The University of British Columbia Data Library: An Overview

LAINE G.M. RUUS

THE UNIVERSITY OF British Columbia Data Library represents, I believe, a unique organizational model among data libraries (by which I mean to include, as well, data archives and data banks) in the manner in which it is jointly operated by the library and Computing Centre of the university. How it came to be as it is is a result of its historical development; it continues to function as it does due to the success of the original model.

In 1963/64, a Statistical Centre for the Social Sciences was established in the university's Faculty of Arts, primarily through the efforts of the departments of economics, political science, and anthropology and sociology. The purpose of the center was to provide statistical and programming consultation to faculty and graduate students in the Faculty of Arts, i.e., to act as an intermediary between the social scientists and the Computer Centre. By 1965 the Statistical Centre had entered into membership agreements with the, then, Inter-University Consortium for Political Research (ICPR) and the International Survey Library Association (ISLA), the membership arm of the Roper Public Opinion Research Center, then at Williams College. Through these memberships a small local collection of punched cards, magnetic tapes and codebooks was built up, in the Department of Political Science. This Political Science Data Bank was administered by a senior faculty member, and his students on a part-time basis. It was used only by a few faculty and students in the department, and consequently, small as the collection was, it was underutilized. The directors of the Statistical

Laine G.M. Ruus is Head, Data Library, University of British Columbia, Vancouver.

Centre were not primarily interested in the management of a data library facility; therefore, the Political Science Data Bank remained in the department for a number of years.

Dr. Jean Laponce, who was, for a time, chairman of the Statistical Centre Steering Committee, had adopted the idea that traditional libraries should assume responsibility for the management of collections of machine-readable data files (MRDF)—first promoted, I believe, by Ithiel de Sola Pool.[1] As early as 1965, attempts had been made to have the data bank's acquisitions and memberships funded by the library; this move to transfer the financial burden from the Faculty of Arts to the library was not permitted by the university senate. By the early 1970s, however, Laponce had successfully persuaded the university librarian and the head of the Computing Centre that they should jointly bear responsibility for the management of the local MRDF collection. In 1970, a proposal was drafted for the creation of a Data Library to be jointly operated by the Computing Centre and the library and: "to have the functions of acquiring, organizing, storing and servicing machine-readable data files. It would also assume liaison responsibilities in connection with other data libraries...."[2] This proposal was accepted by the senate, and in 1972 the Data Library was created.

Its first collections consisted of the data files originally collected through the ICPR and ISLA memberships, and a few other files collected from individual researchers, including a small collection of local 1961 and 1966 Canadian census data. The new Data Library was staffed by a research assistant from the Department of Political Science, transferred to the Computing Centre's payroll, and administered by a systems librarian on a part-time basis. In its second year of operation, the Data Library was finally staffed by a full-time professional computer programmer and a part-time reference librarian, and growth of staff and services since has been continual. The present organizational structure of the Data Library is as follows.

Mandate. The basic mandate of the Data Library, as defined in the original proposal, is "to develop collections and services in accordance with the academic requirements of the University, in parallel with the policies of the Computing Centre and the Library."[3] Because both "parent" institutions are mandated to serve the entire local academic community, this interdisciplinary focus applies to the Data Library as well.

Financing. Financing is derived from a number of sources. The Computing Centre provides a full-time programmer analyst, a computer dollar budget, physical plant, and all that is subsumed by hardware

and software support, as well as some minor office expenditures. The library provides a full-time librarian, a full-time clerical staff person, a real-dollar acquisitions budget, general collections, reference and technical services support, cataloging services, and such incidental expenses as supplies. In addition, to supplement the acquisitions budget, cooperative funding arrangements for the maintenance of very expensive subscriptions or memberships have been negotiated such that, for example, the Department of Political Science contributes part of the cost of the annual Inter-University Consortium for Political and Social Research (ICPSR) membership, and the Faculty of Commerce contributes one-half of the annual subscription costs of CRSP and COMPUSTAT data bases. Thus, the Data Library is in the fortunate position of deriving its primary funding from two of the most stable and secure budgets in the academic environment.

Staff. As noted earlier, the Data Library has a full-time staff of three. Training of these staff has been primarily carried on in-house to complement the experience and academic backgrounds of the three incumbents. Thus the librarian, who had no previous training in statistics or programming, attended both ICPSR summer school sessions in statistics and data management, as well as various local courses. The programmer, on the other hand, whose background included both statistics and programming, received only in-house training. Duties have been divided between these positions such that the librarian has been primarily responsible for acquisitions, reference, library-related technical services, and administration; the programmer has been responsible for user consultation, programming for public and internal purposes, and computer-related technical services; the clerical staff member is responsible for technical services vis-à-vis the codebook and reference collections, tape management, circulation, etc. In addition, for special programming or data management projects, part-time students are periodically hired whose academic backgrounds are suited to the nature of the project in question.

