Toward Integration of Catalog Records on Social Science Machine-Readable Data Files Into Existing Bibliographic Utilities: A Commentary

SUE A. DODD

FORMAL RECOGNITION OF THE NEED for bibliographic control over computerized information has slowly been evolving within the library and information science profession over the past several years. A major landmark that helped to focus increased interest in the cataloging of social science data files was the inclusion of chapter nine on "Machine-Readable Data Files (MRDF)" in the second edition of the *Anglo-American Cataloging Rules* (AACR2).[1] Publication of these rules in 1978, coupled with a number of other events, including the compilation of a machine-readable catalog (MARC) format for machine-readable data files, provided the important links that would facilitate the integration of bibiographic records into local automated systems and eventually into national information systems.

The most recent cataloging code (AACR2) and the MARC format for MRDF provide the standards required for describing and creating automated records, which in turn can be applied to many different purposes, such as shared cataloging, acquisition systems, and the building of a union list on all available MRDF. The primary purpose of this paper is to provide a commentary on the significant steps that have contributed to this current level of bibliographic control and to outline some of the remaining problems still to be considered before MRDF bibiographic records can successfully be integrated into existing bibliographic utilities. (A bibliographic utility as referenced in this paper is an organization that maintains a large bibliographic data base in an online mode via communications lines enabling it to offer computer

Sue A. Dodd is Associate Research Librarian, Institute for Research in Social Science, University of North Carolina, Chapel Hill.

base support to any interested users, including designated network participants.)

Overview and Definitions

If it is true that advances in modern information technology have far exceeded our ability to control data products generated by a computer, then it is equally true that the proliferation of the various types of data files has impeded our ability to apply a consistent vocabulary to describe and distinguish one from the other. Differing vocabularies have emerged depending on which segment of the information community is speaking. To the bibliographic-oriented portion of the information profession, "data bases" are machine-readable bibliographic files, whether produced by a library or by the American Chemical Society. When information other than that represented by the bibliographic journals and indexes was marketed online, then the terms *bibliographic* and *nonbibliographic* emerged. However, the use of this terminology encouraged others to offer another approach. Sessions suggested: "Although the terms bibliographic and non-bibliographic data seem clear enough, the negative term can be eliminated and a clearer relationship between the two kinds of information established by referring to *primary* and *secondary* data files."[2] To Sessions, *primary* would be equated with original (or primary source of) information. The computerized version of the census data, for example, would be considered primary, while the resulting printed census volumes plus the bibliographic references to census documents would be considered *secondary*. Primary data sources in the social sciences predate the "online revolution" and bibliographic data bases. In fact, census data were the first to be represented in punched card form and the first to be computerized by means of UNIVAC I in 1951.[3] Public opinion data represented by the established pollsters were another early source of computerized data, and as the collections of public opinion data increased, data archives and libraries were established to maintain these collections. The earliest and largest collection of public opinion data in the world today, which dates to 1935, is that of the Roper Public Opinion Research Center, founded in 1946.

To social scientists, a "data file" or "data set" will most often refer to a set of numeric values that can be manipulated by a predesigned statistical routine. Characteristically, data from numeric files are statistically manipulable and subject to quantitative analyses. Such files are manipulated using different forms of statistical software, such as tabu-

lation programs and econometric modeling programs. Numeric data can result from surveys of households and/or individuals, from scheduled censuses, from administrative records or economic reports, from test scores, and from other sources of statistical information. Data obtained from surveys or censuses can be classed into two groupings: summary data and microdata.

Summary data are aggregations of individual record data. Totals and frequency distributions show numbers of persons, families, housing units, corporations, vehicles—whatever the unit of enumeration is—distributed by their various characteristics for different geographic areas. A subset of summary data are the so-called time series data files. Time series are observations of discrete variables—such as the price of wheat or grain, the GNP, or the employment totals of an industry—for periodic intervals, such as months or years. Microdata are unaggregated data, produced from basic household and person unit record data (i.e., the actual responses of each person who completes a questionnaire).

A major producer of social science data is the federal government. A significant proportion of federal social and statistical data is disseminated by five general-purpose collection agencies: the Statistical Reporting Service of the Department of Agriculture, the Bureau of Labor Statistics of the Department of Labor, the Bureau of the Census of the Department of Commerce, the National Center for Education Statistics (NCES) of the Department of Education, and the National Center for Health Statistics (NCHS) of the Department of Health and Human Services.

With the appearance of cataloging rules for computerized data and the integration of such data into traditional library collections, still another vocabulary emerges. The generic term for computerized information that has been offered by the library community is "machine-readable data files." According to AACR2, a machine-readable data file is defined as any information encoded by methods that require the use of a machine (typically, but not always, a computer) for translation. The justification for the selection of this term by the ALA Subcommittee for Cataloging Rules for Machine-Readable Data Files is documented in their final report:

> Frequently-heard designations are those introduced by the word "data"; "data record," "data set," "data file," "data base," "data bank," etc. To many these terms convey a sense of size, a "data item" being the smallest unit, and "data base" or "bank" implying the largest accumulations. Between these extremes, "data set" and "data file" are sometimes used interchangeably, but "data file" is more unambiguously defined as a collection of related records to be treated

> as a unit, while definitions for "data set" vary according to computer languages, glossaries, and individual usage. However, any designators which do not take into account the means of access to the information do a disservice to the catalog user, as any of the terms introduced by "data" could conceivably apply to information in another medium.[4]

AACR2 further defined the generic term by stating that "*machine-readable data file* embraces both the data stored in machine-readable form and the programs used to process that data."[5] Consequently, the term *machine-readable data file* or its acronym MRDF as used throughout this paper will stand both for data files and program files.

