost or is ”at risk” of being lost. ICPSR created the LEADS database to document information about the thousands of social science studies that have been conducted over the last 40 years. The database assumed its name because each of the records is a “lead” describing potential data for archiving at ICPSR. The goal of this paper is to describe the creation of the LEADS database and provide preliminary results regarding the scope of social science research data that are “at risk” of being lost.

The largest share of social science research is conducted with federal support. The National Science Foundation and the National Institutes of Health historically have supported a significant share of social science data collections and the trend continues today (Alpert, 1955; Alpert 1960; Kalberer, 1992). Thus, by focusing on gathering information from grant awards made by NSF and NIH it is possible to enumerate much of the social science data collections that exist today. Also, NSF and NIH keep electronic records about grant awardees that can be and have been culled into a single database useful for understanding the scope and breadth of social science research that has produced research data. We call this the LEADS database. The LEADS database contains information about research grant awards made by the NSF and the numerous institutes at the NIH

that fund social science research. The database documents, tracks, and identifies for possible archiving - original, social and behavioral data collections funded by NIH and NSF.

Background

Data sharing has been an important topic of debate in the social sciences for more than twenty years, initially spurred by a series of National Research Council Reports and more recently the publication of the National Institutes of Health Statement on Sharing Research Data in February 2003 (NIH 2003). Despite this formal written statement from NIH and a similar one from the National Science Foundation (NSF-SBE n.d.) that give official support for the long held expectations placed on grantees to share their research data, little is known about the extent to which data collected with support from NIH or NSF has been shared with other researchers. The limited work done suggests considerable variability in the extent to which researchers share and archive research data. LEADS will fill this gap in knowledge and create a research database for answering these questions.

NIH's policy is designed to encourage data sharing with the goal of advancing science. The benefits of sharing data have been widely discussed and understood by researchers for years. An important part of Kuhn's (1970) scientific paradigm is the replication and confirmation of results. Sharing data is at the core of direct replication (Anderson et al. 2005; Kuhn 1970; Freese 2006). The foundation of the scientific process is that research should build on previous work, where applicable, and data sharing makes this possible (Bailar 2003; Louis, Jones & Campbell 2002). The argument has been made, and there is some evidence to support it, that sharing data and allowing for replication makes one's work more likely to be taken seriously and cited more frequently (King

et al., 1995). In fact, Glenditsch, Petter, Metelits, and Strand (2003: 92) find that authors who make data from their articles available are cited twice as frequently as articles with "no data but otherwise equivalent credentials, including degree of formalization."

Additionally, the nature of large datasets virtually guarantees that a single researcher or group of researchers will not be able to use the dataset to its full potential for a single project. It may be the case that those who collect the data are not the best at analyzing them beyond basic descriptive analyses (Bailar 2003). Sharing data in this way ensures that resources spent on data collection are put to the best use possible and the public benefit is enhanced.

Finally, the use of secondary data is crucial in the education of undergraduate and graduate students (Fienberg, 1994; King, 2006). It is not feasible for students in a semester-long course to collect and analyze data on a large scale. Using datasets that have been archived and shared allows students to experience science firsthand. Instructors can use the metadata accompanying shared data to teach students about "good science" and the results obtained from even simple analyses to illustrate the use of evidence (data) in support of arguments (Sobal 1981).

In recent years, several national scientific organizations, such as the National Science Foundation and the National institutes of Health, have issued statements and policies underscoring the need for prompt archiving and sharing of data. These statements from leading agencies supporting research demonstrate that the data sharing ethic is an explicit part of our scientific norms and integral to maximizing the impact of research dollars.

Preserving Social Science Data

Data are currently shared in many different ways ranging from formal archives to informal self-dissemination. Data are often stored and disseminated through established data archives such as

the Inter-university Consortium for Political and Social Research, the Odum Institute for Research in Social Science, the Roper Center, The Henry A. Murray Research Archive or the custodial electronic records program of the U.S. National Archives and Records Administration (NARA). These data generally reach a larger part of the scientific community. Also, entries in formal archives typically include information (metadata) about the data collection process as well as any missing data imputations, weighting, and other data enhancements. These archiving institution have written polices and explicit practices to ensure long-term access to the digital assets that they hold that include replication copies stored off-site and a commitment to the migration of data storage formats.

