Social Science Data Archives

CAROLYN L. GEDA

IN THE RECENT PAST, a variety of new facilities to support research and instruction in the social sciences and related areas of inquiry have emerged. Among the most important of these new organizations and facilities are those concerned with the collection, processing, documentation, preservation, and dissemination of computer-readable research data. Variously called data banks, laboratories, and libraries, these organizations are usually referred to by the common term *social science data archives.* Such organizations vary in the scope of their activities and in specific practices, but all serve the function of facilitating use of computer-readable empirical data in social scientific research and instruction. To an increasing degree, moreover, their resources are called upon also to assist in the processes of forming and evaluating public policies. Although social science data archives serve the rather specialized purposes and goals of the social sciences, they suggest the growing importance and value of computer-readable information, and they constitute a source of valuable experience and expertise for archivists confronted with a rising flood of computer-readable records.

Methodological and technological innovations were the primary factors leading to the development of social science data archives. With the development and subsequent refinement of the sample survey—or, less accurately, of public opinion polling—as a data collection technique, human behavior could be studied by using samples of populations rather than entire populations. As a consequence, researchers and practitioners were no longer limited to such materials as government censuses and reports.[1] At the same time, increased availability of electronic data processing equipment allowed the use of extensive bodies of data and complex statistical methods of data analysis that could not be widely or effectively employed when only human labor could be utilized.

Although survey research methods were used before the turn of this century, the real proliferation in the use of these methods occurred during the thirties and forties as the private and

[1] York Lucci and Stein Rokkan with Eric Meyerhoff, *A Library Center of Survey Research Data* (New York: School of Library Service, Columbia University, 1957).

public sectors employed survey research methods for the collection of data.[2] The United States government contributed substantially to the development of survey methodology with the use of random or probability sample survey techniques as early as 1937, for estimating national unemployment rates. These techniques were transferred to the Census Bureau in 1943 and subsequently became the Current Population Survey in 1947.[3] Concurrently, commercial polling and market research agencies, although employing more scientifically primitive techniques, were and continue to be an ongoing source of new survey material.[4] Finally, during this same period social scientists were refining interviewing and sampling techniques and the design of questionnaires. Scholars pioneering in this area were ultimately responsible for founding survey research institutes and centers, the earliest of which were the Bureau of Applied Social Research, Columbia University; the National Opinion Research Center, at the University of Denver (now located at the University of Chicago); the Survey Research Center, at the University of Michigan; and the Survey Research Center, at the University of California at Berkeley.

During the fifties and sixties, the impact of technology was felt in the form of the computer revolution. With the availability of electronic data processing equipment, *machine-readable* or punched card data could be analyzed more efficiently, accurately, and comprehensively without the large-scale investment of human labor that was formerly required. The results of these analytical efforts were available almost exclusively in published form; they included only a portion of the data findings and were usually presented as simple *marginal* distributions.[5] Social scientists were acutely aware of the limitations inherent in published materials and were experiencing mounting frustrations with the lack of access to the raw data.

Raw data from government and private sector sources provides an impressively rich research resource for social scientists and practitioners. This resource permits exploration of data well beyond the original research or collection focus, an activity commonly known as "secondary analysis."[6] Original collectors of data rarely extract all of the research value from the data, partly because of the focus that the primary interests of the data collectors take, and partly because of limited resources and time. Use of these materials for secondary analysis becomes increasingly critical with the rising costs incurred in conducting a well-designed sample survey. Researchers involved in the collection of sample survey data must have skills in survey techniques as well as access to necessary facilities such as those provided by research institutes which will draw the sample, ad-

[2] P. Young, *Scientific Social Survey and Research* (Englewood Cliffs, N.J.: Prentice-Hall, 1944).

[3] Paul R. Voss, "Population Data in Social Science Data Archives: The Survey Holdings ot the Roper Public Opinion Research Center." Paper presented at the annual meeting of the Population Association of America and the Association for Population Family Planning Libraries and Information Centers, Montreal, April 1976.

[4] Rensis Likert, "The Polls: Straw Votes or Scientific Instruments," *The American Psychologist* 3 (December, 1948): 556–57.

[5] Robert E. Mitchell, "Information Storage and Retrieval: Information Services," in *International Encyclopedia of the Social Sciences,* David Sills, editor (New York: McMillan, 1967), pp. 304–14.

