6

Establishing and Operating a Social Science Data Archive

JOSEFINA J. CARD
JAMES L. PETERSON

The Need

A recent National Research Council (NRC) report (Fienberg, Martin, & Straf, 1985) as well as a series of editorials in publications of the American Sociological Association (Baron, 1988; Card, 1989; Hauser, 1987) have lauded data sharing in the social sciences as a worthwhile goal, while pointing out that there are many obstacles in the way. In this chapter we describe how data sharing can be facilitated by data archives, and illustrate with an example from the social sciences. Although we develop the argument from a social science perspective, similar conclusions could well be drawn for other fields of research as diverse as biology and astrophysics.

The Virtues of Data Sharing

The benefits of sharing original research data and supporting documentation are widely acknowledged. Hauser (1987) lists the following: "reinforcement of open scientific inquiry; the verification, refutation, or refinement of original results; the promotion of new research through existing data; encouragement of more appropriate use of empirical data in policy formulation and evaluation; improvement of measurement and data collection methods; development of theoretical knowledge and knowledge of analytic techniques; encouragement of multiple perspectives; provision of resources for training in research; and protection against faulty data."

Economic as well as scientific factors contribute to the need for more data sharing. Social scientists have become increasingly aware of the value of large and representative samples, of longitudinal research, of replication of studies, and of obtaining a broad array of measures. All of these factors add to the cost of data collection. At the same time, the availability of public funds for research has been constricted and in some cases even eroded. The consequence has been a growing recognition that a small number of well-planned, fully funded, multipurpose surveys may well serve the needs of social science better than a plethora of small, narrowly focused studies. To serve these needs, however, large surveys need to be readily available to researchers for secondary analyses. The National Longitudinal Surveys (NLS) of labor market experience provide an excellent example of a data collection program that has been analyzed by several score of researchers and has spawned well over 200 articles and research reports (Bielby, Hawley, & Bills, 1978, list nearly 300 references in their review of the NLS over a decade ago). Analogous economic factors operate in other fields of science: Astronomers must share data produced by interplanetary space probes; climatologists draw on the common base of meteorological data. Additional, unexpected benefits also arise from data sharing. When the Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP), a pioneering social science archive funded by the Office of Population Affairs, was launched in 1982, attention focused almost exclusively on the format and documentation of the machine-readable data files. The original data collection instruments (typically questionnaires and interview schedules) were made available merely to complete the "supporting documentation." A surprising number of requests have been received for the instruments from researchers not interested in acquiring the accompanying machine-readable files. Easy access to original measurement instruments has proven very useful to investigators in the early stages of research planning. A side benefit for the field is that comparability of findings is enhanced when similar measures or instruments are used by different teams of investigators.

Similarly, when DAAPPP was set up, attention focused on its potential as a research resource; replications, refutations, secondary analyses, meta-analyses, and the like were envisioned. A surprising number of users have obtained the data for classroom use. The educational potential of the archive is only now beginning to be recognized; materials could and should be developed to teach statistics as well as substantive

AUTHORS' NOTE: Portions of this article first appeared in the January 1989 issue of the American Sociological Association's newsletter *Footnotes*.

concepts with real and state-of-the-art data, instead of hypothetical problems in printed textbooks.

The essential functions of a data archive, therefore, are to identify and preserve the best data sets in a field of inquiry, to promote the secondary use of these data sets through increased accessibility, and to do so more efficiently than the alternative informal system of data sharing between individual investigators. As important by-products a good archive may also promote standards of quality for data collection and documentation, improve the teaching of both substance and method, and broaden and strengthen the network of analysts working in the field.

Barriers to Data Sharing

The major deterrents to data sharing lie in the financial and time costs to researchers at both the supply and demand ends of the sharing process and in perceived threats to professional pride and reputation. Data donors incur costs in preparing their data for public distribution; in copying and transmitting the data to an archivist or to other recipients; and in providing technical assistance to users who later acquire the data. At the receiving end, data users incur search, access, and learning-to-use costs, all of which can be formidable if data bases have been inadequately documented and have not been cataloged or indexed in some way.