The term *machine-readable* is easily understood, especially when it can be equated with computer-readable, but the term *data file* still warrants more explanation. A data file is defined here as any organized collection of automated records that are related in some way and treated as a unit, e.g., a payroll file with one record for each employee, showing his rate of pay, annual leave, deductions, etc. In most cases, the reader should conceptualize a singular MRDF to be an "inert file"—that is, existing alone as a separate entity on any number of data carriers such as a magnetic tape. It is this "inert file" of computerized information that conceptually becomes the "item in hand" to be described.

The opposite of an inert file may arbitrarily be defined as a "dynamic file" or a "dynamic data base." A dynamic data base is one that is characterized by its fluid and constantly changing nature. It may be represented by economic time series, or bibliographic data bases, and may be corrected, revised retrospectively, updated, merged, partitioned, and blocked into subfiles without changing its bibliographic identity. Even though these data files are associated with online systems, many are also available on a magnetic tape subscription basis and could conceivably become part of a library's collection of informational resources represented by a serial catalog entry.

Events Contributing to Bibliographic Control of Social Science MRDF

Although the early abortive attempts in 1957 to involve traditional libraries in the acquisition and management of social science data files have been well documented,[6] it was not until the early 1970s that the library profession began to take a bibliographic interest in MRDF. In January 1970, the Executive Committee of the ALA's Cataloging and Classification Section instructed the Descriptive Cataloging Committee to form a Subcommittee on Rules for Machine-Readable Data Files.

Their mandate was to recommend methods of describing data files that would be compatible with existing cataloging procedures for other media. This effort grew out of the perceived need to apply some type of bibliographic control to data files that were actively being collected by academic and research institutions. In the absence of any local data archive or center, these materials after their initial collection and application were often brought to the library for processing. Faculty members were asking librarians to acquire MRDF for them, just as they would request an important book or reference work. According to Byrum, the establishment of the subcommittee marked the formal recognition of the need for standards by which libraries could assist in the control and access of data files which academic and other institutions had already begun to collect as an additional and increasingly important resource of educational and research value.[7]

First under the direction of John D. Byrum, Jr., chief of the Descriptive Cataloging Division, Library of Congress, and later under the direction of Elizabeth Herman, Technical Services Department, University of California at Los Angeles, the subcommittee met twice a year for five years, drafting position papers and making recommendations on every component of the catalog bibliographic record. Their final report was filed in January 1976, and it was this document[8] that laid the groundwork for chapter nine in the second edition of AACR, which in turn introduced rules for cataloging MRDF for the first time.

National Bureau of Economic Research (NBER) Workshop

The subcommittee made an effort to gather feedback from nonlibrary audiences who represented the data processing or data producing community. An important forum for an exchange of ideas and information on bibliographic aspects of MRDF took place in 1974 with the National Bureau of Economic Research's conference on "The Computer in Economic and Social Research," and its workshop on "Documentation of Large Machine-Readable Statistical Data Sets." The focus of this workshop included an evaluation of the recommendations of the subcommittee's work to date. Early cataloging examples were presented at the workshop, along with a checklist of descriptive bibliographic elements.

An additional focus of the workshop was a discussion on the content and format of literature citations for social science data files. It was recognized that an accurate and complete literature citation for social science data files would benefit the researcher and potential user alike and would pave the way for social science data files to be included

in printed bibliographies, end-of-work references, and indexing and abstracting works such as the *Social Sciences Citation Index*. Today, guidelines on how to create a bibliographic citation are available to the reader,[9] and a major journal in the social sciences—*Social Forces*—now carries in its "authors' guide" section instructions with examples on how to cite a MRDF in the literature.

Early Cataloging Efforts

During the five years that the subcommittee met, the members had written and verbal contacts with data librarians who were beginning to catalog data files. Even as the subcommittee was meeting and debating on the bibliographic elements of MRDF, several research-oriented libraries and data centers were beginning to compile catalog records on data files held by their respective institutions. In Canada, the University of British Columbia and the Public Archives of Canada (Ottawa) took the initiative; and in the United States, it was Yale University (Social Science Data Library) and the University of North Carolina at Chapel Hill (Social Science Library). In other cases, like the University of California at Los Angeles, it was the automated 1970 census records that became the first MRDF to be included as part of a library's collection.[10]

IASSIST

The ALA subcommittee was not the only group working to define standards for MRDF. Others included the Computer Media Working Party of the Library Association's Media Cataloging Rules Committee, the American Society for Information Science (ASIS) Special Interest Group for Non-Print Media, and the Association of Educational Communications and Technology (AECT) Cataloging Committee. Another newly formed group—the International Association for Social Science Information Services and Technology (IASSIST)—was established with a special subunit devoted to promoting cataloging and classification procedures for social science data files. As chairperson of this organization's Classification Action Group, I directed a special project aimed at testing the feasibility of cataloging MRDF. The participants were members of IASSIST who had expressed an interest in classification, although many were neither librarians nor catalogers. A brief manual based on the position papers of the ALA subcommittee was given to the participants, and each was asked to select six MRDF of numerical, text or program files and proceed to catalog these files, keeping records on time spent, problems encountered, etc. The rationale for this project was based on two major assumptions: (1) that data

files and programs are underutilized, and the lack of knowledge on the availability of existing data files has hampered the academic community and other interested parties in the ongoing process of scholarly research; and (2) that there is no standard format for providing information on the availability of data files which would make possible one central source of information. The library profession's historical commitment to standards and to providing bibliographic information on a variety of media provides a natural background to study the feasibility of applying library cataloging and classification procedures to MRDF.

