6

William M. K. Trochim

Resources for Locating Public and Private Data

Samuel Johnson said, "Knowledge is of two kinds: we know a subject ourselves, or we know where we can find information upon it." This paper is concerned with the latter of these kinds of knowledge. Because data archives and individual researchers routinely store data, often in machine-readable form, the first task for the secondary analyst is to locate these resources. Extensive documentation on the existence of data sources is already available and the following enumeration should be valuable to secondary analysts. While one's awareness of the existence of data archives makes secondary analysis an attractive prospect, we will see that this attractiveness needs to be balanced against some of the difficulties inherent in the search and acquisition processes.

Data Archives: Issues and Sources

Many of the troublesome issues in using data archives have been discussed in the literature. Questions of accessibility, confidentiality, and the use of such data for secondary analysis have been raised by Boruch and Reis (1980), Hedrick, Boruch, and Ross (1978), and

Schoenfeldt (1970), among others. The federal government offers discussions of data archiving and access in work such as the Federal Statistical System Project (1978; see Chapter Two), and in the many publications of such groups as the Office of Federal Statistical Policy and Standards, the General Accounting Office, the National Technical Information Service, and the National Archives. Organizations of data users such as the Federal Statistics Users' Conference and the International Association for Social Science Information Service and Technology (IASSIST) also discuss these problems.

One of the major decisions facing the secondary analyst is how narrowly to define the search strategy. While overspecifying the topic in the early stages of a search may reduce wasted time, less restrictive specification allows one to discover unexpected but relevant data sources. Even when a researcher knows of a specific data base, a general search for related sources may yield separate indicators of the same phenomenon or data that can be used to supplement or check the validity of the major analyses. Cordray (1978) and Lipsey, Cordray, and Berger (in press) provide examples of how archival data can be used in this way. For example, if one is studying the effect of a compensatory education program within a local school system, one could search for data from other school systems in order to determine the importance of such factors as local idiosyncracies, historical forces, and developmental growth patterns. However. such broader specification of the data needed may predispose the researcher to alter the research question to fit the data that are available (Schoenfeldt, 1970). While researchers cannot always determine in advance how narrowly they should define their search, researchers must attempt to use the data archive rather than be used by it.

To use available archives, researchers need to understand the characteristics of various data bases. For example, some data archives contain actual data, while others simply have information that leads one to data sources. Bibliographical data files usually contain abstracts of publications or information about research project reports. If one suspects that needed data may be available from an individual researcher or report, a search of bibliographical data bases might be the quickest way to locate the data. Nonbibliographical data bases hold actual data, usually in machine-readable form. Archives also differ in the following characteristics: originator (government, research institute, private organization); duration of collection (continuing data collection, individual studies); restrictions on the accessibility of the data; provisions for preserving the confidentiality of respondents and the proprietary rights of the data collector; adequacy of documentation; and cost of use.

Given the wealth of data that are available from a wide variety of sources, no single search is likely to discover all the available data on a particular topic for secondary analysis. The following guidelines for a search should be tailored to fit the individual researcher's needs.

Specification of Needs. Often, the only way to discover where data are located is through various subject indexes that describe data archives and their holdings. For example, one might begin a search for data on electricity pricing by looking under the keywords *energy, electricity, prices,* and so on. During the search, the researcher may discover that the keywords *prices* and *energy* are too broad, while *utilities, peak-load pricing,* and others are more directly relevant. Through trial and error, one's search becomes more narrowly focused.

Initial Familiarization. Using the list of keywords, one begins to search the guides, catalogues, and lists of data archives and organizations that may hold appropriate data. While these sources may indicate the potential usefulness of a given archive, they will seldom supply information that is detailed enough to determine whether specific data are available.

Initial Contacts. Once one has identified likely sources of data, one needs to determine if the data file is truly appropriate for the specific analysis. At this stage, the most effective strategy is to contact individuals who are familiar with the archive in question. They should be able to provide information about specific data holdings, the form of storage, acquisition procedures, and the availability of other relevant data. Some files have restricted access or are available in specific formats. A government department or a private research firm may allow only their own employees to use the archive. Such restrictions can be readily determined by a simple phone call.