A second tier of data archives have more narrowly focused collections around a particular substantive theme such as the Association of Religion Data Archives (www.thearda.com). The data in these kinds of thematic archives are not necessarily unique, though some of their holdings are, but the overlap between archives makes data available to broader audiences than might be captured by a single archive. ARDA, for instance, has a broader non-scientific audience who are interested in analysis and reports as well as the micro-data files for reanalysis. These archives expend resources on the usability of the collection for the present day and make some kind of a commitment to long-term access through migration and back-ups. Another tier of archives are designed solely to support the scientific notion of replication. Journal-based systems of sharing data have become popular in Economics and other fields as a way of encouraging replication of results (Anderson et al. 2005; Glenditsch et al. 2003). The longevity of these collections is sometimes more tenuous than the formal archives particularly if the sustainability of their archival model relies on a single funding source.

Some examples of less formal approaches include authors who acknowledge they will make their data available upon request or who distribute information or data through a website.

Researchers often keep these sites up to date with information about findings from the study, and publication lists in addition to data files and metadata. These sites are limited to those who know about the study by name or for whom the website has shown up in an internet search (see also Berns, Bond & Manning 1996). Typically, the commitment to preserving the content lasts only as long as the individual has resources available.

Thus, an underlying continuum of "risk of loss" has emerged with respect to the sharing of research data. Major archives and to a lesser extent smaller specialty archives have the most explicit commitment to preserving electronic social science data. Other data sharing solutions including journal-centered archives and various self-dissemination strategies carry considerable risk of loss given that the materials may not be refreshed or findable over time.

The Reluctance of Researchers to Archive Data

The time and effort required to produce data products that are useable by others in the scientific community is substantial. This extra effort is seen by many as a barrier to sharing data (Birnholz & Bietz 2003; Stanley & Stanley 1988). In addition to the actual data, information must be added to assist secondary users in identifying whether the data would be of value to them and in the analysis and interpretation of results. Such metadata includes complete descriptions of all stages of the data collection process (sampling, mode of data collection, refusal conversion techniques, etc.) as well as details about survey question wording, skip patterns and universe statements, and post-data processing. All of these factors allow subsequent researchers to judge the quality of the data they are receiving and whether it is adequate for their research agenda. Therefore, substantial effort is required of those sharing data, while the benefits accrue to the secondary user.

Another significant barrier in the sharing of data is the risk of breaching the confidentiality of respondents and the potential for the identification of respondents (Bailar 200);. The issue of protecting confidentiality has become more salient as studies collect information about social context, which may include census tract or block group identification to allow researchers to link the data collected with information about the context. Not only are data about social and community contexts being collected and included in datasets but also global positioning coordinates and information about multiple members of a household, all of which could make identification of any single individual easier. Additional information about biomakers and longitudinal follow up are also hallmarks of new data collection efforts. Both methodological innovations make it more difficult for Institutional Review Boards to allow for the wide redistribution of data.

Other reasons individuals give for withholding data include wanting to protect their or their students' ability to publish from the data as well as the extra effort involved in preparing data for sharing (Louis et al. 2002). Retaining the ability to publish from one's data seems to be a significant concern among scientists, both for fear of others "scooping" the story and that others will find mistakes in their attempt to replicate results (Anderson et al. 2005; Bailar 2003; Freese 2006; Bachrach & King 2004).

Current publication and academic promotion practices act as another barrier to sharing data – or, put another way, those who "hoard" their data are likely to be rewarded more than those who "share". There are often few, if any, rewards to sharing data, especially given the expense in terms of time and effort required to prepare clean, detailed data and metadata files. Researchers are not typically rewarded for such behavior, particularly if the time spent on data sharing tasks infringes on one's ability to prepare additional manuscripts for publication. Academic culture does not support the scientific norm of replication and sharing with tangible rewards. (Anderson et al. 2005;

Berns et al. 1996). As an example, in discussing the notion that researchers might share not only data but also analytic/statistical code, Freese (2006:11) notes that a typical reaction to a "more social replication policy would be to expend less effort writing code, articulating a surprisingly adamant aversion to having [one's] work contribute to others' research unless accompanied by clear and complete assurance in advance that they would be credited copiously for any such contribution." It is unlikely that attitudes about data sharing will change without strong leadership and examples set by senior scientists and the commitment of scientific institutions such as universities and professional societies who facilitate and enforce such sharing (Berns et al. 1996).

