USER SERVICES IN A DATA LIBRARY

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

There seems to be some confusion as to just what is meant by user services. Although the term is relatively new to the library literature, the concept dates back to at least 1876 when, amongst others, Samuel Swett Green was beginning to argue "the disireableness of...personal intercourse between librarians and readers." Dictionaries of library science or librarianship do not yet define the term, nor do the library administration texts that I have consulted.

David Nasatir (1973) has outlined the components of user services in a data library as consisting of dissemination of data files; analyses on demand for users; training of users; consultation on such subjects as mathematics, statistics, methodology, data analysis, etc; the conducting of training programs; a "current awareness" function acquainting local users with parallel research being done elsewhere as reflected in the catalogues of holdings of other archives; and the creation of machine-readable codebooks. Alice Robbin (1977) has defined user services as consisting of reference (or finding the right data for the right user) training; data and documentation reproduction and dissemination; data preparation, processing, and analysis; project planning; and instruction and orientation. These then are my terms of reference when I speak of user services in the data library/archive context.

Samuel Rothstein (1961) defines reference service as "the personal assistance given by the librarian to individual readers in pursuit of information..." Given this definition, his "reference" = our "user services." His three levels of reference service are minimum, middling, and maximum.

His "minimum" or conservative level of service is that in which the librarian is merely a guide to the use of the collection, and the user is encouraged to be as self-sufficient as possible through user instruction and orientation techniques--a policy of "laissez-faire."

"Middling" service offers a great deal more service to the individual user (especially, in an academic library, to faculty and graduate students), including in-depth searching of the literature, compiling bibliographies, and generally doing a fair amount of the user's work for him. On the other hand, "conservative" service is given to the undergraduate student, on the assumption that learning to use the library and its resources is part of the educational process, and the student should therefore not be spoon-fed.

"Maximum" or liberal service relates primarily to the situation of a special or corporate library, where the librarian's *raison d'etre* is to do that part of the research that involves the literature, leaving the part that involves the laboratory (or whatever) to others qualified in other fields. That is, the emphasis is on the delivery of information, rather than on delivery merely of books, journals, etc., which might contain the information. The information delivered is authentic, relevant, and founded "on the impeccable scholarship of the librarian" (Rothstein, 1961).

Ti.e application of this scheme to the data library/data archive spectrum is not difficult.

On the conservative side there is, for example, the new, local-service data library, quietly growing in the bowels of some university structure--in an academic department or computing centre. Because at first it must concentrate on developing internally (you must have a collection before you can provide services based on it), this data library offers only basic services to users: acquisition of machine-readable data files (mrdf) on request, epecially if the researcher knows already where a file is to be had; maintenance of data files as they are received from outside sources without much, if any, cleaning or checking for coding, wild punches, inconsistencies, etc.; access to codebooks as they are received from the supplier; and some basic consultation on problems involving data processing and statistical procedures. Difficult problems are referred to experts: computing problems to the programmers in the computing centre, statistical problems to the statisticians in the statistical centre, data identification problems perhaps to the librarians in the library.

The middling level of service is exemplified by the same local service data library several years later, when it has grown both internally and in its services. Because the collection is now large, with many massive and

intricate mrdf, the data library has developed an on-line inventory describing in great detail the collection, to assist not only users but the data library staff as well in identifying and locating specific files. Data files acquired in a "dirty" state are cleaned, because it is easier to clean them at once, while one still has contact with the principal investigator, than several years later when a researcher needs the data and it is found to be unuseable. Codebooks are routinely converted to machine-readable form, because this is the most satisfactory way of ensuring that any user can get access to a copy at any time. (And how else does one economically supply a copy of a Gallup survey codebook to a political science class of 60 on one day's notice?) Special programs are developed, and special-purpose subfiles of data files are prepared, to make things easier (and cut down on hand-holding) for novice researchers, and more especially, as the only way to provide adequate service for 200 freshman commerce students who every year descend on the data library for their annual excercise with CRSP stock price data. (Although the staff has grown, it still cannot provide consulting to 200 students, and the only way to give them satisfactory service is to make the procedure as "idiot-proof" as possible.) Regular orientations, and some impromptu ones tailored to special courses, are given to introduce novices to the facility and its services, and there may even be a manual describing how to search the inventory, how to mount data files, special programs developed for often-used files, etc. There is now a variety of staff expertise that can be called on for consultation. But the "toughies" are still referred to the experts in other departments, and there is still strict adherence to the basic principle that the user should do the actual work himself, because this is after all part of the educational process, and the "true" researcher learns to be self-reliant.