Secondary Contacts. Once it appears that an individual or archive holds appropriate data, one should seek additional information before attempting to acquire the data file. Helpful information includes: more detailed documentation of the desired data, considerations of the mode of transfer (punched card, magnetic tape, or computer listing), and the institutional capabilities. Magnetic tape files differ in the number of tracks and bits per inch appropriate for the tape. These and other matters are sometimes covered in catalogue description of archives, but should nevertheless be verified prior to making the acquisition request.

Accessibility Problems. Acquiring the data by no means ensures a successful analysis. Inevitable difficulties arise in setting up tapes, writing or using programs to access them, and so on. As with all phases of archive retrieval, it is useful to develop personal contacts with people who have experience using the data.

Analysis and Supplemental Analyses. As the analysis proceeds, one sometimes discovers that additional data are needed. Analysts' knowledge of the availability of related data and their personal contacts should improve their ability to gather additional data quickly and efficiently. For example, if the research question requires the use of an indicator of poverty, but the analysis reveals potential inadequacies in the index due to contamination, changes in reporting practices, and the like, the original list of archives or personal contacts may reveal other measures that could be substituted or used to validate the integrity of the initial measure.

General Sources and Guides

Probably the most difficult steps in the process we have described are the first two. Once specific data sources have been located, one is increasingly able to rely on the expertise of individuals who know the data. Because it is often difficult to initially locate sources of data, we devote the remainder of this chapter to a discussion of general guides and catalogues. While our listing is by no means completely comprehensive, it does provide sufficient information for a good general data search.

Two types of resources are generally available to one who is beginning a search for data. First, there are printed catalogues, guides, and directories of archives and data bases. These vary in quality and scope. However, knowledge of the major ones will greatly reduce search time. Second, a number of organizations and user groups are able to aid researchers who are trying to acquire data.

One of the most general, accessible catalogues of archives is the *Encyclopedia of Information Systems and Services* (Kruzas, 1978). It lists over two thousand archives and data bases, bibliographical and nonbibliographical, covering all types of archives including federal, independent, and foreign. For each archive, the *Encyclopedia* includes the address, phone number, name and title of director, number of staff, related organizations, description of system or service, scope or subject matter, input sources, holdings and storage media, publications, microform products and services, other services, clientele and availability, projected publications and services, remarks and addenda, and the name of a person to contact. The volume also contains eighteen indexes including an index of data base producers and publishers, a listing of data collection and analysis centers, a publications index, and geographical subject indexes. It can be used to locate individual researchers who may have relevant data (through bibliographical data bases), organizations that will perform data searches, and data archives themselves. Because it attempts to be comprehensive and is updated periodically, the *Encyclopedia* is probably a good starting point for a general data search.

Another general guide for locating data for secondary analysis is *Statistics Sources* (Wasserman and Paskar, 1977). This volume is a subject guide to data on industry, business, social conditions, education, finance, and other topics; it covers the United States and the international community. The volume is divided into two sections and primarily references data stored in printed form in government documents, periodicals, and trade journals. The first section is a selected bibliography of key statistical sources including guides, almanacs, U.S. governmental publications, census guides and publications, the statistical abstract of the U.S., yearbooks, and international sources. It describes many of the sources mentioned in this chapter as well as many more subject-specific guides and catalogues. The second and major portion of the book contains listings of specific sources by subject. For example, under the heading "France—unemployment" are two listings: the *Labour Force Statistics* of the Organization for Economic Cooperation and Development and the *Statistical Yearbook* of the Statistical Office of the United Nations. Similarly, the heading "Children—orphans" lists the *Social Security Bulletin* of the U.S. Social Security Administration as a data source.