Policies about Data Sharing

Most institutes and organizations that finance research, especially data collection, have a policy about sharing data once the initial project is completed. The National Institutes of Health (NIH 2003) and National Science Foundation (NSF-SBE n.d.), for example, require a clearly detailed plan about data sharing as part of research proposals submitted for review. Plans must cover how and where materials will be stored; how access will be given to other researchers and any precautions that will be taken to protect confidentiality and when the data is made public. These requirements are not, however, evaluated in the review process nor are there formal penalties for non-compliance after the award. Most professional organizations also include a statement in their "best practice" or ethics guidelines that addresses the issue that research reports should be detailed enough to allow for replication and that researchers should also make available their data and assistance in these replication attempts when the requests are made (e.g., American Sociological Association, American Psychological Association, American Association for Public Opinion Research).

In addition to such general statements that data collected with public funds must be shared with other researchers and that individuals should be willing to assist others replicating their work, some fields, such as Economics, have taken steps to make the data sharing policy more concrete. In an attempt to allow for direct replications as well as full-study replications, the American Economic Review and other major economics journals have instituted the practice that any article to be published must be accompanied by the data, programs used to run the analyses, and clear, sufficient details about the procedures prior to publication (Freese 2006; Anderson et al. 2005). The requirement to include not only the data but also statistical code written to perform analyses requires that individual researchers thoroughly and carefully document decisions made during the analysis stages of the project and allows other researchers to more easily use these as starting points for their own work. This has led to an increased use and citation of work that has been published in journals where this type of information is required (Anderson et al. 2005; Glenditsch et al. 2003).

In summary, while the social sciences share in the normative expectation that research data must be shared to foster replication and reanalysis, there is little to suggest that it is a wide spread practice. Federal institutions and professional organizations underscore these normative expectations with implicit and explicit sharing policies. The advantages of sharing data with the research community are large and cumulative. Yet, with the exception of leading journals in Economics, there are few cases in which these normative statements are coupled with penalties or incentives to reinforce them. The institutional, financial, and career barriers to data sharing are substantial as noted. What remains an open empirical question is the extent of data sharing across social science disciplines. The LEADS database is an attempt to begin to understand the extent to which social science research data have been preserved, lost, or remain “at risk.”

To create the LEADS database, we have employed a systematic approach to identify the most significant studies of the past 75 years, many of which are at risk of loss. Studies identified for the LEADS database have come from two sources: research grant awards made by the National Science Foundation (NSF) and the National Institutes of Health (NIH).

Evaluation of NSF Grant Activity

Information about research projects that the NSF has funded since 1976 can be located by searching the online Award Abstracts database (http://www.nsf.gov/awardsearch/). The database includes abstracts describing the research and names of principal investigators and their institutions. Both completed and in-process research are included in the database. For awards prior to 1976, limited information is available from historic database records maintained by NSF at the same website location. Compared to the pre-1976 awards, more information about awards made between 1976 and 1988 is available, but coverage is still not as complete as for awards from 1989 to the present.

For possible inclusion in the LEADS database, 17,194 grant records (spanning 1976-2005) were downloaded from the National Science Foundation Web site using wildcard matching where the search terms (SOC*, POLIT*, and/or STAT*) appeared somewhere in the grant award record. These awards were made by 53 NSF programs and span the years 1976 to 2005. Next, we applied screening criteria to the records. To be considered for inclusion in the LEADS database, the grant record must describe research activity that is related to the social/behavioral sciences. Second, the grant must propose original/primary data collection or assembly of a new database from existing (archival) sources. Also, grant records referencing secondary data sources are coded for information pertaining to data that may or may not have not been archived.

Table 1 shows the distribution of activity proposed in the 17,194 awards after the initial screening. The largest number of awards were excluded from LEADS because they included no data (including training/workshop/conference activity, secondary analysis of existing data, and no data) or included data collection, but were not social/behavioral – this constituted over half of the reviewed records (56.8%, n=9,783). A substantial number of records data either had no abstract (14.8%) or were flagged by screeners as being ambiguous with respect to the screening criteria (13%). Over 2,500 awards were found to be related to the social/behavioral sciences and describing a primary data collection activity (n=2,537).