And at the "liberal" far end of the spectrum, I envisage a special purpose data library embedded in a research institution, or a corporation, or a government department. Much of the work of this data library is involved in the actual creation of new data files, the maintenance of on-going data bases, and the secondary analysis of existing files. The data library saff here are experts in computer programming, statistics, sampling, etc.; these experts are part of the research team of any project, handling such details as the technical aspects of research design, the research instrument, actual data gathering, and, later, data analysis and interpretation. Here there is no question of the user doing his own work: The level of service given depends on the expertise of the staff; there is no need for such ancillary services as orientation tours and courses, of "idiot-proof" data files and documentation for novices, for there are no novices. The data base management system describing the collection is designed for maximum efficiency, and data files and documentation are machine-readable and very clean, because this is the most efficient means of maintaining and updating them.

Each of these levels of service has its own immediacy of purpose; which level of service a data library or archive approaches is dependent on its user community and the constraints placed on it by available funding and staff. Certainly the level of sophistication possible in the maximum service archive is far beyond the developmental capabilities of the small, local-service data library operating on two and a half people. But the techniques can be transported, and techniques which result in greater efficiency for the end user will also generally result in greater efficiency for the data archive staff as well. My dream, from the point of view of our small, local-service data library serving an academic institution, is eventually to make user services so efficient that no user need darken our door again, except to persuade us to acquire a file that we do not already have. There are, I think, several components to this. The first is to make information retrieval as efficient as possible, then to make documentatior access as efficient as possible, and finally, to make data access and software access as efficient as possible. These efficiencies need not, of course, be implemented in quite this order; one normally does the easiest things first. But I am going to treat them in the order in which most users approach them.

Efficiency of information retrieval becomes critical when the data library's collection grows beyond the point where any staff member can remember what every file in the collection contains. It then becomes necessary, not only for the sake of users, but also for the sake of staff, to be able to quickly find all particulars about any given data file--not just the principal investigator, title, or date of collection, but the individual variables contained in it. The consensus of practice seems to point to some form of data base management system (dbms). It is a little difficult to ascertain who has such systems operating. We know that Roper Center has had an in-house information retrieval system in operation for some years. We also know that ICPSR has offered (in October of last year) on-line access to its inventory through TELENET. SPIRES seems to be favourite among MTS installations, being in use at the University of Alberta Computing Centre, Stanford University Libraries, and U.B.C. Library. Other systems are being used by the University of Wisconsin-Madison and the University

of Washington. The Data Clearinghouse for the Social Sciences in Canada also had developed a data base management system, although its present fate is unclear. Among these systems, the amount of information offered on any given file varies immensely. None that I have seen, however, includes quite the detail envisioned for the data documentation system that IFDO is supporting, which contains 16 pages of (mostly optional) fields per record. The characteristics most desirable in a dbms for this purpose would seem to be that it:

- be flexible enough to handle a variety of file formats and interrelationships.
- be easy and cheap to maintain.
- have a powerful and flexible searching capability.
- be EASY to teach users to search.
- be available at all times that the computer is up.

In other words, as a user service, it should be capable of informing a user at a remote location of what files a collection contains, and should provide sufficient information to access the data without his visiting the data library.

There is another aspect to the question of efficient information retrieval, the retrieval of information vis-a-vis data files not held in the local collection--that is, the identification and location of data files held elsewhere--for eventual acquisition. This is not a problem that can be solved at the level of the local service unit. Present services are dependent on local collections of the inventories of other data libraries or archives, government departments, etc; literature searching; and the intuition of the local data library staff. What is needed is a union catalogue of the holdings of all known disseminators of mrdf, and some efficient means of access to information on what new data files are being created. The movement by ICPSR and the Roper Center towards on-line remote access to their inventories is a mojor step towards information retrieval. The data base developed by the Data Clearing House for the Social Sciences in Canada, which included the holdings not only of Canadian data libraries/archives, but also the holdings of government departments, the commercial sector, and private individuals, did for a time provide a union catalogue of Canadian holdings of Canadian data files, albeit in a restricted subject area. The data base was unfortunately not made remotely available, and the DCH is now defunct (no cause and effect relationship is implied). What is needed is a similar product at the international level.

