INDEXING MACHINE-READABLE DATA FILES FOR A SOCIAL SCIENCE DATA ARCHIVES

by

Jacqueline McGee

Rand Corporation

"It is still true that the best retrieval system is the expert human mind."[1]

This paper was presented at the IASSIST Annual Conference, May 19-22, 1983, in Philadelphia, Pennsylvania.

INTRODUCTION

In the recent past much has been written and discussed about the problems of cataloging and bibliographic control of social science data. Many of these problems may have been resolved with the implementation of the Angle-American Cataloging Rules II, Chapter 9 (AACRII) and the MARC format for bibliographic control (2). However, there are a number of reasons these solutions may not yet be universally implemented.

For instance, the AACRII and the MARC format may be very familiar to library staff, but all archives are not staffed by librarians. Many archives are suffering from a shortage of staff and financial resources. Federal agencies produce a major portion of the data archived and used for secondary analysis and these agencies are also financially depressed. Researchers and programmers who use these Machine-Readable Data Files (MRDF) are not as aware of the problems related to the acquisition or storage of data and their interests do not necessarily correspond to the interests of the data archivist. Technological changes occur so frequently procedures may become obsolete by the time implementation occurs. And finally, so many new commercial firms are installing social science numeric data bases online and the interests of these firms do not lie in the same directions as those of the data archivist. To assist the novice who may be overwhelmed by some of these problems, it is the hope of the author this paper will provide some examples of simple record keeping.

Rowe and Byrum previously described a user-oriented system for the documentation and control of MRDF (3). This system was comprised of four parts. First, a standard catalog entry, second, a data abstract or description form, third, documentation codebooks and lastly, the records of physical and logical characteristics of the data set. It is not the purpose of this paper to offer an alternative system for the documentation and control of MRDF but to provide a practical example of implementing such a system. This example will provide an illustration for the person who has just received

responsibility for the safekeeping of a collection of MRDF or to establish an archive and isn't sure where to begin. The first item in the system by Rowe and Byrum, the standard catalog entry, was described before the Anglo-American Catalog Rules II, Chapter 9 were implemented. Data librarians located in a traditional library are already familiar with the rules for cataloging, but may not be familiar with the AACRII, Chapter 9.

The Anglo-American Cataloging Rules II, Chapter 9 describes the standard rules for cataloging MRDF. It is not within the scope of this paper to argue the pros and cons of the acceptability of the AACRII. There can be no doubt that a uniform standard defining MRDF is necessary in order to alleviate present confusing practices and the proliferation of titles for one data file. Certainly implementation of the AACRII and the agreement of the MARC format were giant strides in the cataloging and bibliographic control of MRDF.

STANDARD CATALOG ENTRY

Rowe and Byrum state "Standard catalog entries, constitute the primary records by which computer-readable data files should be controlled and accessed." It is with this one area of their discussion that I disagree slightly. The standard catalog entry requires extensive staff time and financial resources and need only be considered as necessary under certain conditions; if the required resources are available and may be allocated to such an endeavor; if the data archive or data bank is situated in a library or a library is available and willing to participate; if the data holdings are original data from the institution responsible for the establishment of the archive.

It is hoped non-originating archived data will be catalogued by the originating institution. However, the federal government is responsible for a major portion of the data files held in many archives and current fiscal restraints on most federal agencies probably will not permit such a project in the near future. There is, however, an ongoing cataloging project at Michigan's Inter-University Consortium for Political and Social Research (ICPSB) which may resolve the problem of cataloging federal data (4). ICPSR is certainly one of the largest, if not the largest, of the data archives in the United States. When this project to catalog their holdings is complete, it may be possible to consider a union catalog.

DATA ABSTRACT OR DATA DESCRIPTION FORM

It is the Data Abstract or Data Description Form described by Rowe and Byrum which should be given priority in the development of an archive record system. The data abstract or data description form in a standard format is an absolute necessity and should be the core of the documentation for the archiving of MRDF.

Aldrich has proposed a similar Abstract Form for the documentation of federal MRDF (5). It was from a description by Aldrich the following example was derived. Changes made in the form were for the benefit of the user and do not reflect a disagreement with her proposed standards. Since this form includes an abstract summarizing the data set or file being archived, this document shall be referred to here as a Data Base Profile. With some slight variations, the items contained in the form generally should contain the items shown on the next page.

If it is not possible or feasible for the archive or library to catalog the holdings of the archive according to AACRII at least the information supplied in the Abstract or Data Base Profile form will conform to the standards for describing MRDF. If at some future time cataloging is possible, the information for the catalog

Abstract Form for the Documentation of Federal MRDF

FILE #:	an identifying number for the individual archive
FILE NAME:	a title
FILE SOURCE:	producer, distributor, processor
PRINCIPLE INVESTIGATOR:	primary researcher
TYPE OF FILE:	survey data, microdata, administrative records, process records, geographic records, software
UNIVERSE:	total universe the records describe
SAMPLE SIZE:	number of observations or records
SAMPLE UNIT:	household, person or unit being measured
RESTRICTIONS:	none, or any restrictions placed on the distribution
ABSTRACT:	a summary description of the data set or file. Each abstract held in the archive should contain the same information. Care should be taken not to omit any portion of the required information and the information should appear as closely as possible in the same paragraph. This assures an easier search for the individual looking for specific information as to size of the sample, purpose of the study, key variables, etc.
REFERENCES:	a descriptive listing of the hard-copy documentation available for use with the data file, i.e., code-books, survey instruments, dictionaries, etc.
RELATED PRINTED REPORTS:	known reports where the data is described or where the data file has been used
TAPE SPECIFICATIONS:	the physical characteristics of the data file, the tape numbers, the logical record length, the block-size, data set names and density

entry will be readily available.