Beyond these dollar and time costs there are technical and human obstacles to resolve. The Baron (1988) editorial describes a few of these, chiefly: (a) *what to share*—the possibilities range from printed field notes to machine-readable raw data to machine-readable and/or printed correlation matrices; (b) *how to document*—many areas of inquiry have not yet developed institutionalized criteria for data collection, measurement scale construction, or data base construction; and (c) *whether to share at all*—to share is to give up "monopoly power" over a data base in which one has invested much time and energy, and is to submit one's work to the close and perhaps critical scrutiny of other researchers.

Meeting the Need: Centralized Data Archives

To facilitate data sharing on the supply side, both Hauser (1987) and Baron (1988) recommend that, at a minimum, incentives be provided

and additional funds made available to data donors. Clearly these are necessary. What is not necessary, however, is for researchers to incur *all* of the preparation, documentation, dissemination, and technical assistance costs required by data sharing. A possible alternative, described in this chapter, is to create a center that works with researchers to do much of this work for them. In the process of providing this service for a wide range of researchers, center staff acquire increasing expertise to do the job efficiently and well. Resulting benefits lie not only in lower overall cost to the research community, but also in the development of higher quality data sets and more standardized documentation. This, in turn, lowers the demand side costs of search, access, and learning for data consumers. The data are thus used more heavily, and by a wider constituency of users, amortizing the generally heavy data collection and recording costs more broadly.

Criteria for Evaluating a Centralized Data Archive

How might one determine whether a centralized data archive is successfully meeting the needs of data sharing? One approach would be to assess whether the archive's productivity—the number, quality, and usability of products; the service, technical assistance, and training provided to potential users; the benefits to science, practice, and/or policy-making that resulted from use of the archive—justified the costs incurred in setting up and operating the archive. One could ask the following evaluative questions:

- Does the archive contain the best existing data in the field?
- Are the data accompanied by documentation that is accurate, complete, and easy to use?
- Are there easy-to-search-and-use indexes that help the user decide which variables in which data sets are relevant to a problem at hand?
- Are the archive's dissemination procedures adequate: Do potential users know that the archive exists and what the archive's contents are?
- Is the cost of archive products perceived to be reasonable?
- Are the archive products purchased and used widely yet appropriately?
- Is the training and technical assistance offered by the archive helpful to users?
- What is the magnitude of cost savings (to donors and users) that has resulted from the archive's work?
- What research papers have resulted from analyses of data obtained from the archive?
- Has the archive been useful for teaching, practice, and/or policy-making?

We describe below one centralized data archive that has been in operation for seven years. The center's design and operational procedures address most of the issues raised by the NRC report and the ASA data sharing editorials. We proceed to assess this archive in terms of some of the above criteria and to highlight generalizable set-up procedures, operational methods, and potential benefits of centralized data archives.

An Illustrative Success Story

The Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP) was established in 1982 by the U.S. Office of Population Affairs. Its primary mandate was to assemble, process, and make publicly available those data best able to shed light on the problem of teenage pregnancy in the United States: the problem's incidence, prevalence, antecedents, and consequences, as well as preventive and ameliorative interventions. To date, data from 105 different studies, many of them longitudinal data bases, have been included in both mainframe and microcomputer formats. Data and documentation from the first 82 studies are now available on a single CD-ROM (compact disk, read-only memory)[1] for use on microcomputers.

DAAPPP addresses the obstacles to data sharing in the following ways:

(1) *What to Share:* A multidisciplinary advisory panel consisting of six outside scientists[2] selects which studies are included, using strictly objective criteria of technical merit, substantive utility, and policy or program relevance. Large, nationally representative data bases in the public domain, as well as smaller data bases collected by individual investigators, are included. The archive's focus is on machine-readable data. The original "raw" data and supporting documentation are acquired from the original data holder. When scale scores, indices, or other variables constructed from the raw data are included in the machine-readable file, the methods for deriving such variables are described in a printed user's guide, whenever possible. The sharing of raw data in machine-readable form allows public access to variables beyond those used in publication(s) by the original data holder.

