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 socal 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.

Another type of data archive catalogue is common to a group of archives which serve primarily as regional, provincial, or state-wide resource facilities. These are designed for consumption by individuals beyond the local computing environment as well as scholars from many departments on the local campus. Such catalogues may group entries according to the subject classifications mentioned above, although length and detail of information contained in the entries may vary considerably.

For example, the *Directory* issued in 1978 by the Data and Program Library Service at Wisconsin lists studies according to a reasonably detailed classification scheme, although each individual entry has very limited information. The University of British Columbia *Data Library Catalogue* (1974) has data files arranged alphabetically by title and individual entries are described with considerable detail. The *Annotated Listing of Data Holdings* of the Social Science Data Library at North Carolina organized entries according to a fairly detailed subject classification and at the same time included lengthy descriptions of each data file.

The final type of data archive catalogue is the one usually generated by the data library primarily to serve users of the local computing environment. Like those mentioned above, method of organization and detail in terms of individual data file descriptions vary considerably. Most are mimeographed, multilithed, or reproduced by some inexpensive method; many in fact are produced by line printers from machine-readable bibliographic files of various types. In my filing cabinet are such documents as the *Indiana University Political Science Data Archive Holdings as of May 1, 1971*, *Annotated Listing of Data Holdings, Polimetrics Laboratory, October 1975*, *File Inventory* (Latin American Data Bank, February, 1973), *Compendium of Data Holdings* (Center for Quantitative Studies in Social Science, Univerity of Washington, 9/76), *Data File Descriptions* (Public Affairs Information Service, University of Missouri, 5/2/74), "untitled printout" (Project IMPRESS, Dartmouth College, November 1978), and *Data Holdings (1976) with updates* (University of Iowa).

That is a brief review of the *types* of documents describing individual data archives. Now let us try to determine how useful each of them would be if we had to use them to search for a data file with previously-defined characteristics.

In my view, the usefulness of these catalogues is solely dependent upon (1) the scheme used to arrange the order of the entries in the catalogue; and/or (2) the types of indices that are appended to the list of data files. Data file descriptions or indices can be arranged by different schemes and below are listed twelve such schemes that have been used in various data catalogues. The schemes are listed according to my own judgement as to their usefulness as a reference tool; "1" is most useful; "12" is least useful.

1 subject or substantive categories (and subcategories) produced from keyword descriptors assigned to the data file independent of words in the title

2 unit of analysis or universe sampled

3 KWIC or KWOC of words appearing in the title

4 geographical area or location of sampled units

5 year or time period to which the data refer

6 author or principal investigator

7 data collection agency

8 depositor

9 first word of title

10 year data collected or archived

11 sample size or number of units

12 study number or archive accession number

No doubt, one may wish to disagree about the priorities suggested in this list, although I do hope that most would agree that "subject or substantive categories produced from keyword descriptors assigned to the data file independent of words in the title," is one of the most important reference schemes. Of course, if we would all heed Sue Dodd's advice on the construction of titles, No. 3 could be equally as important. It should be noted that I have placed author, principal investigator, data collection agency, and depositor indices some distance down the list. This is for a very good reason. In the old days (the late 60's) it was frequently the case that data files were very closely identified with a particular individual or research team, and thus references so arranged made a good deal of sense. Today, however, the number of data sets residing in archives is so large that personal references are starting to fade and the need for references which key on subject classifications, unit of analysis, and geographical area seem clearly of more importance.

A second area of note is "year or time period" which I have ranked No. 5. My academic training as an historian is, no doubt, part of the reason for the place this item holds on the list, although *time* has become a more important factor in social science research in the past few years. Time is important because of advances in the way we ask survey questions and in the way samples are drawn from populations; older surveys may be less likely to contain the kinds of questions a researcher wants on a certain kind of population. Also the time periods over which certain kinds of surveys have been taken, or certain kinds of questions asked, continue to grow longer, and we are now starting to build up impressive sets of files for longitudinal analyses. In addition, the analyses undertaken on these sets of longitudinal files can themselves be used to build up chronologically-ordered, aggregate files for sophisticated time-series analyses.