[6] Herbert H. Hyman, *Secondary Analysis of Sample Surveys: Principles, Procedures and Potentialities* (New York: John Wiley & Sons, 1972).

minister the questionnaire through a field section employing trained interviewers, code the data from the completed questionnaire, and produce the data in machine-readable form. Obviously, the process of sample survey data collection requires substantial funds. The very high level of funding required for such data collection limits the number of studies which can be funded by national foundations and in turn the number of researchers fortunate enough to be engaged in this type of primary research. These considerations emphasize again the need to share data and have it readily available for secondary analysis.

Replication of original research through retesting hypotheses formulated by the original investigator is feasible only if the raw data is available. The availability of the raw data permits researchers to employ the statistical techniques used in the original research and to verify the resultant data findings. Students also can replicate original research in substantive courses when the raw data are available. In addition, a broad collection of independent sample surveys on the same population will provide further opportunities to retest existing hypotheses and to support or refute newly formulated ones.

These are but a few of the benefits that can be realized through the sharing of machine-readable data. Until the late fifties, however, there were few systematic attempts to preserve data and even fewer that came to fruition. In fact, punched cards produced by some public opinion polls and market research corporations were already being destroyed in the thirties and forties due to maintenance costs and/or the need to provide space for data being currently produced. The 1930 Census materials, approximately eight million punched cards, were destroyed on the reasoning that hard copy or printed form tabulations existed and the punched cards were no longer needed.[7] Several commercial research firms, however, were pursuing the possibility of establishing repositories for these data. The earliest of such efforts was that of the Elmo Roper Organization which led to the creation of the Roper Public Opinion Research Center in 1946, based at Williams College, to house the Roper surveys.

In the late fifties and early sixties, social scientists were primarily responsible for creating social science data archives. The merger of the broadening research interests related to quantitative data, and financial support from agencies such as the National Science Foundation, made their creation possible. These archives can then be viewed as service centers established by social scientists to respond to their research and scholarly needs.

Social science data archives are almost exclusively based within university or academic environments. The most common type of archives is a local one within a university department or computer center whose user community is predominantly that institution's faculty and students.[8] Other ar-

[7] Charles M. Dollar, "Computers, the National Archives, and Researchers," *Prologue* (Spring 1976): 29–34.

[8] For further discussion of types of archives, see Ralph L. Bisco, "Social Science Data Archives: A Review of Development," *The American Political Science Review* 60 (March 1966): 93–108; Ralph L. Bisco, editor, *Data Bases, Computers, and the Social Sciences* (New York: John Wiley & Sons, 1970), pp. 1–15; and Warren E. Miller, "The Development of Archives for Social Science Data," in Mattei Doggan and Stein Rokkan, editors, *Quantitative Ecological Analysis in the Social Sciences* (Cambridge, Mass.: MIT Press 1966), pp. 521–31.

chives, although still within the university structure, are based around primary research centers which are the major suppliers of their data. Examples of these organizations are the National Opinion Research Center, at the University of Chicago; the International Data Library and Reference Service of the Survey Research Center, at the University of California at Berkeley; and the Louis Harris Data Center, at the University of North Carolina. In Europe, several countries have an additional type called "national social science archives" which are funded by the governments and/or national social science research councils, such as the Danish Data Archive in Copenhagen; the Social Science Research Council Survey Archive at the University of Essex; the Steinmetz Archives within the Information and Documentation Centre for the Social Sciences in Amsterdam; and the Zentralarchiv für empirische Sozialforschung at the University of Cologne.

The vast majority of the archives are funded by their parent institution. Unlike the situation in the early sixties, little funding is currently available through external agencies such as the government or a foundation for general archival activities. Occasionally, funds may be secured to develop or accession data in a specific substantive area such as the Criminal Justice Archive and Information Network located with the Inter-university Consortium for Political and Social Research (ICPSR), which was funded by the Law Enforcement Assistance Administration (LEAA). Another pattern of funding, frequently tried but generally found to be inadequate, is charging for services. Such fees cannot support an archives, but are generally used for the reimbursement of direct costs for a given service. The final form, funding by membership, is used by only two archives in the United States, the Roper Public Opinion Research Center, connected with the University of Connecticut, Yale University, and Williams College; and the ICPSR at the Institute for Social Research, at the University of Michigan. Institutions pay an annual fee in return for a given set or level of services. This form of funding shifts from individual scholars to their institution the financial responsibility for accessing data and services.

