ACCESS TO SOCIAL SCIENCE QUANTITATIVE DATA

Sharon E. Henry
Executive Director
Data Clearing House for the Social Sciences
151 Slater
Ottawa, Ontario
K1P 5N1

ABSTRACT

Access to quantitative machine-readable social science data in Canada is discussed in this paper. Three major issues are addressed: 1) why access to these data is important 2) what are the major problems in accessing these data and 3) the role of the Data Clearing House for the Social Sciences in assisting users to obtain access.

RESUME

Le propos de cet exposé concerne l'accessibilité aux données quantitatives en sciences sociales au Canada. Trois problèmes principaux sont présentés: 1) pourquoi cette accessibilité est importante 2) quels sont les problèmes principaux reliés à l'accessibilité des données et 3) le rôle du Centre d'échange de données en sciences sociales auprès des utilisateurs de données.

INTRODUCTION

The decades following World War II saw such an explosion of information that the social scientist of today is faced with a surfeit of data (Nasitir, 1973). Paradoxically, unless the social scientist can obtain access to what data exists, a dearth of usable data results. In this paper, the following three issues concerning access to social science data in Canada will be discussed:

1. Importance of access
2. Major problems in obtaining access
3. Role of the Data Clearing House for the Social Sciences.

The term "quantitative data" will be used in this paper with reference to data in machine-readable form, that is, "a body of information coded by methods that require the use of a machine (typically but not always a computer) for processing" (American Library Association, 1976). In the social sciences, these data are usually the products of surveys, behavioral research, simulation, content analysis, recordings of transactional or administrative processes ("process-produced data") or official governmental inventories such as the census (Hanis and Mitchell, 1977). Analysis by the originator of the data, is defined as primary analysis. Secondary analysis which represents much of the analysis in the social sciences, is the result of examining data originating from another person or organization.

IMPORTANCE OF ACCESS

The sophisticated analyses of primary and secondary social science quantitative data are increasing in the academic, private, and government sectors of society. These data are being used for a variety of reasons such as to plan and justify government and corporate policies in response to social and environmental pressures as well as to further intellectual investigations by academic researchers.

The machine-readability of data facilitates their use in secondary analysis in a way that publication in traditional form does not. The same information may be organized, summarized, analysed, and criticized from a variety of perspectives with equal facility (Nasitir, 1973). The Canadian research community and its professional organizations are becoming increasingly aware of the importance of secondary analysis of data and of the researchers' responsibility to place information about these data as well as the data themselves in the public domain (Canada Council, Consultative Group on Survey Research, 1976).

Access to social science data previously collected, furthers

the comparison of methodology and the data themselves. It can lead to the examination of certain variables over a period of time, such as the way people vote(time-series). Multinational studies comparing the same or similar variables in different nations can be developed. The cross-disciplinary use of data is encouraged, e.g. data gathered for an environmental impact study can be of interest to sociologists in addition to environmentalists. Secondary analysis of data also avoids duplication of effort in study design, data collection and data encoding and input - all of which are expensive processes. Reduction of duplication of data collection can reduce the response burden on certain saturated survey populations.

Therefore, to permit effective utilization of scarce human and monetary resources, and to enhance the richness of analysis, it is important to have access to quantitative data in the social sciences. In the long run, the knowledge that one's data will be used for secondary analysis will be a motivator to improve the technical quality of the data.

MAJOR PROBLEMS IN OBTAINING ACCESS

There are four major problem areas in obtaining access to machine-readable quantitative data in the social sciences: 1) identification of data, 2) acquisition of data, 3) technical compatibility, and 4) data comparability. Each of these problem areas contains a number of issues.

These problem areas have been discussed at recent meetings of several international organizations: the International Federation of Data Organizations for the Social Sciences (IFDO), the International Association for Social Science Information Service and Technology (IASSIST, 1977), Unesco's Ad Hoc Committee on Social Science Information (Unesco, 1978), and the European Coordination Centre for Research and Documentation in the Social Sciences (International Conference on Information and Documentation in the Social Sciences, 1977). At the national level, the Canadian Consortium for Social Research (CCSR) and the Social Science Federation of Canada (SSFC) sponsored a conference in May, 1977 on "Problems in Social Science Data Creation, Archiving, and Management". One of the resolutions of this conference led to the formation by SSFC, of a Canadian Data Organization Committee (CANDOC) to discuss problems relating to quantitative social science data.