By testing the ALA subcommittee's rules and providing the initial cataloging experience, it was expected that the most immediate outcome of this committee's work would be to help pave the way for the inclusion of MRDF catalog records into the local or most appropriate public facility. The effort yielded over forty individual catalog entries from nine participants representing the following institutions: National Opinion Research Center, Data Use and Access Laboratories (DUALabs), Yale University Social Science Data Archive, Drexel University Graduate School of Library Science, University of Pittsburgh Social Science Computer Research Institute, Rutgers University, and the National Archives Machine-Readable Archives Division. The variety of MRDF represented in the project was significant as well as interesting. The types of MRDF included text files, bibliographic data bases, census and census-related files, survey data, panel studies, time series, aggregate data banks, longitudinal files, serials, computer software programs, mathematical models, online program lessons, educational data packages, and simulation games. All of these different kinds of MRDF with their unique characteristics were successfully cataloged within the scope of the subcommittee's recommendations and with the guidance provided in the manual.[11] While this project helped to establish the feasibility of cataloging MRDF by many different parties with varying degrees of cataloging experience, it also helped bring to the surface some of the problems that have come to be associated with the overall cataloging effort.

National Cataloging Conference

In March 1978, a national Conference on Cataloging and Information Services for Machine-Readable Data Files was held at Airlie House, Warrenton, Virginia. It was funded by the National Science Foundation and was organized by DUALabs. This was the first concerted effort at a national level to develop standards and to suggest cooperative efforts in establishing bibliographic control for MRDF. This meeting brought

together key persons and organizations having an active interest in establishing a framework within which a national program of cataloging and information services could be developed. A heavy emphasis was placed on the problems of federal data producers and publicly available data.

While the conference did not attempt to provide solutions to the problems associated with applying standardized bibliographic control procedures for MRDF, there was general consensus that such procedures and related information services are urgently needed to improve user access to machine-readable data resources. The conference concluded with a "call for action" and with several recommendations, including the following:

—that "the AACR2 rules should be tested on a broad range of MRDF to determine the feasibility of using these rules as a standard for cataloging."
—that any resulting procedures should be directed toward an automated system of bibliographic records for MRDF;
—that the Library of Congress should be encouraged to design and establish a MARC format for MRDF;
—that products and services which could be derived from such a cataloging effort be defined; and
—that the feasibility of integrating the resulting catalog records into existing network systems be investigated.[12]

Federal Task Force

Immediately following the Airlie House cataloging conference, the Office of Federal Statistical Policy and Standards (OFSPS) took action to establish a mechanism for using the staff resources contributed by various federal agencies. The result was the establishment of a federal task force, which in turn, would coordinate federal efforts to develop acceptable standards for cataloging MRDF. Under the task force's direction, a small interagency working group developed standards for statistical data files as they apply to creating bibliographic citations and abstracts. These procedures are presently being applied by several agencies in an effort to produce more informative and reliable directories of federal MRDF.

In October 1979, the Bureau of the Census issued a new inventory of their holdings, entitled *Directory of Data Files.*[13] With this *Directory*, the bureau has incorporated the task force's standards for citation and abstracts. The citation may be characterized as a "minicatalog" entry which includes the International Standard Bibliographic Description

(ISBD) punctuation. This effort was a tremendous breakthrough on the federal level, and it made the link between descriptive practices of the federal sector and the existing bibliographic standards of the library and information science community much closer.

In a related effort, the OFSPS established an Interagency Committee on Data Access and Use. This committee, in turn, initiated a multiagency project to adapt these same standards and produce a comprehensive directory of federal statistical data files. The *Directory of Federal Statistical Data Files,* issued jointly by the Machine-Readable Products Division of the National Technical Information Service (NTIS) and OFSPS, contains more detailed bibliographic information than past efforts. (For a further description of the directory, see Duncan's article in this issue of *Library Trends.*)

OFSPS has now moved to the Office of Management and Budget (OMB), and it is expected that OMB will issue a Statistical Policy Directive on Standards for Abstracts of Public Use Statistical MRDF. Such a directive would help to institutionalize the *Directory* as a regular periodic publication and to establish uniform standards among federal statistical agencies. Standing behind the directive will be *Technical Paper No. 3: Procedures for Preparation of Abstracts of Public Use Statistical Machine-Readable Data Files.*[14]

Cataloging Manual

The cataloging manual that had been used in the IASSIST-sponsored cataloging test had to be revised. After the ALA subcommittee issued its final report, the Joint Steering Committee of AACR2 made further changes and recommendations. With the appearance of chapter nine in the second edition of AACR, a new manual was planned, and its scope was extended to include basic procedures for proper bibliographic control and additional levels of recordkeeping associated with library management of data files. Its objectives were broken down into five broad areas: (1) to provide guidelines for establishing bibliographic conventions for MRDF (especially for those data producers or distributors in need of guidance or structure in this area); (2) to suggest integrated levels of recordkeeping for MRDF; (3) to bring into sharper focus the AACR2 rules as they relate to cataloging of computerized files; (4) to provide notes, examples and interpretations of MRDF cataloging which would otherwise not be available; and (5) to provide working tools for those cataloging MRDF for the first time.