Much of the data stored in archives are collected by nongovernmental research organizations, many of which retain a copy of the data as part of their own archival system. While few, if any, nongovernmental organizations have a formal policy of making data available to the public, they can direct the secondary analyst to appropriate data files or evaluation studies. Two general catalogues list nongovernmental research groups. The *Research Centers Directory* (Palmer, 1979) contains 6,268 listings of nonprofit and university-related research institutes, centers, foundations, laboratories, bureaus, and other research groups. The directory is organized into sixteen basic subject areas and includes an institutional index, an index of research centers, and a detailed subject index. The *Consultants and Consulting Organizations Directory* (Wasserman and McLean, 1976) offers information on over 5,000 for-profit firms and individuals engaged in consultation for business and industry. The entries are arranged alphabetically and the volume contains a subject key, index by state, and an enumeration of principal persons in each organization. Using the subject indexes of these two directories one could, for a given topic, locate nonprofit groups and for-profit firms that are or have been engaged in work related to the secondary analyst's area of interest.

Agencies within the federal government are major producers, either directly or through contracted research, of statistical data. Several guides describe data systems within the federal government. Although not intended as a reference tool, *A Framework for Planning U.S. Federal Statistics* (U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, 1978) provides a description of the federal statistical system and the data bases held by various agencies. It does not describe in great detail the specific data holdings of each agency but does explain the general categories of the agencies' holdings. In addition, it covers each of eighteen substantive areas (for example, criminal justice, education, energy, health, housing and community development, income maintenance and welfare), describes the agencies holding data on these topics, comments on the quality of the data, and recommends future data collection policy. The *Framework* is especially useful in limiting the number of federal agencies within which one needs to search. More specific information on the data that are held must be obtained directly from the agencies.

Four other sources, spanning the entire federal system, document some of the available resources—including evaluation studies, longitudinal surveys, and ongoing demographic and economic data collection efforts. The National Technical Information Service publishes the *Directory of Computerized Data Files and Related Software* (U.S. Department of Commerce, National Technical Information Service, 1978). This document describes machine-readable data bases and computer programs and indicates how they can be acquired. The *Directory* is divided into forty-six subject fields and includes a detailed subject and agency index. The National Archives publishes the *Catalog of Machine-Readable Records in the National Archives of the United States* (U.S. National Archives and Record Service, 1977), which describes ninety-nine files created by a wide variety of federal agencies. While neither the National Archives nor the National Technical Information Service comprehensively index federal holdings, they are both likely to become major depositories in the future. The General Accounting Office (GAO) issues *Federal Information Sources and Systems: A Directory for the Congress* (U.S. General Accounting Office, 1976), which covers automated information systems, catalogues, and listings issued by over sixty federal agencies. This directory emphasizes information for administrative rather than research use. It includes, for example, an on-line literature search system for medicine (MEDLARS and MEDLINE) and various departmental management information systems. Although most data banks containing purely statistical information apparently fall outside the information resources designated by the GAO, the directory is nonetheless a vehicle for primary and secondary analysis of administrative information. In addition, the Congressional Information Service publishes the *American Statistics Index* (Congressional Information Service, annual), which is intended as a master guide and index to all the statistical publications of the federal government. It consists of an index to and abstracts of any governmental reports that present statistics. Its major usefulness for secondary analysis is in providing access to tables and charts that can be analyzed or cited and in directing the analyst to persons or agencies that hold the data discussed in reports.

In addition to these federal guides, there are catalogues that describe the holdings of a particular agency or agencies. These include: *Selected Federal Computer-Based Information Systems* (Herner and Vellucci, 1972), *Standardized Microdata Tape Transcripts* (U.S. Department of Health, Education, and Welfare, National Center for Health Statistics, 1976), *Some Statistical Research Resources Available at the Social Security Administration* (U.S. Department of Health, Education, and Welfare, 1979), *Research Microdata Files* (U.S. Department of Health, Education, and Welfare, Social Security Administration, 1978), the *Directory of Federal Agency Education Tapes* (U.S. Department of Health, Education, and Welfare, National Center for Education Statistics, 1976), the *Directory of Automatic Data Processing Systems in the Public Health Service* (U.S. Department of Health, Education, and Welfare, Public Health Service, annual), and the *BLS Data Bank Files and Statistical Routines* (U.S. Department of Labor, periodical), among others. Most agencies involved in government archives issue some type of catalogue of their holdings. For example, the Bureau of the Census, by far the largest data collector, issues a wide variety of documents on holdings and suggestions for access.