[Table 1 about here]

The historic, pre-1976 records from NSF included an award title and limited other information (PI, year), but no abstract. All historic records were downloaded from the NSF web site (n=96,403) and the limited information that was available was reviewed by the trained screeners. They excluded titles that were clearly not research grants (e.g. workshops and training) or referenced a topic that was clearly not social science. Thus, just over 1,000 historic awards were screened in because they had a social science title (n=1,102) and an additional 2,917 were screened as possibly being related to the social sciences. The pre-1976 records provide no information regarding the intention to produce/collect data.

Evaluation of NIH Grant Activity

For NIH awards, we mined NIH's CRISP database. CRISP (Computer Retrieval of Information on Scientific Projects) is a searchable database of federally funded biomedical research projects conducted at universities, hospitals, and other research institutions

(http://crisp.cit.nih.gov/). Current and past NIH awards made between 1972 and 2006 are currently accessible through CRISP. Users, including the public, can use the CRISP interface to search for scientific concepts, emerging trends and techniques, or identify specific projects and/or investigators. The database, maintained by the Office of Extramural Research at the National Institutes of Health, includes projects funded by the National Institutes of Health (NIH), Substance Abuse and Mental Health Services Administration (SAMHSA), Health Resources and Services Administration (HRSA), Food and Drug Administration (FDA), Centers for Disease Control and Prevention (CDCP), Agency for Health Care Research and Quality (AHRQ), and Office of Assistant Secretary of Health (OASH).

The LEADS review process of the NIH awards was similar to that used with the NSF awards, with one exception. NIH awards were screened and included in the LEADS database when they met the following criteria: social science (including behavioral) and original quantitative data. This strategy differs from the NSF award review in that strictly qualitative studies were not identified as such and excluded from LEADS. For the award years 1990-2001, all NIH institutes available from CRISP were downloaded and screened. For all other years, only the following institutes were reviewed: NICHD, NIA, NIMH, NINR, AHRQ, NIAAA, NIDA, Clinical Center, NIDCD, FIC, NCI, NHLBI, NIDDK. In all, 218,759 awards were screened and 7,626 selected as meeting the two review criteria. Table 2 includes a summary of NSF and NIH screening results combined.

[Table 2 about here]

Enhancing the Records in LEADS

Because we expect the list of at-risk data to be very long, standard selection criteria were developed to help ICPSR set priorities with respect to archiving data identified in the LEADS database. The need to gather information for selection has guided the activities ICPSR has undertaken following the initial screening process. One major selection criterion is the importance of a given study; we defined as important those studies whose data will advance knowledge, bearing in mind that it is not always evident in the present what data we might need in the future. The extent to which a dataset is assessed as being at-risk is another of the selection criteria. We also evaluate the risk of losing the content of each of these studies should acquisition and archiving not take place quickly. Based on a combination of these factors, we assign a priority ranking.

Using the information recorded in the LEADS database, ICPSR enhanced each grant record with information that can be used for selection using the following procedures: (1) generating updated contact information for the PI of the study; (2) determining whether the study is archived at ICPSR, Roper, Odum, or Murray; (3) asking principal investigators whether data have been produced, shared, or archived, and whether they are still available or accessible (PI-follow-up), (4) reviewing other awards obtained by the principal investigator, and (5)
collecting related citations using online citation searches

ICPSR selected for PI follow-up a set of awards that numbered a total of 10,905 awards. This follow-up set was selected based on a number of criteria: (1) the set includes all of the screened in records except for a subset of records that were being used for separate projects at ICPSR and (2) the set was expanded by including all of the NSF records where the screener was uncertain whether data were collected that might be in scope. Using a set of semi-structured questions, ICPSR collected information about archival status, availability of data, the format of the data files, and the storage media (e.g. punched cards, tapes or something more contemporary).

ICPSR also solicited from the PIs a description of the unique and/or special qualities of the study and resulting data set.