At the bottom of my list are things like "the year the data was collected" and "study number" and "archive accession number." These may be important things to know about a machine-readable data file, but they are not, in my view, of any help as a reference item in locating such files. These references are extremely useful for internal archive purposes, but do little to help those outside the archive find appropriate data. They are tools related to the acquisition process, not the data reference process.

The aforementioned priorities in data reference, however, would seem to be somewhat contrary to the actual practice of individual data archives. Upon examination of the data catalogues from thirteen different archives prepared over the past six years, one finds that six of the archives have used the internal study number as either the method of ordering the entries in the catalogue or as the distinguishing feature of a separate index. Another six have used the first word of the title, nine have used author or principal investigator indices, five the geographical area, and only two have used date of the study or time period.

On the other hand, five have used a good subject classification as the basis for ordering entries in their catalogues, while only another three have included useful subject or substantive indices. Only three provided indices to the unit of analysis--a reference which I think is extremely useful.

One does not wish to name names here, but it seems appropriate to award a first prize to the Steinmetz Archive for the most indices--nine to be exact. Five of the archives provided no indices at all with their data catalogues, although three of these ordered their data entries in such a way as to make the publication somewhat more useful than they might otherwise appear.

This discussion of the usefulness of various indexing schemes should not be taken as specific condemnation of the data catalogue of a particular archive, but has been developed to demonstrate that individual archives may not have always given adequate thought to the way in which descriptions of their holdings will be used by those outside their local computing environment. In addition, some of this talk may prompt further discussion along these lines and perhaps eventually some recommendations.

Directories

A data directory is distinguished from the data catalogues or the lists of holdings just described by one important feature -- it attempts to provide a useful reference tool to the machine-readable holdings of more than one social science data archive or data library. By this definition, few so-called directories would remain in this category. One would be selections from the "Directory of Directories," which appeared in the IASSIST Newsletter in 1977; others are the National Technical Information Service's *Directory of Computerized Data Files and Related Software*, the *Directory of Federal Agency Education Data Tapes*, and Vivian Sessions' *Directory of Data Bases in the Social and Behavioral Sciences*. All of these directories apparently contain references to data held by more than one repository, although in some cases these separate repositories may all be in the United States federal government.

The Sessions volume is probably more widely known among data librarians, so it may be appropriate to discuss its utility as a data reference tool. Published in 1974, and clearly out-of-date now, it is nonetheless of considerable interest because of what it attempted to do. The "major thrust of this directory," according to the Introduction, was "the identification of the nonbibliographic data bases." The title of the volume would suggest a concern with the social and behavioral sciences, but a number of factors undermine the promise of the title.

First, data bases were apparently defined as any systematic collection of data, primarily but not exclusively in machine-readable form. Second, whether these data bases were in the social and behavioral sciences was left to the reporting agency. Third, those data centers that were included in the volume (and there were over 600 of them) were also self-ascribed. Fourth, major subject classifications in the index represented a strange mixture of academic disciplines and sub-categories with the practice of public administration. And fifth, "it was inevitable that the primary organization of this directory is by data center."

When these five factors are evaluated alongside indices concerning institutional names, data center personnel, and geographical location, one comes up with a rather limited reference tool. The volume has little to do with what would normally be thought of as the "social and behavioral sciences" and in fact concentrates on municipal, county, regional, and other governmental agencies directly involved in the planning or administration of public programs. It is also not really a directory of the data bases, but rather a directory of places that collect and/or store data for purposes other than those normally ascribed to the physical sciences.

To my knowledge, no directory of this type has been attempted since publication of the Sessions volume. Had one been published, we can imagine that we would have wanted it to be organized and indexed in much the same manner as a catalogue describing the holdings of a single data library. The entries might be organized according to some general subject classification scheme, with four or five indices focusing on: (1) a more detailed subject classification scheme; (2) unit of analysis; (3) words appearing in title; (4) geographical area; (5) time period of study; etc.

Anyone considering a project to provide a general guide to machine-readable data in the social sciences or a directory on some special subject area would be well-advised to study the Sessions volume as well as the methods of organization and indexing found in data catalogues of the major data archives and libraries.