Identification of Data

The problem of identification of data involves the questions of inventorying, cataloguing and standardization. Inventorying means establishing and listing the social science data files that exist in Canada in the academic, private, and government sectors. The users of this inventory would be information specialists, researchers, and

administrators.

To satisfy the needs of these three major user groups, the cataloguing or recording of information about the data should contain general identification-type details as well as specialized details at the variable level. In addition, the cataloguing record should contain the information required by the traditional library standards such as the Anglo-American Cataloging Rules (AACR) and the Canadian MARC (MAchine-Readable Cataloguing) formats.

Maximum flexibility of the information would be facilitated if the details were themselves recorded in machine-readable form, supported by a system that would permit on-line searching and a variety of printed products. The principle of describing the data only once and distributing this information would eliminate duplication of effort and standardization in the identification of data (White, 1976). Perhaps even a Cataloguing-in-Production program for data files would be started (Dodd, 1977).

Acquisition of Data

There are several aspects to the problem of acquisition of the data once it has been established that they exist. The first aspect concerns privacy, an individual's right to know what data files exist pertaining to him or her, the researchers' right not to release data versus the researcher's obligation to do so, and the right of the individual to be guaranteed anonimity when data are released.

One must also face the problems of where to obtain the data (union catalogue aspect of including multiple listings) and the cost. To solve these problems, is it necessary to have the data stored centrally or with the present technology, is it sufficient even across the great distance of Canada to transmit large files on-line?

Technical Compatibility

The problem of technical compatibility involves such questions as how the data are stored (e.g. on tape, disk, or cards), have they been cleaned (e.g. accuracy validated), what software is needed to use the data (e.g. SPSS), are the data in a non-standard format, are the data well documented (e.g. codebook, list of variables) from a technical point of view, is a special computer configuration needed, and are the data archived somewhere to ensure their continued availability?

Data Comparability

The test of the usefulness of any system established to solve the problems previously listed, is whether or not the data can be

compared or re-analysed by another researcher. This presumes that the data documentation is sufficient to permit understanding by another researcher and that similar methods of sample selection, data collection, and data coding (e.g. standardization of social indicators) are used. The need for work in this area to further the comparability of data has been identified by the Canada Council Consultative Group on Survey Research and, more recently, by the CANDOC committee.

ROLE OF THE DATA CLEARING HOUSE FOR THE SOCIAL SCIENCES

The Data Clearing House for the Social Sciences was established to assist in the solution of a number of the problems listed above. It is an independant, private, not-for-profit organization founded in 1973 by the Association of Universities and Colleges of Canada (AUCC) and the Social Science Research Council of Canada (SSRCC), now the Social Science Federation of Canada. The Data Clearing House functions as a central comprehensive source from which to obtain information about social sciences quantitative data holdings found in the private, government and academic sectors of Canada.

The four main objectives of the Data Clearing House are:

- The preparation of an index of quantitative social data holdings that exist in machine-readable form and are to be found in Canadian universities, as well as in non-profit research agencies and other bodies conducting social science research.

- The collection from federal and provincial government departments of a continuing description of their holdings and the performance of a liaison role between individual scholars and government departments.

- The provision of information in response to individual inquiries, referring the inquirer to the source but not attempting to provide the inquirer with the actual data.

- The provision of technical information necessary for the more effective use of the data.

To accomplish its objectives, the Data Clearing House developed CANISISS, a computer-assisted data base containing descriptions of social science data holdings and centres. The data base was developed using the National Inventory System of the National Museums of Canada which is based on a modified and enhanced version of ISIS (Integrated Set of Information Systems).

Using the concept of the unit record, the Data Clearing House

offers the following bilingual services from CANISISS (Henry, 1977): annual Social Science Data Inventory, annual Directory of Social Science Data Centres, on-line referral service and current awareness service. Activities of the Data Clearing House and other research organizations in the field, are highlighted in the Bulletin.

Specialized services are provided on request. Examples of these services include the compilation, cataloguing, and production of special subject-oriented or institutional catalogues of data files and the provision of specialized reference services. One of the special projects being undertaken by the Data Clearing House is the production of a Directory of Federal Data Sources for Statistics Canada. The Data Clearing House is also working on phase I of the development of a clearing house of statistical information in Statistics Canada publications.