(2) *How to Document:* Documentation is primarily in the form of machine-readable program statements created by archive staff for use with SPSS (formerly the Statistical Package for the Social Sciences) data analysis software. The program statements name and label each variable, specify the variable positions in the raw data file, and identify missing values. The program statements not only document the

file; when used with SPSS they create a system file capable of easy and powerful analysis. Users of other statistical analysis software programs can edit the program statements provided to suit the statistical program of their choice. A printed user's guide to the machine-readable files describes the purpose and contents of the data set, evaluates its quality and completeness, and alerts the user to idiosyncratic features discerned by archive staff while preparing the data set for public use. The archive's documentation is produced with input from the original investigator; the original investigator also reviews and approves the documentation prior to public release. Because of recent advances in optical character recognition software, which is able to scan a printed page and transform it into a machine-readable text file, it may soon be possible to provide machine-readable files in lieu of paper documentation for questionnaires, measurement instruments, study descriptions, and codebooks.

(3) *Minimizing Costs to Donors as Well as Users:* For the data donor, the costs of data sharing are limited to getting the data and documentation in a form understandable to experts on the archive staff, and then answering occasional questions that arise in the course of the archive staff's preparing the data base for public use. The archive receives the original data and documentation in whatever form the original investigator is most comfortable with. Data and documentation have come to DAAPPP in a very wide range of formats and "levels of finish." Data have been transmitted as ASCII, EBCDIC, dBASE, SPSS, and SAS files (SAS is another statistical program for social sciences); and have come on mainframe tapes, floppy diskettes, Bernoulli cartridges, and even punched cards. These data have been accompanied by codebooks in forms ranging from penciled notes to letter-perfect, machine-readable works of art, and have been accompanied by study descriptions in forms ranging from rough notes to a rich batch of publications. Researchers transmit this information *once* to a professional archivist who is very familiar with social science data. This done, the burden of processing, documenting, and disseminating the data file, and of providing assistance to users, shifts to the central source. News of the availability of archive data is disseminated by way of occasional publications, press releases, and conference presentations, plus a quarterly newsletter circulated free of charge to all who request to be put on the archive's mailing list. A limited number of user-training workshops are also offered (free of charge upon request) at universities, research institutes, government offices, and professional conferences around the country.

(4) *Providing Incentives:* Positive incentives (accompanied by friendly persuasion and patience) are used with potential data donors, and have been found sufficient. Upon selection of their data sets by the advisory panel, potential donors are informed—by formal letter and, whenever possible, by a telephone call from someone on the archive staff whom they know—of the "honor" of the selection, based strictly on criteria of scientific merit and utility to the field and to policymakers. As the archive's reputation has grown and its procedures become commonly known, and as an increasing number of government funders have begun to require the public release of data within two years of project termination, a growing number of researchers have been volunteering their data sets and requesting consideration for inclusion.

Requirements

The requirements for such knowledge sharing are quite reasonable. From funders, there are dollar requirements: $5,000 to $12,000 per study, depending on the size of the sample, the number of variables included, and the quality and completeness of the documentation received from the original investigators. This amount covers acquisition, processing, documentation, dissemination, and technical assistance. It is considerably less than the cost of the competing approach of asking each team of investigators to prepare its own data for public use, disseminate the data, and then provide technical assistance. Additional savings are gained by the fact that an independent body of scientists decides what is worth sharing; preparation and dissemination costs are thereby only allocated to those studies with sufficient scientific merit to deserve public distribution.

From data donors, good will and some data preparation time are required. DAAPPP has found a gratifying amount of the former. Though DAAPPP only compensates data donors for direct costs associated with copying and mailing their data and documentation to the archive, no one has refused to turn over requested data because of lack of time to prepare such, although the archive has had to be patient (waiting for investigators' downtime or for time to write "that one last paper") for a few data sets. From archive staff, a combination of substantive and computer expertise is required, along with a willingness to serve to facilitate colleagues' research efforts. In return archive staff get a broad overview of a field and a deeper insight into its underpinnings. Original data comprising a field are seen both in their glory and with their blemishes.

Issues, Problems, and Challenges

There are various practical issues and challenges involved in organizing and setting up a data archive, and in managing its day-to-day operations.