Collections Policy

As is, I believe, common, the collections policy is rather vague and ad hoc. The original mandate made only one stipulation regarding collections—that the Data Library "develop collections...in accordance with the academic requirements of the University, in parallel with the policies of the Computing Centre and the Library." A policy has evolved over the years, however, which can be outlined as follows: the

Data Library will collect automatically all significant Canadian data files, such as census data, election studies, and other major social surveys, public opinion poll data, etc. All other MRDF are acquired on request, tempered by considered need, potential for future use, and, of course, budgetary constraints. In addition, the library will function as a data archive in the sense that an attempt is made to acquire any original MRDF produced by local researchers, or offered for deposit by outside researchers (depository MRDF), and every effort is made to ensure that these are maintained for posterity. Because of the breadth of the mandate, it has not been necessary to restrict collections to those applicable to any one or limited number of disciplines. Other restrictions, however, are applied of a technical or contractual nature. To clarify, data files are not acquired or maintained which cannot at the very least be made available to any member of the local academic community, nor are MRDF acquired which contain confidential data, such as names, addresses, Social Insurance Numbers (SINs), etc., such that individual privacy is violated. Further, MRDF are not maintained which lack adequate documentation, or which are so "dirty" as to be useless for secondary analysis.

Composition of the Collection

Because of the previously mentioned factors, the collection at the present time cannot be called homogeneous, nor is it one of any great depth. It contains a rather comprehensive collection of Canadian census data, a good collection of Canadian Gallup Poll data, and most major Canadian public opinion surveys. On the other hand, there is a small collection of texts in English, French, German, Latin, Greek, and Amerind languages, but no major collection of texts from any one language, author or time period. Our fairly large collection of Canadian socioeconomic time series data is not yet balanced by an adequate collection of Canadian financial data. A small collection of health-related data is not balanced by an adequate collection of vital statistics data. And then there are many "odd-ball" files, such as digitized maps, a catalog of digraphs, and a large collection of satellite imagery of the northwest coast of North America. This sporadic collection is, in essence, then, nothing more than a direct reflection of our mandate—it reflects the research and teaching needs of the local academic community over the past two decades.

Technical Services

These tend to be of a preservative orientation rather than creative. This is dictated by the mandate, by the focus of the raison d' être of the Data Library as a library rather than as an archive, and by the primary user community which is the local academic community, rather than a wider national or international one.

For the purpose of maintenance, the Data Library collection consists of three parts: a collection of magnetic tapes on which are stored MRDF, documentation (i.e., codebooks) and some special-purpose programs. Each collection requires a separate set of technical procedures. They are as follows:

—MRDF are acquired, stored on tape, cataloged, and documented in the data base catalog. When necessary, MRDF are also "cleaned" and "translated" into system files formatted to the requirements of statistical packages such as SPSS or OSIRIS, as required by users.
—tapes are routinely copied, cleaned and rewritten.
—codebooks and other documentation are copied, or (if machine-readable listed to paper or COMfiche (if also very large). They are bound, assigned call numbers, and otherwise processed for circulation. Only very recently has it been possible, given time and staff constraints, to begin to create machine-readable documentation and cleaned (OSIRIS format) MRDF of depository files.
—software may be written for two reasons: (1) to rationalize and support in-house procedures, and (2) to make data access and retrieval by users as efficient and foolproof as possible. Such software is extensively documented, in standard Computer Centre user documentation.
—auxiliary technical services include maintenance of a reference collection, circulation system, personnel records, etc.

User Services

User services account for the major portion of Data Library staff time. The mandate stipulates very clearly that the Data Library's primary users are the local academic community, i.e., faculty, students and staff. However, increasingly in the past decade, the obligation of the university to serve the community at large has been reflected in the service policies of both the library and Computing Centre, and therefore also of the Data Library. Where possible, direct access to the holdings is provided to users in the government, commercial and private sectors;

and MRDF are copied on request for outside individuals and institutions. Our ability to provide these services is, however, very restricted by our contracts with MRDF suppliers, both individual and corporate.

The ideal (and therefore, almost by definition, seldom encountered) user transaction can be considered to consist of seven stages. For the sake of convenience, I will consider our user services within the framework of these seven stages.