Numerous federal agencies publish periodicals that discuss statistical systems, including the *Vital and Health Statistics Publications Series* of the National Center for Health Statistics, the *Review of Public Data Use* of the National Technical Information Service, the *Bureau of the Census Catalog* of the Bureau of the Census, and the *Statistical Reporter* published by the Office of Federal Statistical Policy and Standards. In addition, an external agency, the Federal Statistics Users' Conference, publishes a *Newsletter*. The problem, at least within the federal government, is not a lack of information, but rather, how to locate the appropriate sources. The best approach is to locate the appropriate agencies and then contact them or the regional office for further information on sources. A helpful document in this regard is the *Federal Statistical Directory* (U.S. Office of Management and Budget, 1974)—a telephone directory of persons engaged in statistical programs and related activities of the federal government.

Outside the federal system there are few sources (excepting the aforementioned *Encyclopedia of Information Systems and Services)* that provide comprehensive calatogues of data archives. Almost every archive issues its own catalogue and, once the researcher has identified likely sources, individual catalogues may be obtained. For example, the Data Library at the University of British Columbia published a basic catalogue in 1974 and updates it annually; the Interuniversity Consortium for Political and Social Research annually publishes a complete *Guide to Resources and Services.* Similarly, once appropriate profit and nonprofit research firms have been located through some of the sources listed earlier, their annual reports and bulletins can be acquired and checked for pertinent information. In addition, the Association of Public Data Users publishes the *Data File Directory* (Association of Public Data Users, 1977), which describes a large number of independent data sources and archives. However, these are available only to association members. Finally, the United Nations publishes a catalogue of its data holdings entitled the *Directory of International Statistics* (United Nations, 1975).

A number of organizations either will provide information about available archives or will conduct a data search. The National Technical Information Service has set up an Information Documentation Center with an independent firm, DUALabs, Inc. DUALabs will, for a fee, search the federal statistical system for data sources, usually within five days of the request. Because of the complexity of the decentralized federal statistics system, a search of this type by experienced people is often advisable. DUALabs issues a report for each search as well as giving names of persons to contact within the federal government.

Four major user groups are likely to have useful information for those seeking data. These are the Data Clearinghouse for the Social Sciences, in Canada, and in Europe, the International Federation of Data Organizations for the Social Sciences, the European Association of Scientific Information Dissemination Centers, and the International Association for Social Science Information Service and Technology. Each group publishes a bulletin and holds regular meetings. The addresses for the associations and user groups mentioned in this chapter are listed in the Appendix.

As the technology of data archives becomes more sophisticated, the field continues to change rapidly; archives combine into new networks, gather additional data, and revise their acquisition and transfer policies. Because a researcher alone seldom has the capacity to stay abreast of recent developments, use of the documents, catalogues, and guides described here provides a way for the individual to discover appropriate data sources with a minimum of effort.

APPENDIX

- Association of Public Data Users (APDU)
  Box 9287 Rosslyn Station
  Arlington, Va. 22209
- Data Clearinghouse for the Social Sciences
  151 Slater
  Ottawa, ON
  Canada KIP 5NI
- DUALabs, Inc.
  Information Documentation Center
  1601 N. Kent St.
  Suite 900
  Arlington, Va. 22209
- European Association of Scientific Information Dissemination Centers (EUSIDIC)
  P.O. Box 1776
  The Hague
  Netherlands
- Federal Statistics Users' Conference
  1030 Fifteenth St., N.W.
  Washington, D.C. 20005
- International Association for Social Science Information Service and Technology (IASSIST)
  Judith S. Rowe, United States Secretariat
  Princeton University Computer Center
  87 Prospect Avenue
  Princeton, N.J. 08540
- International Federation of Data Organizations for the Social Sciences (IFDO)
  Guido Martinotti, President
  Archivio Datie Programmi Per Le Scienze Sociali (ADPSS)
  Via G. Cantoni
  4 - Milano
  Italy

References

Association of Public Data Users. *Data File Directory.* Arlington, Va.: Association of Public Data Users, 1977.