Email addresses could be found for 6,565 awards -- taken either from the grant record in LEADS itself (many of these were outdated or missing) or found through an internet search process. Before sending emails, ICPSR searched the shared Data-PASS catalog to verify if the data were in fact already archived at one of the partner archives (based on matching PI name and subject area). There were 215 awards where data were found archived at one of the partners already. The total possible sample size of responses we could have received was 5,848 – after any bounced emails were removed. Of those successfully contacted (defined as an email that did not fail upon sending), 2,548 responded to the email (as of October 31, 2007). Thus, the response rate for awards where a PI email address could be located is 43.6%.

Over 1,800 awards were confirmed by the PI (or another authoritative person who responded) or ICPSR (through the shared catalog search) as having produced research data (n=1,868). Among the awards with confirmed data (excluding 200 awards where the PI did not answer one or more questions), ICPSR found that only 334 awards that produced data had been archived (215 determined by ICPSR and 119 described by the PI and verified by ICPSR) or 20.0% were archived (see Figure 1).

[Figure 1 about here]

When data were not already archived, 49.3% of the awards (n=823) produced research data where the PI still had direct access to a copy of the data (see Figure 1). Thus, ICPSR believes that it has the potential to acquire over 800 data collections that are “at risk” of being lost. Through the Data-PASS operations group, we are approving or declining each unique data set for acquisition and

assigning a priority level. Nearly one quarter of the awards that produced data are already lost (23.9% in figure 1).

Obstacles to Archiving Data

Based on the responses of principal investigators, we compiled a qualitative set of reasons that older data have not been archived or in some instances cannot be archived. Some principal investigators attribute this to problems with data format or documentation. We learned that principal investigators have sometimes destroyed data stored on magnetic tapes because they mistakenly came to believe that there was no way to recover the data from those formats. Also, a lot of time may have elapsed since the study was funded and consequently data cannot be located or recollections about the structure and organization of the data and documentation have faded. Interestingly, several investigators noted that they were unable to secure the funding to archive the data. A common problem for foreign language surveys is that no English language translation exists of the survey or the documentation. Several studies lacked significance for public archiving because the source of the archival materials was available elsewhere, the series was not updated, or the study was itself not successful. However, the most common reason that data have not been publicly archived is that principal investigators have made the data available on a personal or departmental Web site and considered that as fulfilling any commitment to share the data.

Conclusions

The LEADS database contains valuable information about a wide range of social science research data collected with support from the National Science Foundation and the National

Institutes of Health. NSF and NIH awards typically lead to some of the largest investigator-initiated research activities in the U.S. and both institutions have had longstanding expectations that data collected with public money ought to be made available to the public and/or research community. In the social science research community, more so than in other basic disciplines, there have been longstanding avenues for archiving and sharing data through ICPSR and the other archives that make up Data-PASS. Even with this advantage, we find that the majority of social science data are not archived publicly. And, a substantial minority have been lost.

Through the Data-PASS project, ICPSR collected extensive information about the data most valuable to acquire through a resource intensive, but valuable process. We relied heavily on principal investigator cooperation because other possible research methods, such as reviewing citations and final reports to funding agencies, would have been especially time consuming and likely to leave large gaps in knowledge. Identifying investigators who were willing to provide information about the nature of their data sharing and archiving experiences proved to be invaluable – both in generating a list of data to archive and in understanding how much data is "at risk" or lost.

The LEADS database has several methodological limitations that we are working to address. First, we have not completely quantified the extent to which errors were made during the screening review process. Throughout the duration of the project, ICPSR has employed many staff and temporary employees to help review the NIH & NSF awards. In the process of screening the hundreds of thousands of records, a large number of awards could not be fully evaluated based on our inclusion criteria and many were missing an abstract altogether. Thus, we will pursue additional measures to research and include these awards. Also, the selected awards do not necessarily represent a mutually exclusive set of projects. Collaborative projects and continuation projects have not yet been eliminated from the records selected for LEADS, thus the number of selected records

will likely be smaller than the set of NSF and NIH awards combined that met our screening criteria. Finally, ICPSR has examined hundreds of thousands of awards in pursuit of creating the LEADS database. The size and scope of the LEADS database has been and continues to be one of the most challenging aspects of this effort.