Assistance was sought to support the work on this new manuscript entitled *Cataloging Machine-Readable Data Files: An Interpretive*

Manual, and in August 1979, I received funding from the Council on Library Resources (CLR). The grant from CLR was funded under the auspices of its Bibliographic Service Development Program (BSDP), which has focused on the development of a set of strategies aimed at establishing bibliographic control of materials in libraries and sharing the bibliographic data that is produced. What followed was a period of investigation and research into the informational needs of both catalogers and users of MRDF. It was determined that the interpretive aspects of the manual should fall on the side of the many intricacies of computerized information in general and on the unique characteristics of certain classes of MRDF in particular. Experts were consulted in the areas of computer hardware and software, computer cartography, language/text processing, simulation models, federal statistics and survey data. Site visits were completed to the Library of Congress and to academic and research libraries, including Princeton University, Columbia University, and University of Michigan. In addition, an exchange of information among those data centers and libraries engaged in early cataloging efforts was developed. As a result, the manual contains explanatory information plus cataloging examples on many different types of MRDF, including survey data, federal statistical files, cartographic programs, econometric models, computerized dictionairies, Greek text files, and economic time series. Associated terminology is defined and a glossary of MRDF-related terms is provided.

The biggest difference between the cataloging of books and the cataloging of MRDF is that the cataloger normally does not have an "object in hand" which he is able to describe; and even if he did, it would not do him much good. External descriptive labels on magnetic tapes are not permanent, nor do they carry the customary prominence or authority associated with external labels for other media (e.g., sound recordings). According to AACR2, the chief source of information for an MRDF is the internal user-header label (an option available on standard labeled magnetic tape reels). Lacking this label, the chief source of information for an MRDF is the accompanying documentation generated by the creator, producer, etc., of the file. Documentation is a generic term covering a wide range of descriptive items, such as a data dictionary, tape layout, codebook, and user's guide. Both an internal user-header label and documentation external to an MRDF are discussed in the manual, and selected types of documentaion are provided as examples.

With any medium, the quality of cataloging depends on the so-called authority or prominence of the source from which bibliographic

information can be obtained. In the past, very little attention has been given to the importance of providing complete bibliographic information to an MRDF or its external documentation. Many of these external descriptive sources relating to the content and organization of a particular file have little or no file-specific bibliographic information which could provide some authority for the cataloger. Without a standardized title page, the number of useful descriptive elements varies from file to file. In addition, certain numeric data files have no titles at all, while others may have two or three loosely associated titles. Before satisfactory cataloging efforts for MRDF can take place, some external controls must be exercised over the information describing this new medium, and guidelines establishing proper bibliographic conventions must be outlined. To address this problem, chapter three of the manual provides guidelines on how to create a descriptive title, title page, bibliographic citation, and data abstract.

Additional levels of recordkeeping are required to maintain MRDF in any library collection. Some of the MRDF recordkeeping practices currently in operation by libraries and information centers are reviewed, and suggested integrated levels of recordkeeping for MRDF are outlined.

Because the new cataloging rules for MRDF have not been tested on a large scale, it was necessary to match the rules with specific examples. Extensive applications of the rules were tested on a variety of files and programs and the results were documented and explained throughout the text. Specialists at the Library of Congress were consulted on rule interpretations, and LC policy interpretations as they relate to MRDF have been noted.

The first draft of the manual was reviewed in late December 1980 and early 1981. It is expected that the final version of the manual will be published in early 1982.

MARC Format for Machine-Readable Data Files

On June 1, 1979, the Library of Congress Network Development Office, in cooperation with the LC Automated Systems Office, announced that it would begin work on compiling a MARC format for machine-readable data files. The project was under the direction of Lenore Maruyama, and to assist her in this effort, an advisory committee of individuals who were actively involved with standards for bibliographic control of MRDF was established. The mandate of the committee was to provide input and advice on the elements to be included in the format, review the drafts and comments from other organizations or key

individuals on the recommended format, and make recommendations on how the completed format should be updated and maintained.

The final draft of MRDF/MARC format has been designed to incorporate multiple levels of information. The data elements included in the format reflect a broad interpretation of informational needs beyond the traditional catalog record. This concept is outlined by Maruyama in the introduction to the *Machine-Readable Data Files: A MARC Format*:

> The MRDF format has been designed to accommodate the data elements specified in the second edition of the *Anglo-American Cataloguing Rules* (AACR2), but the data elements included in the format have not been limited to those described in AACR2. Also, the explicit identification (or content designation) of these elements has been designed to accommodate a variety of products, e.g., a data inventory...,[a sales catalog,] a union catalog, in addition to a catalog record [in the form of a printed card].[15]

ICPSR's Automated Cataloging System for MRDF

Also in 1979, the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan received a grant from the National Endowment for the Humanities (NEH) to create a multipurpose automated cataloging system for its current holdings. The entire project is intended to be a pilot project which will provide a full-scale test of the cataloging rules for MRDF as prescribed in chapter nine of AACR2, identifying any needed modifications and revisions both in the cataloging code and in the newly formulated MRDF/MARC format. The ICPSR system will implement many of the data elements provided in the format and be operational in an interactive mode via the Michigan Terminal System (MTS) version of the Stanford Public Information Retrieval System (SPIRES). The automated cataloging system will act as a resource data base for information on thousands of available data files relevant to the social sciences. Included among the products that will be available from the system are detailed data abstracts. These data abstracts will be compiled in the consortium's annual *Guide to Resources and Services*.

Cataloging-in-Source

Another important information service to be derived from the ICPSR system is "cataloging-in-source" (also known as "cataloging-during-production"). Although modeled after the Cataloging in Publication (CIP) scheme, this MRDF effort operates outside the jurisdictional directives of the Library of Congress. The importance of

the scheme is that it allows major data producers to provide cataloging information at the early stages of a file's development. The cataloging takes place either at the site of the MRDF production by an in-house librarian or by a professional cataloger at another location. The final results are printed on the verso of the title page of the file's documentation.