Efficiency of documentation retrieval is more involved. Documentation comes in a variety of forms and formats. Some data files have no documentation at all, others have as their documentation only the second paragraph of a personal letter addressed to person X, and yet others have exceedingly complex documentation that is much longer than the data file that it describes. Documentation can be in hard-copy or machine-readable. The utility of machine-readable documentation for user services is obvious. It is the only way that documentation can be made maximally accessible, since in the optimum system, all machine-readable documentation is accessible at all times that the computer is up. Placing extra copies of documentation in a local library does not serve quite the same purpose, since I know of no library that is open as long hours as the computing centre. And a local computer system with a network of remote terminals scattered throughout an institution gives more immediate access than the library, which is invariably at least a twenty minute walk from where the user is (a very important consideration on the West Coast where it rains a great deal). However, to make the documentation machine-readable is not quite good enough. No one wants to read through the whole of the Six Nation Project codebook (Inkeles, n.d.), which occupies 9 linear inches of shelfspace, to find the twenty variables wanted. In order to maximize the efficiency of documentation retrieval, what is needed is some computer-assisted means of searching the codebook, so that one can retrieve just those portions describing variables containing specified key words. The Roper Center's proposed on-line system is designed to retrieve at this level of information. But there are many codebooks which are not amenable to this type of treatment, such as that describing the CANSIM system, which contains over 300,000 time series and as many variable labels. Maintaining a data inventory is one thing, but maintaining this level of information in an on-line data base is, for a small local-service operation, at present quite unfeasable.

Maximizing data access efficiency is dependent on the efficiency of tape mounting procedures in the local operating environment. However, it should be possible to develop a system such that, based on the information given in the information retrieval system, it is possible to gain access to all data files at all times that the host computer is functioning, not just at those times that the data library is open. Desirable features of data access would include: a minimum number of commands to access a file (whether on disc or tape); and maximum simplicity of commands--i.e., mechanical details such as mode, blocking factor, label, reel number, and position should be transparent to the

user. In addition, a data access system can be designed to maintain data security (e.g., check the permit status of computer ID's), collect statistics on tape mounts, perform an SDI function (referral to updated editions of files), and reinforce the moral conditions of data access (such as by cautioning against unauthorized copying for dissemination of files by users).

Of course, efficient access to data and documentation is of little value if one does not also have access to appropriate software. Software acquisition, creation, and support is very much dependent on the institutional structure within which the data archive/library exists. The smaller a local-service data archive/library, the more likely it is to be heavily dependant on the software resources of the host computing centre, with little possibility of influencing the computing centre's decisions as to hardware and software procurements, software to be supported, and future software (or hardware) developments. If the computing centre in question is user-oriented (which is the exception rather than the rule), if it provides comprehensive documentation, user orientation, introductory courses on commonly used statistical packages and programming languages, and extensive programming consultation, then the data archive/library need not provide these services. If, on the other hand, the computing centre is not user-oriented, the problem of providing efficient access to appropriate software becomes serious, and the solutions are neither easy nor cheap. As the amount of computing expertise dedicated to the data library/archive (or employed in it) is increased, however, an individualized level of service can be provided. This may extend to special-purpose housekeeping programs, to improve the efficiency of such invisible services as data file cleaning, rationalizing inventory, tape mounting and security. It may also extend to visible services such as writing a formula for the retrieval of data from a complex data-base, or creating a program to provide uniform access to and manipulation of aggregate data across several censuses. As the amount of computing expertise dedicated to user services is increased yet further, and basic software for common problems has been provided, yet more elaborate services become possible, such as the development of special purpose software.

Once one has all these systems operating, of course, one must teach one's users to use them. Seminars, special class presentations, and short courses seem to be the normal approach. A good addition is a user's manual, written not for the computer programmer (as most computing documentation is), but for the novice user. A history professor, with almost no computer experience, should be able, with the help of your manual, to sit down at a terminal, search your inventory for any files containing, e.g., illegitimacy rates in Ontario in the 19th century, have printed a copy of the (short) codebook on the local high-speed printer, and mount your data file, all on a rainy Sunday afternoon.