Archival Appraisal and Accessioning

Although social science data archives differ in size, user community, type of data held, funding bases, and level of services, there is a set of activities in which all archives are engaged, albeit to different degrees. The most critical activities are data acquisition, processing, preservation, and access or dissemination.[9] Recommendations for the acquisition of specific data may come from the user community (students and researchers); the archives staff itself; from various governing or advisory bodies such as a council, board of directors, or acquisition advisory committee; or from individuals who wish to contribute or acquire data. Since foundation grants made for the purpose of data collection frequently contain a clause requiring that the data be placed in the public domain, recipients of such grants usually contact archives in compliance with this clause. Advisory committees, interdisciplinary

[9] Unlike a machine-readable division of traditional archives such as the National Archives which seeks to preserve the integrity of official records as received, social science data archives will frequently correct data or records upon validation and will almost always reformat the data for ease of access.

in nature as required, may be formed to identify existing bodies of data relevant to a particular subdiscipline or area of research.

Acquisition of data *files*—and, indeed, identification of collections for acquisition—is often far from straightforward. Information on existing data collections is fragmented across government agencies, research institutions, and university computing and data centers. To complicate further the location and identification process, the information in existence does not have a common *format*, does not necessarily contain the same elements of information, and may call the same set of data by different names.[10] Obviously, a union list of machine-readable files does not exist. This means that archives perform an important service by acting as an information center or clearing house with respect to the availability of data.

Some social science data archives, such as the Louis Harris Data Center and the Roper Center, have long-term arrangements with certain producers or suppliers of data and are automatic recipients of data collections from these producers or suppliers. As indicated above, however, most archives acquire data on the basis of user or clientele need and interest or other appraisal criteria. A major consideration for acquisition is the principal investigator's cooperation in making the data available to other researchers. Versatility of the data is assessed on the basis of such considerations as potential interdisciplinary use and research value for new interdisciplinary or disciplinary thrusts. New data may complement an existing group of data within the archives, or represent a new area that the archives is making a concerted attempt to augment. Other criteria reviewed are: whether the data represent a primary source for a particular research area; whether there have been major publications resulting from the data; whether the data will continue to hold research interest; and whether the data will be useful for instructional purposes. When the data were collected and the time period with which the collection is concerned may also be important criteria. Additional criteria may include whether the data concern a single point in time, several points in time (as in panel or longitudinal surveys), or are part of a continuing series. If the data represent a single point in time, there is the question of whether the study will be replicated. If it is a continuing series of data, there might also be the question of who is responsible for updating the material and at what cost.

Technical considerations also must be explored. The magnetic tape on which the data is written must be readable by the computer. The data should be in a technical *format* which can be readily redistributed, or convertible to such a format. Data tied to specific *software* systems generally will be more difficult to handle and disseminate than data not so tied.[11]

Adequate *documentation* for the data is required. The documentation must include a complete description of each variable. (This may be the question text if the documentation is for a sample survey.) Descriptions for each of the code values used to represent the

[10] Sue A. Dodd, "Cataloging Machine-Readable Data Files—A First Step?" *Drexel Library Quarterly* 13 (January 1977): 48–69.

[11] For further discussion of appraisal criteria see Charles M. Dollar, "Appraising Machine-Readable Records," *The American Archivist* 41 (October 1978): 423–30.

responses or numeric values are also required along with the location of each variable on the tape or punched card. These requirements constitute what might be called "minimum documentation." Data generally cannot be used without this information. Valid substantive use of the data frequently requires adequate information on the design of the data collection or data sources and collection techniques employed. Information on the structure of the data file, how missing data were handled, a list of publications or reports based on the data, and any other supplementary material that will make the data easier for a secondary analyst to use is desirable.

Every attempt should be made to protect the privacy and confidentiality of the respondents in surveys. Data collected from a national sample or mass population should have any identifying information removed before the data are made available to the public. In special samples, such as elite groups, the very nature of the data allows for disclosure of the identity of the respondent. Elimination of all possible offending variables would frequently make the data less valuable for research or instructional purposes and in some cases destroy its usefulness. One solution to this problem is restricted access to the data. All requests for analysis are then fulfilled by the archives staff. It is possible that an archives may not want to undertake such responsibilities, or that the principal investigator may not want to make the data available to an archives.