Creating a database of important social science studies that have not been archived has been a longstanding interest of ICPSR. The LEADS database benefited from investments by several funded projects at ICPSR and has developed into a diverse and rich resource for identifying important social science studies that might be lost if they are not archived. ICPSR is beginning to realize the benefits of these investments as the “at risk” data are starting to come to ICPSR to be archived and disseminated. Through the DataPASS project, ICPSR is acquiring studies that meet the selection criteria of being important to the social sciences and are at risk of being lost because they are not in a permanent archiving situation.

References

Alpert, Harry. 1955. The Social Sciences and the National Science Foundation. *Proceedings of the American Philosophical Society,* 99(5), Conference on the History, Philosophy, and the Sociology of Science: 332-333.

Alpert, Harry. 1960. The Government's Growing Recognition of Social Science. *Annals of the American Academy of Political and Social Science,* 327, Perspectives on Government and Science: 55-67.

Anderson, Richard G., William H. Greene, B.D. McCullough and H.D. Vinod. 2005. The Role of Data and Program Code Archives in the Future of Economic Research. The Federal Bank of St. Louis Working Paper Series.

Bachrach, Christine. 1984. Contraceptive Practice Among American Women, 1973-1982. *Family Planning Perspectives* 16:253-259.

Bailar, John C., III. 2003. The Role of Data Access in Scientific Replication. Paper presented at Access to Research Data: Risks and Opportunities. Committee on National Statistics, National Academy of Sciences.

Berns, Kenneth I., Enriqueta C. Bond, and Frederick J. Manning (eds). 1996. *Resource Sharing in Biomedical Research.* Committee on Resource Sharing in Biomedical Research, Division of Health Sciences Policy, Institute of Medicine. Washington, D.C.: National Academy Press.

Fienberg, Stephen E. (1994). Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions. *Annual Review of Public Health* 15:1-18.

Freese, Jeremy. 2006. Replication Standards for Quantitative Social Science: Why Not Sociology? Unpublished manuscript, University of Wisconsin-Madison.

Glenditsch, Nils Petter, Claire Metelits, and Havard Strand. 2003. Posting Your Data: Will You be Scooped or Will You Be Famous? *International Studies Perspectives* 4(1):89-95.

Kalberer, Jr., John T., 1992. When Social Science Research Competes with Biomedical Research. *Medical Anthropology Quarterly*, New Series, 6(4):391-394.

King, Gary. 2006. Publication Publication. *Political Science & Politics*, 39(1):119-25.

King, Gary, Paul S. Herrnson, Kenneth J. Meier, M.J. Peterson, Walter J. Stone, Paul M. Sniderman, et al. 1995. Verification/Replication. *PS: Political Science and Politics* 28(3):443-499.

Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Louis, Karen Seashore, Lisa M. Jones, and Eric G. Campbell. 2002. Sharing in Science. *American Scientist* 90(4): 304-307.

National Institutes of Health (NIH). 2003. *Final Statement on Sharing Research Data.* February 26, 2003. Retrieved September 6, 2006 from http://grants.nih.gov/grants/policy/data_sharing/

National Science Foundation Directorate for Social, Behavioral, and Economic Sciences (NSF-SBE). (n.d.) *Data Archiving Policy*. Retrieved August 21, 2006 from www.nsf.gov/sbe/ses/common.

Robbin, Alice. 2001. The Loss of Personal Privacy and Its Consequences for Social Research. *Journal of Government Information* 28(5): 493-527.

Sobal, Jeff. 1981. Teaching with Secondary Data. *Teaching Sociology*. 8(2): 149-170.

Stanley, Barbara and Michael Stanley. 1988. Data Sharing: The Primary Researcher's Perspective. *Law and Human Behavior* 12(2): 173-180.

Table 1 – Screening Results of NSF Awards (1976-2005)

Type of Grant Activity Proposed	%	N=
Not Social Science or No Data	56.8	9,783
Social Science - Primary Data Collection	14.8	2,537
No Abstract	15.5	2,664
Flagged	13.0	2,232

Table 2: ICPSR's NIH/NSF Awards Screening Results

	# Records Reviewed	# Social Science Data Screened In
Recent NSF (1976+)	17,194	2,537
Historic NSF (Pre-1976)	96,403	1,102
NIH (1972+)	218,759	7,626
Total	332,356	11,265

Figure 1. Archival Status/Availability of Identified Data (n=1,668)