And then, of course, there are the special purpose subfiles, teaching packages, and special purpose documentation for the 200 Commerce students. Designing this kind of service requires more effort to be put into instructor-education than into student-education. It is essential that the instructor have designed the assignments so that they are appropriate to the data files he wants to use. The data library staff should be given ample warning of the assignment. It must know the expectations of the instructor, and the level of consulting he or she will give the students as part of the course. With adequate preparation by the data library staff, the onslaught can be relatively painless, and will have a marvelous effect on tape mount statistics

Throughout, I have been considering only "minimum" and "middling" service levels of the data library/archive. At both levels, the user's loing the job is considered part of the educational experience, and self-reliance is something to be fostered. I shall not consider here the extent and nature of special services offered in the maximum level data archive.

Each of the above components is composed of transparent services and apparent services. Apparent services I define as those which the user can see and appreciate. In this category I include the basic provision of information services; acquisition on demand of mrdf; consultation services of all kinds; provision of some kinds of dissemination services (e.g., relieving a principal investigator of the responsibility of maintaining and disseminating his data file himself); insuring efficient access to documentation, data, and software; orientation activities; provision of a user's manual, and so on. Transparent services are those which the user seldom notices or consciously appreciates, including the extensive level of reference services involved in the identification and location of fugitive mrdf, "on spec" acquisition of mrdf, rationalization of tape library and inventory procedures, routine data file cleaning and machine-readable codebook creation, and creation of special-purpose files to support special-purpose software, such as a map file to support graphic retrieval of census data.

Why this distinction? User services are usually the *raison d'etre* of the facility, and the reason for its continued funding. Apparent services are directly user-oriented. Transparent services, although often refinements of apparent services, are more often staff-oriented, and only indirectly contribute to that desirable phenomenon--the satisfied user. My conclusion is that, at the outset, a data archive/library should concentrate on apparent user services, in order to cultivate an active and supportive user community, and only later in its development, when this has been accomplished, develop transparent user services.

REFERENCE TOOLS FOR MACHINE-READABLE DATA FILES

John G. Kolp
Laboratory for Political Research
University of Iowa

It would appear to be a rather awesome task to try to summarize the current state-of-the-art in data reference; for surely after 16 years of steady growth in archives, reference materials giving access to the data in such archives should be exhaustive. Yet this is clearly not the case, although some attempts to provide useful reference tools in this area have appeared over the past decade. In the brief report which follows, I have singled out for discussion three categories of reference tools for machine-readable social science data which seem well-established.

Those reference tools which appear to play the most prominent role in directing users to appropriate machine-readable data are: (1) data catalogues describing the contents of individual social science data archives and data libraries; (2) directories describing the contents of more than one archive, or directories within special topical areas; and (3) periodicals, like *s s data*, which have attempted to report at regular intervals information on the holdings of social science data archives.

An attempt will be made to examine each category of reference tool from two very personal perspectives: (1) the user consultant who is continually asked to locate data files which must meet a number of very special conditions; and (2) the editor and compiler who must try to locate all known data files relating to certain topical areas and acquire information on the most recent acquisitions of data archives. In the former role, one is frustrated by the inadequacies in reference tools; in the latter role, one is amazed that we have come as far as we have.

Data Catalogues

Lists of holdings, guides to resources, archive directories, or data catalogues are available from most individual data repositories. Probably the most well-known of those documents which describe individual archives would be ICPSR's *Guide to Resources and Services*, issued on a nearly-annual basis since sometime in the 1960's. Others of this *genre* would include the recently-issued *SSRC Survey Archive Data Catalogue*, the *Steinmetz Archives: Catalogue and Guide* ((1978), the *B.A.S.S. Inventaire des Archives Disponsibles* (1975), *Catalogue of Machine-Readable Records in the National Archives of the United States* (1975), and the *Lokaliseringsoversigt* of the Danish Data Archives (1978), to mention just a few.

Physically, documents of this type are soft-cover, book-length descriptions of the holdings of a major archive. These are real publications, meant for broad distribution to a national or international clientele. Entries are often arranged according to some broad subject classification scheme which might include such major headings as community and urban studies, elites and leadership, mass political behavior, social welfare, religion, the international system, legislative and deliberative bodies, etc. The individual entries usually include title, author or data collection agency, population covered and/or sampling scheme, time period of study, number of cases and variables, distribution restrictions, and a brief abstract summarizing the purpose of the study and the focus of major categories of variables.