Stage one consists of basic orientation in response to an initial user inquiry. Whether this be conducted in the Data Library, or in other information dissemination areas, such as the main library's or Computing Centre's information desks, the user has access to basic orientational materials which include information on how to access, as well as generate instruction in the use of, the Data Library's data base of data file descriptions (SPIRES catalog). Occasionally, this stage consists of in-class orientation lectures by Data Library staff.

Stage two consists of the identification of MRDF appropriate to the user. Because there is no union catalog of MRDF, this involves searching for MRDF in the local collection as well as other sources. To facilitate searching of the local collection, brief bibiographic records of each MRDF are maintained in the University Library's card and COM-fiche catalogs. In addition, extensive data file descriptions, including variable summaries, are maintained as an online SPIRES data base, which allows indexed and string searching of the contents of the data file descriptions. In the event that a required MRDF is not part of the local collection, the user receives extensive assistance in searching our reference collection to identify appropriate MRDF and sources thereof.

Stage three, then, is the acquisition by the Data Library of rquired MRDF not in the collection. Acquisition from outside sources is generally a lengthy, time-consuming, and occasionally frustrating process.

Stage four is the provision of documentation. The Data Library attempts to ensure that it has two copies of all codebooks, and at least one copy of all supplementary documentation, pertaining to MRDF in its collection. The SPIRES catalog record includes citations of all documentation, call numbers of codebooks, and other information necessary to access machine-readable codebooks or other system-generated documentation, as well as citaions of published works based on that data file. Novice users receive, of course, consultation on "how does a codebook mean?"

Stage five consists of access to the MRDF. All information necessary to mount Data Library-owned tapes is contained in the *Data Library User's Guide*;[4] the information necessary to access individual data

files—through a cataloged tape mount procedure which is transparent to the user—is contained in the SPIRES catalog record, including such file information as size, format, etc. Files can thus be accessed at all times that the computer is in attended mode. In certain cases of especially complex MRDF, particularly those heavily used by novice users, additional service is provided in the form of special-purpose software written to simplify and rationalize the retrieval of data and its storage in disc files in formats appropriate as input to other statistical programs. Documentation of these programs is also included in the *Data Library User's Guide*.

Stage six consists of the user's analysis. Service at this point is restricted to consultation, especially to novice users. The Data Library does not perform any analyses for users.

Stage seven consists of the user's publication of the results of the analyses. At this point, service consists primarily of consultation on such matters as citation formats for MRDF for inclusion in bibliographies and footnotes, etc.

There are, of course, other types of users who require other types of services, such as those wishing to deposit MRDF, those conducting surveys or otherwise creating MRDF, and even (once) a television game show producer wishing to upgrade the intellectual content of his show. Services to these users include consultation on questionnaire design, encoding, data file formats, creation of machine-readable documentation, etc., and agreements to archive MRDF for posterity.

Current Issues

Issues facing the Data Library today are numerous, certainly such as to preclude the possibility of sitting back and with complacency maintaining the status quo. Fortunately, few issues are administrative.

Staff recruitment and training—the Data Library staff had, over the past eight years, remained almost totally stable; unfortunately, in the past year two-thirds of the staff have been replaced. Because of previous lack of turnover, no staff training routines have been developed, nor has a full staff manual been written. Both of these lacunae must be filled; this can only be done in-house.

Documentation—whereas, at the present time, data files can be accessed by the local user virtually at all times, this is not true of most codebooks. Documentation which is only available in printed form is only accessible during the limited hours that the Data Library is open—our experience with placing copies of heavily used documentation in

other library locations with longer opening hours has not been entirely satisfactory for a variety of reasons. Many machine-readable codebooks, which are technically as accessible as any quantitative data file, are so large that it becomes very expensive merely to copy the file. In any case, the average user does not require a listing of the total codebook; normally he requires documentation only for a selected subset of variables. Thus, copying the codebook to COMfiche is not an entirely satisfactory solution. In order to maximize user access, we hope to develop software capable of searching machine-readable codebook records and retrieving those records satisfying user-specified conditions. A user could thus search a codebook and retrieve those variable descriptors of interest to him. Once all the codebooks are converted to this format, access will be much more efficient. There are, of course, problems associated with this, not the least of which is the choice of a standard format for codebooks capable of encompassing all types, including those that describe microdata, macrodata, textual data, representational data, models, etc.