The first implementation of "cataloging-in-source" was carried out in 1978 by the National Opinion Research Center (NORC) under the direction of Patrick Bova, NORC's data librarian. The NORC effort included not only the proper cataloging for the file itself, but the cataloging for the file's printed documentation as well. Also included on the verso was the proper bibliographic citation for the file, to which a user may refer when citing the MRDF in the literature.

Since 1978, two other major data producers have implemented the "cataloging-in-source" scheme—the Bureau of the Census and ICPSR. An example of ICPSR's cataloging copy on the verso of the title page of the file's documentation is given in figure 1. The MRDF "catalog-in-source" is important because such information is more likely to be accurate, and it assures that the original issue of an MRDF is cataloged.

Informational Needs of MRDF Users

The design of any information system should take into consideration the needs of its potential users. Improved access to information on the existence and availability of MRDF has been at the forefront of all efforts to bring bibliographic control to MRDF. The value of cataloging is ultimately proved not by how well each MRDF is uniquely defined, but by how efficiently the user is directed to the resource he needs. What follows is an examination of the informational needs of social science users and the resulting data elements that have been included in the MRDF/MARC format.

At the Conference on Cataloging and Information Services for Machine-Readable Data Files held at Airlie House, Warrenton, Virginia, in March 1978, a special session was devoted to MRDF user needs with respect to the creation of a national information system for computerized files and programs. The session was a combination of creating "wish lists" and reacting to already existing "catalogs" for MRDF. While the user input group agreed that a more descriptive emphasis than is usually inherent in a traditional catalog entry was essential to users, there was substantial difference of opinion as to how extensive the

CATALOGING-DURING-PRODUCTION

Machine-readable data file plus codebook

Janda, Kenneth
Comparative political parties data, 1950-1962 [machine-readable data file] / principal investigator, Kenneth Janda ; [generated as part of] the International Comparative Political Parties Project, Northwestern University. - ICPSR ed. - Ann Arbor, Mich. : Inter-university Consortium for Political and Social Research, 1979.
1 data file (158 logical records) + codebook (268 p.)
Summary: Data on the characteristics of 158 political parties operating in 53 nations from 1950 through 1962.
1. Political parties. I. Title.

Printed codebook only

Janda, Kenneth
Comparative political parties data, 1950-1962 / principal investigator, Kenneth Janda ; [generated as part of]the International Comparative Political Parties Project, Northwestern University. - Ann Arbor, Mich. : Inter-university Consortium for Political and Social Research, 1979.
268 p., 23 cm.
This codebook is to be used in conjunction with the machine-readable data file by the same title.
1. Political parties. I. Title.

BIBLIOGRAPHIC CITATION

All manuscripts using this data file and/or codebook should contain the following citation:

Janda, Kenneth. Comparative Political Parties Data, 1950-1962 [machine-readable data file]. Principal investigator, Kenneth Janda; [generated as part of] the International Comparative Political Parties Project, Northwestern University. Ann Arbor, Mich.: Inter-university Consortium for Political and Social Research. 1 data file (158 logical records) and codebook (268 p.).

ISBN 0-89138-966-0 Library of Congress Catalog Card Number 79-90467

Fig. 1. Example of Cataloging-During-Production for MRDF

descriptive information in a national information system should be.[16] One contributor to this conference explained:

> Users feel strongly that there must be an emphasis on the data summary or descriptive phase of the presentation of information on data files. The most important section of the cataloging card is the abstract. The abstract should be as detailed as possible and give as much information on the machine-readable data file as is consistent with the limits of the catalog entry. As users we would be content with many fewer details than are suggested by the AACRII cataloging rules and would prefer more extended comment in the files themselves....A simple and related point is the need to have some identification of the genesis of the file and its history. In this way it would be possible to link slightly modified files and to recognize when a data set is similar to one that is already in hand. Additionally, some key-word structure would yield a great deal of information on the data files themselves. ...Finally, we believe it is useful and important to link data files and software, in those situations in which a particular software has been created to manage and/or operate with a particular file.[17]

Another contributor expressed user's needs in this way:

> User requirements center upon data element retrieval through definitive data file description and data base documentation. Comprehensively, this means users require:
> Knowledge of the existence of data.
> Knowledge of the source of data.
> Knowledge of the applicability of data to solving specific problems or analytic needs.

These expressions of user's needs and many others were taken into consideration when formulating the MRDF/MARC format. Examination of *Machine-Readable Data Files: A MARC Format* will indicate data elements beyond those required for describing a monograph or a serial and include those needed to depict the special characteristics of this medium and the particular needs of MRDF users.

Conceptually, data elements for MRDF can be broken down into at least six levels: (1) those needed to *identify* MRDF (e.g., bibliographic elements); (2) those needed to *describe* the contents of MRDF (e.g., descriptive summary, data abstract, or in-depth subject analysis per item or variable); (3) those needed to *classify* MRDF (e.g., appropriate classification codes, indexing or subject descriptors necessary to group like data files together); (4) those needed to *access* MRDF (e.g., physical characteristics such as recording density and computer/software compatibility); (5) those needed to *analyze* or *use* MRDF (e.g., citation of documentation and related reports, how/when the data were collected, unit of analysis, sampling procedures); and (6) those needed to *archive*

or *maintain* MRDF (e.g., in-house records pertaining to the processing, storage and use of the data file).