The products of the Data Clearing House contribute significantly to solving the problems of identification of data. CANISISS and its products contain general-purpose information about the data files as well as a summary and index terms. Information at the variable level is not always provided although a copy of the data documentation is often retained at the Data Clearing House. The Anglo-American Cataloging Rules and the Mini-MARC format for monographs have been followed where applicable. The data base and format are flexible enough to accomodate changes in these standards and in user requirements. CANISISS is accessed on-line at the Data Clearing House. The possibility of remote on-line access is being investigated.

The Data Clearing House is taking the initial steps towards becoming the organization which catalogues the data the first time, establishes the authorities, and distributes the information. However, the Data Clearing House is totally dependant on the willingness of people to report their holdings and must expend considerable resources in seeking holders of data. Recent Canada Council regulations requiring researchers to place their data in the public domain in a recognized archive are a beginning towards making the identification process automatic.

The goal of the Data Clearing House is to have an inventory of all quantitative machine-readable data in the social sciences in Canada or about Canada. Even partial fulfilling of that goal will assist in the rationalization of data cleaning and archiving activities in Canada.

With respect to the acquisition of data, the Data Clearing House lists only data that is available for secondary analysis. Multiple locations for data are recorded . The holder of the data sets the terms of access and the price. Data are not held by the Data Clearing House; only information about the data is retained. There is no central archive of data in Canada nor is it feasible to transmit the data on-line

now.

The Data Clearing House does not duplicate but complements the services of the existing data libraries and archives such as Machine-Readable Archives Division of the Public Archives of Canada, Statistics Canada, Institute for Behavioral Research at York, Social Science Data Library at Carleton and at the University of British Columbia, and the Centre de Sondage at the Université de Montréal to name a few.

Technical details of the data are recorded by the Data Clearing House to facilitate the determination of technical compatibility. A greater specificity of detail would probably be required by the local or regional archive or researcher.

The question of data comparability has been left to the peer group of the data originator. It is properly the work of the discipline-oriented professional associations and committees to assess the quality of the data and their potential for comparison.

Similarly, the Data Clearing House has not taken an active role in the training of researchers to use machine-readable data in the social sciences, considering this to be better accomplished by the professional associations and universities.

CONCLUSIONS

The Data Clearing House for the Social Sciences plays an unofficial coordinating role in the social science data area in Canada. The organization acts as a "broker" of information between the different data producers and data users. This role will continue to be assumed by the Data Clearing House provided that the market for its services continues to grow.

However, to provide access to social science data in Canada, financial support must be secured for the national coordination of data cleaning and archiving at a local or regional level. Technical specifications must be developed and enforced. Further work must be done in the area of data comparability and standardization. Training courses for using data in research and for archiving and cleaning data must be developed. Concerned data organizations are seeking solutions to these problems now so that better access to social science quantitative data can be ensured in the future.

REFERENCES

1. American Library Association. Catalog Code Revision Committee. Subcommittee on Rules for Cataloguing Machine-Readable Data Files. Final Report. Chicago: American Library Association, 1976.

2. Canada Council. Consultative Group on Survey Research. Report. Ottawa: Canada Council, 1976.

3. Dodd, Sue A. "Cataloging Machine-Readable Data Files - A First Step?" Drexel Library Quarterly 13 (January 1977): 48 - 69.

4. Hanis, Edward H., and Mitchell, S. Paula. "Publication Processing of Machine-Readable Numeric Data (MICDAT) Files for the Social Sciences." Canadian Journal of Information Science, 2 (1977): 26 -33.

5. Henry, Sharon E. "Bibliographic Control of Social Science Machine-Readable Data: the Data Clearing House for the Social Sciences Model." American Society for Information Science. Western Canada Chapter. Ninth Annual Meeting, Proceedings (September 21 - 23, 1977): 17

6. International Association for Social Science Information Service and Technology. Newsletter, 1 (November 1976 -).

7. International Conference on Information and Documentation in Social Sciences. Papers. Moscow: International Social Science Council, 1977.

8. Nasitir, David. Data Archives for the Social Sciences: Purposes, Operations and Problems. Paris: Unesco, 1973

9. "Unesco Ad Hoc Committee on Social Science Information." Data Clearing House for the Social Sciences. Bulletin, 3 (Spring, 1978), forthcoming.

10. White, Howard D. "A National Information System for Social Science Data Files." American Society for Information Science. Annual Meeting. Proceedings, 13 (1976), 54.