Defining the Scope of the Archive

It is useful and productive to define an archive's scope according to a topical focus. Being clear about the archive's substantive focus is the prerequisite for developing criteria for what data sets to include, and for developing a constituency of users. In time it may be possible for an archive to encompass several foci. Even so large and diverse an archive as that of the University of Michigan's Inter-university Consortium for Political and Social Research, however, maintains a clear, if broad, disciplinary focus. A substantive focus also helps with defining the target constituencies for archive products, and with development of dissemination, marketing, and training strategies aimed at reaching these constituencies. Organizing an archive around a given topic also helps with obtaining initial funding for the endeavor, as public and private funders generally have well-defined content areas of interest.

Obtaining Financial Support for an Archive

A variety of sources of support are available to establish and maintain an archive. Foundations or federal agencies with programmatic interests in certain topics are often willing to provide the funds needed to obtain and archive data sets within their fields of interest. Increasingly, granting agencies are requiring that grantees deposit their data sets in an archive at the conclusion of the grant period, and are providing funds in the grant to cover the costs.

Once data sets are archived, the costs of maintenance and dissemination can usually be covered by sales or subscriptions. A large archive, such as the University of Michigan's, may recover costs through subscriptions to member institutions: such members pay an annual fee, and can then obtain any data set in the archive at no additional cost. Other archives charge users on a per-data-set basis. These charges, typically a small fraction of the original cost of data collection and archiving, cover the expenses of producing copies of the data and documentation along with a small surcharge for maintenance. Commercial archives, which provide data to businesses, often charge considerably more as their acquisition and archiving costs are seldom covered by grants or contracts. Archives may also subsidize their archiving costs through the sale of ancillary services or products, such as customized data analyses, data books, or teaching products.

Maximizing Credibility

It is important that the archive's staff and operations have scientific credibility. Perceptions of an archive's scientific credibility could influence funders' willingness to provide support for acquisition and archiving, donors' willingness to forward their data to the archive, and users' predisposition to purchase and use data from the archive. It is often helpful to assemble an advisory panel of outside scientists who are experts in the content focus of the archive to provide guidance on which data sets to include in the collection and to provide intermittent review of archive products and procedures.

Standards of quality for both data and documentation are also essential. Indeed, one of the main functions of the archivist is to screen potential data sets and include only those that meet basic requirements for study design, sampling plan and response rate, instrumentation and data collection methods, data reduction, and documentation. One of the services an archivist provides is to bring the documentation of a data set up to minimum standards when this is necessary. By setting minimum standards in these areas, archivists may well promote the wider acceptance and adoption of such standards.

Product Design and Standardization

Much of an archive's utility will depend not only on the accuracy and completeness of its documentation, but also on the ease of use of such. The design of standard, user-friendly documentation is thus an important start-up function of a data archive. Of equal importance is the production of some form of annotated index to data sets that informs potential users of the contents of the archive and provides guidance in selection of appropriate data sets for particular analytic purposes.

Another important start-up task is the design and creation of "value-added" products and services to accompany the archive's data and documentation and facilitate their use. The DAAPPP archive has developed a particularly useful example of such value-added products and services in the form of software that permits search and retrieval of the archive's contents both at the level of the individual data set and at the level of the individual variable. Such software complements and extends the annotated index, allowing for an even more accurate and cost-effective method of selecting appropriate data sets.

Other examples of value-added products and services are: educational materials that reference archive data and/or require use of archive data; printed and/or machine-readable catalogues of archive products; on-line bulletin board informational service for users; training workshops for potential users; technical assistance for donors concerning how to prepare a data set for transmittal to the archive; technical assistance for users on appropriate use of archive data; presentations on how to use the archive at professional conferences attended by potential users; news briefs and journal articles describing the archive's contents and teaching users how to use the archive; and periodic newsletters describing recent additions to the collection.

Requisite Staff Capabilities

We have found it important that archive staff have both substantive as well as computer and statistical expertise. One cannot document a data set well unless one understands the nature of the study; the sampling, instrumentation, and data collection procedures used; which parts of which measurement instruments were administered to a particular subject; the meaning and implications of various kinds of missing values; how and when to use case weights, and so forth. All of these tasks require training, experience, and expertise in the substantive field of the archive.