Selected data elements for these six levels as represented in the MRDF/MARC format are given below:[19]

Level 1: Bibliographic Identity
—corporate or personal author (e.g., principal investigator, program director, etc.)
—title, subtitle, and other title information (i.e., statement of responsibility)
—general material designation (i.e., machine-readable data file)
—edition, plus appropriate statements of responsibility relating to edition
—production statement, including place, organization and date of production
—distributor statement (if appropriate), including place, organization and date of distribution
—size of file (including number of files, number of logical records, and statement indicating the presence of accompanying documentation
—series statement (title and numbering within series, if appropriate)
—notes
—unique identification numbers

Level 2: Data Abstract
—unique identification number
—type of file (numeric, text, computer programs, etc.)
—bibliographic citation of MRDF
—methodology (universe, sampling, unit of analysis, etc.)
—geographic coverage
—time period (chronological coverage of MRDF)
—date(s) of data collection (if unique from other dates)
—summary (subject matter description)
—derived source of data (if derived from printed sources or other MRDF)
—file size (number of observations, cases, variables, and any special file characteristics)
—bibliographic citation of accompanying documentation
—primary publications based on the use of the MRDF
—terms of availability
—contact person

Level 3: Classification
—Library of Congress Classification Number
—Dewey Decimal Classification Code

—LC Geographic Classification Code
—Subject category codes (applied locally)
—Descriptors or index terms (applied locally)
—Geographic headings (applied locally)

Level 4: Access of MRDF (technical information)
—mode of access
—type of data carrier or storage medium
—recording density, blocking factors, etc.
—computer compatibility
—software compatibility
—peripheral requirements
—special formats or system files

Level 5: Analysis or Use of MRDF
—file structure/sort sequence
—condition of data
—restrictions on use
—intended audience or level of expertise
—applications of the file or program
—linkage with other files or programs
—unit of analysis
—sampling procedures
—citation and location of documentation

Level 6: Archiving or Maintaining MRDF
—archival study number
—personal or organizational donor of MRDF
—date received
—date processed and entered into collection
—retention status (if temporary)
—access code (publicly available, restricted, etc.)
—cost for file duplication/dissemination
—frequency of updates or additions
—holdings note (for serials or serial-like MRDF)
—processing history (changes, revisions, modifications, etc.)
—documentaion number or shelf location

These data elements are not meant to be all inclusive, rather they are provided here to demonstrate the feasibility of an integrated approach to MRDF descriptive information.

Local applications of the MRDF/MARC format include the generation of several distinct products from one record, including a bibliographic citation, catalog entry, and data abstract. Examples of these

products along with associated content designations, are given in the appendixes to this paper. (At the time of this writing, no bibliographic utility had incorporated the MRDF/MARC format into its system, but two local applications of the format have been realized—at the Social Science Data Library, University of North Carolina and at the Inter-University Consortium for Political and Social Research, University of Michigan.)

Where Do We Go From Here?

There is no doubt that machine-readable data will play an even greater role in research and development programs of the future. More and more data needed for government and private research will appear in computerized form. Researchers and scholars should not have to spend additional time and dollars locating and acquiring appropriate MRDF. This is a function that can best be provided by a bibliographic utility. Such a utility already has the expertise in online network access and the data-base management programs for MARC-formatted files to offer the following products and services.

Shared cataloging. The machine-readable version of the 1970 censuses has undoubtedly been cataloged and described hundreds of times at as many libraries and data centers. The process of shared cataloging reduces this work to a one-time effort. Participants in a bibliographic utility system could benefit from the work performed by others.

Authority control. Access to authorized forms of author, uniform titles, author/title series and title subject headings used in bibliographic records would be provided by the utility. The primary purpose of an authority file is to accomplish the collocation function of the catalog, that is, to enable the catalog to relate and display together works by the same author, on the same subject, and in various editions regardless of the media.

Acquisition system. A bibliographic data base maintained by a utility can serve many purposes, including providing sufficient information for ordering available MRDF. Centralized access to such information would greatly reduce the time and effort required to locate and purchase MRDF needed by researchers. Some utilities even provide recordkeeping services related to the order process, including accounting functions.

Private file creation. Utilities may offer each participant the capability to create his or her own file of copied (derived) and original records. Such a file cannot be altered in any way by other participants.

This would allow participants to have interactive access to files representing their own unique holdings and associated local recordkeeping.

Union list. With several libraries and data centers contributing cataloging information on their uniquely held data files and programs, a union list of MRDF could be established. The union list would operate as a centralized inventory of data resources (who has what and where) across the country. Participants compiling the union list would be registering new acquisitions on an ongoing basis, thus providing a constantly updated and comprehensive list of unique MRDF, including a list of libraries and agencies which maintain these files.

Products. Derived products from an MRDF bibliographic data base maintained by a utility might include catalog records, book catalogs (with multiple indexes), data abstracts, distributor lists, orders in process, new acquisitions lists, union lists, special subject bibliographies, and local inventories. Most products can be provided in varying formats, including printed form, microform and machine-readable.

With the cataloging code (AACR2) available for MRDF, with the appearance of a working manual to help catalogers interpret these rules, and with the data elements and content designators defined in the MRDF/MARC format, the "blueprint" is at last in place for the next step—the integration of MRDF records into any of the existing bibliographic utilities. The benefits of utilizing existing bibliographic utilities to provide information on available MRDF is evident. However, other problems must be addressed before such a step can be implemented.