Data Periodicals

Periodicals issued on a regular basis provide one mechanism whereby information on the recent acquisitions of social science data archives can be transmitted to users of data reference materials. s s *data* is the prime example of such a periodical and the one that will be discussed in the remainder of this paper. For those not aware of its history and purpose, let us begin here.

s s *data* began publication in September, 1971, under support from a two-year grant from the National Science Foundation. Its purpose, as stated by G.R. Boynton in the initial grant proposal, was to fill a much-needed gap in reference information on the holdings of social science data archives in the United States and abroad. The grant proposal also envisioned that s s *data* was only a stop-gap measure--something to fill the information void for two or three years until a better system was developed by leading professionals in the data library field.

The plan suggested that s s *data* would reference on a quarterly basis the new acquisitions of all academic socal science data archives in the United States and as many Canadian and European archives as possible. It was estimated that this would be 40 or 50 new data sets each quarter or about 180 per year. While this figure probably over-estimated the acquisitions of academic archives in the United States (exclusive of the Roper Center), it was clearly an under-estimation of potential world-wide acquisition activity. The audience for this newsletter was thought to be first and foremost, social scientists, and second, data librarians and individuals involved in reference activities.

That is what s s *data* was supposed to be; what was it initially and what has it become over the past 8 1/2 years?

First, the periodical could not cover comprehensively the acquisition activities of all academic data archives in the United States for two basic reasons: (1) it was never possible at any one point in time to know which data archives were in existence and which ones were not; and (2) among those identified at any one time, the degree of cooperativeness in providing appropriate reference materials for publication varied a great deal. Of the approximately 25 archives who were initially contacted about their participation in the newsletter, only 18 provided a positive response to that request. Some never were heard from and apparently had gone out of business; others simply refused to answer their mail, although it was clear that the archive was still in operation. Those archives that did agree to participate provided information on their holdings at irregular intervals or in spurts.

During the first few years, for example, archives at Wisconsin, Iowa, Illinois, York, and ICPSR and the Roper Center provided information on a fairly regular basis. As other archives joined the list of participants it became more common for an archive to simply send a copy of its latest data catalogue and entries would be selected from these for each issue of the newsletter. A few, like the State Data Program at Berkeley, the Drug Abuse Epidemiology Data Center, and ICPSR established a regular procedure for transmitting information on new acquisitions, although all except the Drug Abuse archive have stopped doing so. New acquisitions at the Consortium, for example, are now only identified through new catalogues, news releases, announcements, etc., which are received second-hand from other faculty and staff at Iowa. Most archives do send the most recent issues of their data catalogues, and it is through these that most of the entries in *s s data* are obtained. Occasionally, however, batches of new acqusitions and announcements come drifting in from archives which are not otherwise heard from on a regular basis.

Because of the various problems involved in obtaining up-to-date information on the recent acquisitions of data archives, the newsletter has evolved into something a bit different than was originally envisioned. While *s s data* still tries to identify and report new acquisitions, increasingly the focus has shifted to the reporting on data files related to common areas of investigation. Two years ago, this shift in emphasis was offically noted in the newsletter--previously data sets had been reported according to the discipline of the principal investigator. Now they are reported according to subject areas within disciplines, or by areas of concern in public policy formulation, or by new fields of study that represent interdisciplinary approaches. For example, rather than simply grouping a set of files under *political science*, they are now listed under such headings as *elections*, *political attitudes*, or *legislative elites*--the latter being an interdisciplinary grouping of studies conducted by both historians and political scientists. Some recent public policy listings have included the *elderly*, *housing*, *conservation*, and *transportation and travel*; while some additional interdisciplinary listings have included *mobility studies*, *legal studies*, and *demography*.

In addition to listing data files under topical headings, an attempt has been made to reference (when possible) related data files which have appeared in previous issues. A recent issue, for example, included a section on *national legislative elites* which not only included descriptions of nine data files but also referenced sixteen other files appearing in past issues. A few readers have commented that they find this new method of listing and referencing past entries to be very helpful, but the majority of readers have remained silent on the issue.