Once the decision to accession a study is made, the data are located through examination of catalogs of holdings or data inventories provided by archives and agencies; newsletters, journals, and directories; and direct contact with other archives and individuals. Acquiring the data may take a prolonged period of time, perhaps several years, and may not always be possible because of the lack of documentation for the data, or its prior destruction. With some data collections, principal investigators have invested significant amounts of personal time, effort, and resources in the process of coding and making the data machine-readable. Under these circumstances the principal investigators may be unwilling to deposit the data with an archives until their research is finished. Researchers may extend these proprietary rights over their data to include their graduate students working on dissertations or research projects. Therefore the data may not be available for accessioning into an archives until these latter projects are completed. This attitude is reinforced if the original researchers and students feel that open access to the data may result in being "scooped" by a secondary user of the data. It is also possible that the principal investigators intend to replicate the survey in the future or extend the data collection backward or forward in time, and consequently they view the data as increasing significantly in research value. Again, they may prefer to retain control over the data until they have had the benefit of analyzing it in the future as well as the present.

Control over the data can be a major factor if the principal investigators feel that the data may be misused, resulting in the publication of erroneous findings. Acute concern about misuse of data may arise if analysis could lead to the identification of respondents.

Some collectors of data are interested in distributing their data themselves. Direct contact with prospective users allows principal investigators to stay abreast of research areas being ex-

plored by secondary users and to offer consultation to these users as required. There may also be the expectation that the fees charged for the data will defray some of the original costs or may be used for future data collection. However, investigators generally are very willing to have their data accessioned after completing the primary publication since to distribute their own data can be very time consuming and usually involves disruption of other scholarly activities.

Computer-based Archival Activities

It is rather rare that data acquired by an archives are well documented and free of error. Therefore, upon receipt the data are tested for errors and inconsistencies. To perform these checks and subsequent corrections requires computational facilities for data management and organization. The checks may be minimal and simply verify that the code descriptions and locations of variables given in the documentation match the actual locations of the variables and code values on the tape. They may, however, be quite extensive, including a check of marginals against those reported in publications as well as a check for invalid, or "wild," codes for each variable. Inconsistencies within contingent or related variables are also checked. If new measures, indexes, or scales have been created from raw variables by the original investigator, these may be replicated and errors noted. Following the data checking process, the data are corrected by referring to the original sources of the data or contacting the original investigator for resolutions. If neither of these is possible, corrections will be made on the basis of the best available information or the errors will be noted in the documentation. Although the majority of the errors may not have represented problems for the original investigator, they may be obstacles for secondary analysts and students. The data may also be converted to consistent coding conventions. When the archives undertakes the responsibility of cleaning the data, it relieves future users from this costly and time consuming process and thus prevents duplication of effort.

After the data are checked and corrected, it may be necessary to reformat or reorganize them to a technical format that increases compatibility with different computational systems. Once this work is done, several back-up copies or duplicate copies of the data are made to secure the data against loss or destruction.

Concurrently, the documentation may be checked for comprehensiveness. Supplementary documentation will be produced by consulting the collectors of the data and publications based on it. Frequently, the documentation will be made machine-readable to allow it to be distributed with the data on a magnetic tape. *COM* fiche can also be produced quite inexpensively from this form of documentation and will act as another form of back-up or preservation of the documentation. Needless to say, extensive checking of the data and the creation of comprehensive documentation requires considerable resources both in terms of money and of skilled staff. Archives cannot perform these procedures on all data acquired and must, therefore, decide which data will be cleaned and to what level, given their available resources. Not all archives check data to the degree mentioned above. Some data are acquired, stored, and made available in their original form. Most archives have established a classification system for their data which alerts

potential users to the level or degree of checking that has occurred.

A side benefit of the documentation and cleaning process is the generation of standards for formatting, cleaning, coding, and documentation of data. The resultant product often can be used by original collectors as a guide for handling their own data.

In addition to the data acquisition, processing, and dissemination activities, data archives may be involved in consultation and training. Consultation and training activities may encompass *ad hoc* technical advice on the local computer and software facilities; advising students and researchers on data available for their substantive interests; sponsoring training programs and special seminars on the use of the facilities; and annual training programs such as have been established at the ICPSR through its Summer Training Program. The last program was originally established to train students and retool faculty in methodology and quantitative techniques.