Identification of MRDF—not all data archives and libraries publish catalogs or inventories of their holdings. Nor is there a comprehensive bibliography of those that are available. Thus, the identification of MRDF in other collections is a formidable task. This is a problem that we alone can do nothing more about than to disseminate as widely as possible our own catalog (in which admittedly we have been delinquent—our SPIRES catalog will be published in COMfiche format upon completion of retrospective conversion of all MRDF descriptions), and to promote in every way possible the compilation of a comprehensive bibliography of "finding aids" and, eventually, a union catalog of MRDF. This will not solve the corollary problem of identifying MRDF not in existing data library collections, especially those held by private individuals. The only presently viable solution to the problem of identification is the promotion of universal use of standard bibliographic citation formats in all publications.[5] Steps now have been initiated in Canada, through the Social Science Federation of Canada, to persuade publishers and editors to adopt as part of their editorial policy appropriate formats for MRDF citation. We can only hope for results and push for the adoption of these standards in Canada and elsewhere.

Access—this is often problematic whenever one is not dealing with a regular disseminator of MRDF, whether it be an individual or corporate body. Individuals are often very protective of their data until such time as they have published—sometimes ten or more years later. After

publication, the individual principal investigator often has not bothered to ensure the long-term maintenance of the raw MRDF, and/or not maintained adequate documentation, and thus the data are "lost"; corporate bodies are also often equally unconcerned about the long-term maintenance of raw MRDF. Government, in the Canadian experience, can impose extreme constraints to protect the privacy of the individual. Such policy recently, in combination with other factors, almost resulted in the total restriction of access to MRDF to be generated by the forthcoming Canadian census. Had this situation not been averted, it would have had disastrous consequences for Canadian demographic and social research. Possibly one solution to this type of problem might be the adoption and publication of a uniform code of ethics for data archives/libraries, including such issues as user access, dissemination and redissemination of MRDF and accompanying documentation, and the resolution of the thorny problem of copyrights vis-à-vis MRDF; some preliminary work in this area has already been done.[6] Certainly, government departments should be encouraged to standardize their MRDF dissemination policies, some of which are permissive, and some totally restrictive.

New services—large online quantitative data bases are proliferating on the international computer networks. Some data base producers offer to lease to individual users or institutions periodic "batch" editions of these data bases. As opposed to costly online subscriptions, this is a viable alternate mode of access for academic data libraries whose users often do not require "tomorrow's figures today." However, even the reduced academic rates charged for these data bases are high, and this method of updating data bases is inefficient. Some means of rationalizing access must be developed in order to eliminate unnecessary duplication, effort, expense, and time delays, but yet facilitate access to the data in a format allowing statistical analysis of the data without rekeying. We are currently considering the system developed at the University of Western Ontario for online batch access to the CANSIM data base, which utilizes the facilities of both local and national networks.

A somewhat different issue is the proliferation at various institutions of data archives serving specialized non-social science users, i.e., serving specifically one of the humanities disciplines or the hard sciences. The establishment of several types of data service facility within one parent institution will lead to further duplication of staff, facilities and services, the avoidance of which has been the major argument for the creation of data archives and libraries in the first place. Certainly there is a need for improved communication among these various types of data service facilities.

What of the Future?

The past five years have seen, locally, the wider promotion of the Data Library's services, development of needed software, production of the *Data Library User's Guide*, and the development of the SPIRES data base of MRDF descriptions; on the interinstitutional scene, this period has seen, among other things, increased awareness of a data archive or library "profession" and the creation of the International Association for Social Science Information Service and Technology (IASSIST), the establishment of regular courses in data library management, and the development of cataloging and citation formats. I hope that the next five years will see at least some of the developments outlined herein, including the widespread use of MRDF citation formats, the publication of a bibliography of the literature on MRDF management, a directory of existing data archives and libraries, an efficient liaison with online vendors of quantitative data bases, and the incorporation of these services into the normal sphere of the data library. Also, I hope to see increased communication between our data archives for the social sciences, for the humanities and for the sciences, and the coordination of these types of services at all levels.

References

1. Pool, Ithiel de Sola. "Data Archives and Librarians." In *INTREX: Report on a Planning Conference on Information Transfer Experiments*, edited by Carl F. J. Overhage and R. Joyce Harman, pp. 175-81. Cambridge, Mass.: MIT Press, 1965.

2. Kennedy, J., and Stuart-Stubbs, B. "The Data Library: A Proposal." Mimeographed. Vancouver: University of British Columbia, [1970].

3. Ibid.

4. Amos, David. *UBC Data Library: Data Library User's Guide*. Vancouver: University of British Columbia Computing Centre, 1980.

5. Dodd, Sue A. "Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines." *Journal of the ASIS* 30(March 1979):77-82.

6. Robbin, Alice. "Ethical Standards and Data Archives." *New Directions for Program Evaluation* 4(1978):7-18.