Copies of the Data Base Profile may be stored in computer format as well as in hard-copy. If the data is stored in computer format it would be possible to devise a simple online search capability. If the data librarian wishes to be bibliographically correct, study the AACRII and include in the Profile the pertinent information from the AACRII as well as the information required or deemed necessary for the institution housing the data library (7). Many of the elements of the Data Base Profile may be utilized to produce a catalog. For instance, each abstract when extracted from the Profile provides summaries of the archive holdings. The Rand Data Facility catalog uses the abstracts in such a way.

Each abstract then written, therefore, should include the following (6):

- Data base identification number
- Source
- Name
- Date the information was collected
- Subject
- Geographic level of the data (lowest)
- Population or sampling unit
- Number of observations or number of logical records
- Key variables

Indicies may then be derived from the information given on a Data Base Profile.

SOURCE AND NAME INDEX

The Source and Name Index is derived from the File Name and the File Source as given in the Data Base Profile. Often these items are sorted as separate indices; an author index and a title index.

An index should lead a user to the information he is seeking with as little effort as possible, and so we have combined these indices.

On the Data Base Profile and in data file records we use the most correct name for a data file. The correct name may be drived by using the AACRII rules. Since we also wish our individual indices to assist the user in his search we also include in our local archive index those aliases or acronyms when they are commonly used, if the index is to prove useful.

In order not to have a great many "see....." included in the index where aliases or acronyms or common usage names are listed, the correct identifying number of a particular data file is used as the pointer.

KEYWORD INDEX

This index is certainly one of the most difficult to construct. A thesaurus would be helpful; however, the keyword index discussed here was developed from individual data files. As mentioned earlier, one of the mandatory sections of the Data Base Profile is a list of key variables. At the time of archiving a new data file, the abstract is written and the key variables extracted. These variables or subject categories are added to the end of the keyword index. A copy of the keyword index is kept online. New keywords are added at the end of the old index and a short SAS program sorts the keyword index by words or by identifying file numbers whenever necessary.

GEOGRAPHIC LEVELS AND MAJOR SUBJECT VARIABLES IN THE KEYWORD INDEX

Using Census Bureau designations, the lowest geographic level of the data is assigned as a second element of the keyword index. By using the lowest geographic level it is possible when searching for data to weed out those data files not useful to the researcher.

A third element for the keyword index is a prescribed list of major subject variables. For each data file the keyword index will include at least

one, but not more than three major subject categories. It is then possible to produce tables listing data files by geographic areas and major subject categories. Librarians may want to use the Library of Congress Subject Headings (LCSH) List.

The Data Base Profiles may be produced as printed copies to be given to researchers interested in using a particular data file and can be used as documentation for bibliographic citations in research papers for the creating of a catalog of holdings.

The Data Base Profiles stored in a partitioned data set at Rand are soon to be converted to a total online system using the IBM INFO/SYS CSD data retrieval system.

The OZ INFO/SYS has the capability of handling multiple data bases and has been utilized at Rand recently to create an information system for a library of software models as used in one department (8).

With the data stored in OZ it will be possible to search the system for a particular data base by title, by keword, or keyword combinations. Requesting data bases by keywords produces a "hit list" of all data bases that contain the requested keywords. The hit list can then be accessed in either of two forms: one is an abreviated form that lists only a few specifics about the data bases on the list (including the data base number and title); the other is a full description of the data base. It is possible to get print copies of an OZ screen, of the individual data base descriptions and of the hit list of keywords.

References

(1) Michael Phillipson, "The Classification, Storage and Retrieval of Survey Data, Reader in Machine-Readable Social Data,: H. White, ed., Information Handling Services, Englewood, Colorado.

(2) See Sue Dodd, "Cataloging Machine-Readable Data Files: An Interpretive Manual," American Library Association, December, 1982.

(3) John D. Byrum, Jr. and Judith Rowe, "An Integrated, User-Oriented System for the Documentation and Control of Machine-Readable Data Files," Library Resources & Technical Services 16(3), Summer, 1972.

(4) Carolyn Geda, "Marc Formal Applied to Machine-Readable Data Files: A Pilot Project of ICPSR," Paper prepared for delivery at the annual meeting of the Society of American Archivists, September 1-4, 1981, Berkeley, California.

(5) Barbara Aldrich, "Proposed Standards for Bibliographic Entry and Abstracts for Federal Machine-Readable Data Files," Redraft 2, October, 1978.

(6) Don Trees, "The Rand Computation Center: Rand's Data Facility: A Guide to Resources and Services, " BCC-1555/18, Santa Monica, April, 1981.

(7) Anglo American Cataloging Rules II, Chapter 9, Machine-Readable Data Files, P201-216.

(8) As described by William Fowler, The Rand Corporation, Santa Monica, California, 1983.