Problems Related to This Effort

Compiling an Expanded Record for MRDF. The question has been raised whether catalogers of MRDF can be persuaded to provide information beyond the briefest bibliographic record—especially since the number of characters required to compile and expanded bibliographic record for MRDF has been approximated at 2500 characters. There are several reasons offered here as to why catalogers should be persuaded to create such a record. First, catalogers of printed materials must deal with a large volume of works, and there is usually a backlog of works to be cataloged. By comparison, the volume of MRDF to be cataloged will be low. With a low volume of input, the cataloger should theoretically have more time to compile an expanded record. Without such a record, the identified needs of MRDF users will not be met. Second, catalogers of printed materials normally are not required to look beyond the title

page for cataloging information. However, catalogers of MRDF will of necessity be required to examine documentation beyond the title page to extract information not only to identify a particular MRDF but also to provide information on its nature and use. It is predicted that catalogers will find it necessary to provide more information in the note area of the bibliographic entry than they would for other media. The result will be that much of the information that goes into compiling an MRDF catalog entry may also be used to prepare a data abstract. In practical terms, there appears to be no reason why these same data elements should have to be compiled twice. Third, with automated bibliographic systems, we are no longer bound to the three-by-five card mentality, nor to the concept of meeting only one informational need. By thinking in terms of multiple applications of one system, the shared benefits go up and the cost of duplication goes down. The intended design of the MRDF bibliographic record as described here is to serve as an "organic record" from which several products can be derived without duplication of effort.

Lack of expertise. Before there can be widespread cataloging of MRDF, participating catalogers must be given the opportunity to become more familiar with this technical medium. Workshops and training sessions must be developed and offered as needed.

Lack of professionalization. At the present time there is no professional group within the library profession to speak to the needs of MRDF catalogers nor to be a vocal group for changes related to the cataloging code or the MRDF/MARC format. IASSIST (through its Classification Action Group) is the only visible group currently addressing these needs, but it has a limited voice in the library world. An organization similar to the International Association of Music Libraries should be organized for data librarians. This librarian group could also represent the needs of the user, publicize problems and promote sharing.

Inspired participation. Several research libraries are already cataloging MRDF (including Yale, Princeton, UCLA, and the University of British Columbia), but other libraries and centers maintaining MRDF must be persuaded to participate in the effort to compile bibliographic records for their unique holdings. Such participation is crucial to the goal of a union list for MRDF. Also crucial to the effort for social science researchers is the participation of the federal data collecting agencies and the support of the Office of Federal Statistical Policy and Standards (OFSPS).

Conclusion

The cost of computers and communications technology is declining steadily. With smaller computers and better software becoming more readily available, and with scholars familiarizing themselves with these tools and their applications, a new dimension to the information explosion is now apparent; and with it comes an increasing demand for access to more and better-documented data files. Before a data file can have value, however, it must first be communicated to the potential user. Communicating the availability of usable data is an inseparable part of research and an integral part of librarianship. In the near future, libraries will have no choice but to become more involved with computerized files and programs. The nature of this involvement might well depend on demonstrated need, creative planning and available resources; and while it is not yet feasible to expect libraries to provide a full range of services related to MRDF, they are prepared to provide better access to information on the availability of data files. Taeuber sums it up this way: "While it is difficult to single out one function which is more important than any other, if libraries participated in the data revolution in no other way, preparation of a union list of data resources would be a major contribution to research. This could be a first step in increased library participation while training for the technical functions proceeds."[20]

ACKNOWLEDGMENT

The author would like to thank Ann S. Gray, research assistant in the Social Science Data Library at the University of North Carolina, for her work in preparing the examples contained in this paper.

Appendix A

MARC-Formatted Bibliographic Record for an MRDF

Appendix information taken from the *Directory of Data Files* prepared by the U.S. Bureau of the Census, Washington, D.C.

did *****	C-40
stn	14-01-FOURTH-USCEN-70
til#ac	Census of population and housing, 1970 housing summary statistic file 4#conducted by the Bureau of the Census
edn	DUALabs ed.
pro	Arlington, Va.#Data Use and Access Laboratories (DUALabs) #1972
col#ae	3 data files (ca. 210000, 228000, 90000 logical records) #1 codebook
noq	This is a series of summary statistic files each containing detailed housing characteristics by geographic area based on the 1970 census sample questionnaires. Each file contains records which correspond to an individual geographic area.
noq/2	Also known as: Fourth Count Housing Summary, 1970 Census of Population and Housing
tof	Numeric (Summary statistics)
nos	This file, known as the "fourth count" has three individual files, each containing identical subject matter. Some of the variables included in the tables are total number of housing units, year structure built, tenure, race of head, gross rent or value of unit, persons per room, heating, air conditioning equipment, and presence of the following: clothes washer, clothes dryer, dishwasher, food freezer, television set, and battery operated radio. Separate tables with similar information are provided for the Spanish population.
tim	Data contained in the files pertain to the date of the census, April 1, 1970, except for selected items which relate to historical periods.
for	Data is 'compressed'. Use DUALabs' MOD-series program to process. Use DUALabs' DDLIST program to produce a listing of DUALabs' Data Descriptor List
not	Title from: Directory of Data Files prepared by the Bureau of the Census
chn	14-13
acc	I
uni	The universe consists of all housing units. The data are based on 5-, 15-, and 20-percent samples.