Readership is another factor which has changed dramatically over the past eight years. During the two-year grant period (1971-1973), the number of subscribers surpassed 1000. The newsletter was distributed free to subscribers in the United States during this period, and most readers were individual social scientists attached to colleges and universities. Less than a quarter were probably institutional subscriptions. When NSF support ended in 1973, subscription fees of $2.00 for individuals and $4.00 for institutions were initiated and these have remained the same since then. The number of subscribers dropped rather markedly as soon as an actual fee was charged and in recent years has stabilized at about 400. The character of subscribers, however, has changed considerably. About one-third are IASSIST members who receive subscriptions as part of their annual membership in that organization. Another 50% are institutional subscribers, including primarily university and college libraries, public and private research centers, data archives and libraries, and a few metropolitan public libraries. The remaining 20% are individual subscribers, such as social scientists, information specialists, and planners and researchers in the private sector. Today, *s s data* serves the data reference community and not primarily the individual researcher, social scientist, or community planner.

In concluding this paper, what remains to be done is to reflect on the usefulness of *s s data* and the potential role of such periodicals in the future. *s s data* has not been and never will be able to cover comprehensively all the acquisitions of data centers in the United States, let alone those in Canada or Europe or the Third World. Second, even when references appear in *s s data* they rarely contain enough detail to immediately initiate the data acquisition process, if desired; numerous technical details are missing as well as specific information on all variables. Third, reference centers with a fairly complete set of data catalogues might find little need for periodicals like *s s data*. Fourth, although *s s data* has published descriptions of nearly 1000 data files over the past eight years, one would need to have a complete set of back issues and a cumulative index to easily locate references to particular kinds of files.

Given these criticisms (and all of them and more have been leveled at *s s data* since it began publication), what is the present and future role of such a periodical? The only real contributions that *s s data* can make in its present format are: (1) it can *highlight* the recent acquisitions of archives who are willing to provide such information; (2) it can gather together data files relating to specific disciplinary, interdisciplinary, or public policy areas; and (3) it can relate sets of data files to those of similar focus that have been described previously. If those functions are enough to justify its continued publication, then the newsletter can go on for some time in the future. However, it is important to raise the possibility that the print medium is now an out-dated mode of communication in the field of data reference. On-line systems at individual archives and projects whereby archives exchange data descriptor tapes certainly hold the promise of a much improved data reference system. Networking and other developments, such as improved cataloguing of machine-readable data files, also offer interesting possibilities.

It is hard to imagine that one system will dominate the field of data reference in the future, and therefore one important function that organizations like IASSIST can play is to recommend ways in which new and old reference systems can be effectively integrated in the future.

ON-LINE REFERENCE TOOLS FOR THE HARD SCIENCES

Gordon H. Wood
Canada Institute for Scientific and Technical Information
National Research Council of Canada
Ottawa, Ontario

I Ir .roduction

It is one thing to know or suspect that collections of machine-readable numerical data pertinent to one's discipline or problem may exist; it is something else to find that data, gain access, and use it profitably in one's research. The purpose of this paper is to review briefly the methods presently used to generate and access numeric data bases relevant to the so-called "hard sciences." Special emphasis will be given to the areas of on-line data retrieval and manipulation--areas where the state of the art in the "hard sciences" is generally conceded to be ahead of that in the "soft sciences." The reader wishing an inventory and description of the many scientific/technical data bases that are available worldwide is referred to the references at the end of the paper.

For the sake of clarity, it is useful to define a few terms as they will be used in this paper.

A.(Scientific/Technical) Numeric Data Base

An ordered collection of numbers whose values:

1. correspond to various properties, parameters or attributes of elements, substances or systems;
2. are critically evaluated by experts prior to their being included in the data base.

B.Numeric Data Base System

A numeric data base system consists of one or more machine-readable scientific/technical numeric data bases as defined above plus:

1. programs for searching, retrieving and organizing the data according to user selected criteria and, usually,
2. programs to manipulate the data.

(In general the latter property is what sets a numeric data system apart from a simple handbook or compendium. For example, a search routine may retrieve data giving the co-ordinate positions of the atoms in a given crystal. A simple command permits the user to calculate the various interatomic distances and the angles between the bonds joining the atoms. Another command generates a two dimensional drawing of the crystal projected along any desired axis or plane.)