Data archives may also have access to support facilities or staff such as computer programmers for the development of needed data processing and management programs. Additional computational programs may be required for faculty research or classroom instruction. These programs may be maintained by a data archives for general use and distribution along with data resources.

Future Implications

Anyone who uses, is affiliated with, or writes about data archives laments the lack of documentation and bibliographic control of machine-readable data files. This problem includes the lack of a union list of machine-readable files and inability to retrieve information at the item or variable level. As stated earlier, archives generally produce guides to, or abstracts of, their data, but there is no common procedure or format that is generally followed. The state of affairs is well summarized by John D. Byrum, Jr., and Judith S. Rowe:

> Data archives of all types are proliferating everywhere but data librarians are finding themselves floundering in their attempt to organize and document their data holdings. No rules exist; no generally accepted plan appears in the literature.[12]

This statement was made in 1972. Byrum and Rowe proposed four levels of documentation: standard catalog entries, data abstracts or descriptions, data documentation, and descriptions of the physical and logical characteristics of the machine-readable file. There is no doubt that if the recommendations had been followed, the level of documentation for machine-readable files would be improved. Somewhat preceding these recommendations, the American Library Association in 1970 formed a Subcommittee on Rules for Machine-Readable Data Files. This committee developed recommendations for a chapter in the *Anglo-American Cataloging Rules II* dealing specifically with machine-readable materials. Armed with these rules and the manual for cataloging machine-readable materials, also containing guidelines for formatting the information, systematic production of essential information can be undertaken. Only one

[12] John D. Byrum, Jr., and Judith S. Rowe, "An Integrated, User-Oriented System for the Documentation and Control of Machine-Readable Data Files," *Library Resources and Technical Services* 16 (Summer 1972): 338–46. The quotation is on page 338.

major obstacle remains: the necessary funding to perform the task. Archives are simply not sufficiently funded to handle all the activities. In spite of this, the cataloging of machine-readable materials has gained an impressive momentum among archives and, if it continues, we may well be in sight of a national union catalog for these materials.[13]

As these standards, and others referred to earlier, become more prominent and common, the materials received by archives should be more complete and require less staff time and money to prepare for distribution. However, countering this happy effect is the fact that more and more data are being generated. Also, archives no longer are focusing simply on survey materials for acquisition but are broadening their holdings to include almost all types of data being produced, particularly those being produced publicly. Add to this the rising costs to dc virtually anything, especially the paying of salaries, and it is evident that archives will find their decision-making tasks even more difficult. Computer technology, however, is producing systems that are more efficient, compact, and reliable even while they cost less. We have now entered the era of the mini- and micro-computers which indeed are so economical that they are becoming increasingly commonplace. All kinds of devices will have computers imbedded in them to assist normal functions: ovens, thermostats, typewriters, phones, and so forth. Use of personal computers for both office and domestic functions as well as use of computers in small businesses will spread. At a minimum, this will allow individuals to do at their own pace many of the things now done on larger computing systems. It will also increase the demand for shared or public data banks of data for recent surveys, census summaries, prices this week at local stores, educational materials—*ad infinitum.* Clearly, another phase of the growth of computing will result from the work that is being done to upgrade data communication, networks, and remote access systems. These changes are increasing the total demand and market base for computing, which in turn means that manufacturers of storage devices are motivated to spend the necessary money to develop more efficient equipment. It is possible, then, to look toward technology to decrease some archival costs. Ironically, this same technology is creating greater demand for data and indeed contributing to the production of vast quantities of data. During the next decade or so it is estimated that as much as 80 percent of the information processed by the federal government may be machine-readable.[14] We can anticipate that all of the familiar demands for well documented, formatted, and cleaned data will increase dramatically. If these demands are to be met, machine-readable data must be preserved systematically.

[13] A Conference on Cataloging and Information Services for Machine-Readable Data Files, funded by the National Science Foundation, was held 29–31 March 1978 at Airlie House, Warrenton, Virginia. Copies of the conference report are available from MRDF Cataloging Conference Secretariat, DUALabs, 1601 North Kent Street, Suite 900, Arlington, VA 22209.

[14] Charles M. Dollar, "Computers, the National Archives, and Researchers," p. 30.