Other archiving tasks require experience with computer hardware and software. The archivist must be capable of performing the following with machine-readable data and documentation files: edit the files; transform raw data files into system files, and/or system files into raw data files; develop user-friendly facilitators for using the data, such as statistical program control statements, or search and retrieval software to accompany the archive; perform statistical checks to assess the correspondence between the original and the archived data; and spot-check the internal completeness and consistency of the data. All these tasks require familiarity with computer hardware and with word processing, spreadsheet, and statistical software, as well as a basic knowledge of statistics.

Hiring and keeping a competent staff of archivists is a challenge to data archive managers. As just described, a relatively high level of diverse skills is required. Yet the day-to-day job activities are frequently repetitive and tedious. An archivist is not creating new data, but merely transforming existing data and documentation into a standard form usable by the general public. On the positive side, the periodic completion of a product of high quality—a well-documented data set in standard format—provides enormous satisfaction to competent archivists. It is important that data archive managers, as well as scientists at large, acknowledge the professional contribution of archivists by such means as including archivists' names in citations of the data base.

Professional Credit and Citations

Several individuals and institutions generally collaborate in making a data set available for public use. There is the sponsor or funding agency; the original investigators who collected the data, transformed such data to machine-readable format, and produced the original, privately held documentation; the archivists who augmented, cleaned, and clarified the original documentation, putting it into a standard format to make it understandable to the general community of interested researchers; and finally, the distributor or marketer of the data set. It is important that professional societies develop standard forms for citing data sets that acknowledge the unique contribution of each of these collaborating professionals and institutions. We have standard ways of citing publications that acknowledge editors, translators, and publishers in addition to the authors. We need to develop analogous standards for citing all those who contribute to the creation and distribution of a publicly available, machine-readable data set.

Facilities: Requisite Hardware and Software

A well-functioning archive will need to own, or at least have ready access to, the following computer-related equipment and software:

- a microcomputer for each archivist (a 286 or 386 with a hard disk drive of least 40 megabytes is recommended);
- access to either a mainframe computer for processing large data sets, or an in-house, networked workstation with high processing speed and a hard disk of at least 300 megabytes (if justified by the number of archivists and/or the size of the data sets being processed); and
- the requisite word processing, spreadsheet, and statistical software to process and document the data files.

The Future

There is likely to be a burgeoning of topically focused social science archives in the future. As the competition for scarce research dollars gets stiffer, and as the cost of designing and conducting national surveys increases, the cost-effectiveness of these data archives is enhanced. Already, the operators of DAAPPP are in the process of establishing data archives on the American family, on social research on aging, and on criminal justice statistics. These new archives will be using the same conventions, standards, and technology as DAAPPP.

Archives of the future will increasingly focus on microcomputer, as opposed to mainframe computer, media. State-of-the-art microcomputers capable of storing and processing vast amounts of data at very little cost are now within the reach of most academic departments and research institutions throughout the country. The advent of microcomputer CD-ROM technology, capable of storing 550 million bytes of information on a single disk only 4.72 inches in diameter, promises radical changes in how scientific research is conducted and taught.

We may envision a time when large cross-disciplinary archives develop through the merging of multiple special-purpose archives. Such archives might be patterned on the university or large metropolitan library. Although no single collection is likely to become close to encyclopedic, movement toward comprehensiveness will be facilitated to the extent that data storage and retrieval become inexpensive and standardized, developments that have already begun to take place.

Notes

[1] Nicknamed NATASHA, for National Archive on Sexuality, Health, and Adolescence.

[2] Wendy Baldwin (chair), Michael Donahue, Brent Miller, Kristin Moore, Alice Robbin, Freya Sonenstein. Past members: George Cvetkovich, Craig Peery, and Maris Vinovskis.

References

Baron, J. N. (1988, February). Guest editorial—Data sharing as a public good. *American Sociological Review, 53*(1), vi-viii.

Bielby, W. T., Hawley, C. B., & Bills, D. (1978). Research uses of the National Longitudinal Surveys. In *A research agenda for the National Longitudinal Surveys of Labor Market Experience* (part V). Washington, DC: Social Science Research Council.

Card, J. (1989, January). Facilitating data sharing. *ASA Footnotes, 17*(1), 8.

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

Hauser, R. M. (1987, December). Guest editorial—Sharing data: It's time for ASA journals to follow the folkways of a scientific sociology. *American Sociological Review, 52*(6).