Boruch, R. F., and Reis, J. "The Student, Evaluative Data, and Secondary Analysis." In L. Sechrest (Ed.), *New Directions in Program Evaluation: Training Program Evaluators*, no. 8. San Francisco: Jossey-Bass, 1980.

Congressional Information Service. *American Statistics Index*. Washington, D.C.: Congressional Information Service, (annual).

Cordray, D. S. "Making the Case for the Use of Patchwork Analyses in Quasi-Experimental Evaluation Research." Unpublished doctoral dissertation, Department of Psychology, Claremont Graduate School, 1978.

Federal Statistical System Project, President's Reorganization Project. *"Issues and Options."* Reproduced report, Office of Management and Budget, Washington, D.C., 1978.

Hedrick, T. E., Boruch, R. F., and Ross, J. "On Ensuring the Availability of Evaluative Data for Secondary Analysis." *Policy Sciences*, 1978, *9*, 259–280.

Herner, S., and Velluci, M. J. (Eds.). *Selected Federal Computer-Based Information Systems*. Washington, D.C.: Information Resources Press, 1972.

Kruzas, A. T. (Ed.). *Encyclopedia of Information Systems and Services*. (3rd ed.) Detroit: Gale Research, 1978.

Lipsey, M. W., Cordray, D. S., and Berger, D. E. "The Use of Multiple Lines of Evidence to Evaluate a Juvenile Diversion Program." *Evaluation Review*, in press.

Palmer, A. M. (Ed.). *Research Centers Directory*. (6th ed.) Detroit: Gale Research, 1979.

Schoenfeldt, L. F. "Data Archives as Resources for Research, Instruction, and Policy Planning." *American Psychologist*, 1970, *25*, (7), 609–616.

United Nations. *Directory of International Statistics*. Statistical Papers, Series M., No. 56. New York: United Nations Publications, 1975.

U.S. Department of Commerce, National Technical Information Service. *Directory of Computerized Data Files and Related Software*. Washington, D.C.: U.S. Department of Commerce, 1978.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. *A Framework for Planning U.S. Federal Statistics*. Washington, D.C.: U.S. Department of Commerce, 1978.

U.S. Department of Health, Education, and Welfare, National Center for Education Statistics. *Directory of Federal Agency Education Tapes*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1976.

U.S. Department of Health, Education, and Welfare, National Center for Health Statistics. *Standardized Microdata Tape Transcripts*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1976.

U.S. Department of Health, Education, and Welfare, Public Health Service. *Directory of Automatic Data Processing Systems in the Public Health Service*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, annual.

U.S. Department of Health, Education, and Welfare. *Some Statistical Research Resources Available at the Social Security Administration*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1979.

U.S. Department of Health, Education, and Welfare, Social Security Administration. *Research Microdata Files*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1978.

U.S. Department of Labor, Bureau of Labor Statistics. *BLS Data Bank Files and Statistical Routines*. Washington, D.C.: U.S. Department of Labor, periodical.

U.S. General Accounting Office. *Federal Information Sources and Systems: A Directory for the Congress*. Congressional Sourcebook Series. Washington, D.C.: U.S. General Accounting Office, 1976.

U.S. National Archives and Records Service. *Catalog of Machine-Readable Records in the National Archives of the United States*. Washington, D.C.: U.S. National Archives and Record Service, 1977.

U.S. Office of Management and Budget, Statistical Policy Division. *Federal Statistical Directory*. (24th ed.) Washington, D.C.: U.S. Office of Management and Budget, 1974.

Wasserman, P., and McLean, J. (Eds.). *Consultants and Consulting Organizations Directory*. (3rd ed.) Detroit: Gale Research, 1976.

Wasserman, P., and Paskar, J. (Eds.). *Statistics Sources*. (5th ed.) Detroit: Gale Research, 1977.