Appendix A—*Continued*

sof	Three files; File A contains approximately 210,000 logical records representing about 35,000 tracts for the U.S. File B contains approximately 228,000 logical records representing 37,500 MCD's for the U.S. File C contains approximately 90,000 logical records representing about 13,000 summary areas for the U.S.
qec	The three housing files in fourth count have different levels of geography. File A presents housing summary statistics for all census tracts. File B presents housing summary statistics by minor civil divisions (or census county divisions). File C presents housing summary statistics for States, counties, places of 2,500 or more, SMSA's and component parts of SMSAs, urban/rural non-farm, and rural farm components.
ref#bef	1970 Census of Population and Housing Fourth Count Housing Summary Tape (Sample)#Arlington, Va. : Data Use and Access Laboratories (DUALabs), 1972#This is DUALabs' version of the Bureau of the Census' documentation and technical guide for this file.
ref/2#abe	United States. Bureau of the Census#1970 Census Users' Guide, Part I and II#Washington : U.S. Government Printing Office, 1970
ref/3#abe	United States. Bureau of the Census#Data Access Descriptions No. 22, Fourth Count Summary Tapes from the 1970 Census of Population and Housing, March 1971#Washington : U.S. Bureau of the Census.
sul#ayx	Housing#-- United States#-- Statistics
suh	Census#-- Population and Housing Data
aec	United States.#Bureau of the Census
aec/2	Data Use and Access Laboratories (DUALabs), Arlington, Va.
tie	Fourth count housing summary, 1970 census of population and housing
doe	81/04/09#ag
ced	d-636 - d-638

Appendix B

Catalog Entry Derived from Bibliographic Record for MRDF

Census of population and housing, 1970 housing summary statistic file 4 [machine-readable data file] / conducted by the Bureau of the Census. -- DUALabs ed. -- Arlington, Va. : Data Use and Access Laboratories (DUALabs), 1972.
3 data files (ca. 210000, 228000, 90000 logical records) + 1 codebook.

This is a series of summary statistic files each containing detailed housing characteristics by geographic area based on the 1970 census sample questionnaires. Each file contains records which correspond to an individual geographic area.
Title from: Directory of Data Files prepared by the Bureau of the Census.
Also known as: Fourth Count Housing Summary, 1970 Census of Population and Housing.
Data is 'compressed'. Use DUALabs' MOD-series program to process. Use DUALabs' DDLIST program to produce a listing of DUALabs' Data Descriptor List.
Primary reference: 1970 Census of Population and Housing Fourth Count Housing Summary Tape (Sample) -- Arlington, Va. : Data Use and Access Laboratories (DUALabs), 1972.
This is DUALabs' version of the Bureau of the Census' documentation and technical guide for this file.
Geographic coverage: The three housing files in fourth count have different levels of geography. File A presents housing summary statistics for all census tracts. File B presents housing summary statistics by minor civil divisions (or census county divisions). File C presents housing summary statistics for States, counties, places of 2,500 or more, SMSA's and component parts of SMSAs, urban/rural non-farm, and rural farm components.
Data contained in the files pertain to the date of the census, April 1, 1970, except for selected items which relate to historical periods.
Summary: This file, known as the "fourth count" has three individual files, each containing identical subject matter. Some of the variables included in the tables are total number of housing units, year structure built, tenure, race of head, gross rent or value of unit, persons per room, heating, air conditioning equipment, and presence of the following: clothes washer, clothes dryer, dishwasher, food freezer, television set, and battery operated radio. Separate tables with similar information are provided for the Spanish population.

1. Housing -- United States -- Statistics. I. United States. Bureau of the Census. II. Data Use and Access Laboratories (DUALabs), Arlington, Va. III. Title: Fourth count housing summary, 1970 census of population and housing. Did

Appendix C

Data Abstract Derived from Bibliographic Record for MRDF

MRDF ABSTRACT

ID Number: DID-C-40

Type of File: Numeric (Summary statistics)

Citation: Census of population and housing, 1970 housing summary statistic file 4 [machine-readable data file] conducted by the Bureau of the Census. DUALabs ed. Arlington, Va : Data Use and Access Laboratories (DUALabs), 1972.

Universe: The universe consists of all housing units. The data are based on 5-, 15-, and 20-percent samples.

Geographic Coverage: The three housing files in fourth count have different levels of geography. File A presents housing summary statistics for all census tracts. File B presents housing summary statistics by minor civil divisions (or census county divisions). File C presents housing summary statistics for States, counties, places of 2,500 or more, SMSA's and component parts of SMSAs, urban/rural non-farm, and rural farm components.

Time period: Data contained in the files pertain to the date of the census, April 1, 1970, except for selected items which relate to historical periods.

Summary: This file, known as the "fourth count" has three individual files, each containing identical subject matter. Some of the variables included in the tables are total number of housing units, year structure built, tenure, race of head, gross rent or value of unit, persons per room, heating, air conditioning equipment, and presence of the following: clothes washer, clothes dryer, dishwasher, food freezer, television set, and battery operated radio. Separate tables with similar information are provided for the Spanish population.

Appendix C—*Continued*

File Size: Three files; File A contains approximately 210,000 logical records representing about 35,000 tracts for the U.S. File B contains approximately 228,000 logical records representing 37,500 MCD's for the U.S. File C contains approximately 90,000 logical records representing about 13,000 summary areas for the U.S..

Special Formats: Data is 'compressed'. Use DUALabs' MOD-series program to process. Use DUALabs' DDLIST program to produce a listing of DUALabs' Data Descriptor list.

Primary Reference: 1970 Census of Population and Housing Fourth Count Housing Summary Tape (Sample). Arlington, Va. : Data Use and Access Laboratories (DUALabs), 1972.

Contact Person: Judith Poole, Research Consultant, North Carolina State Data Center, Institute for Research in Social Science, Manning Hall026A, University of North Carolina, Chapel Hill, N. C. 27514 (919) 966-3346.

References

1. American Library Association. *Anglo-American Cataloging Rules,* 2d ed. Chicago: ALA, 1978, pp. 201-16.

2. Sessions, Vivian S. "Primary and Secondary Data Base Professionals: Time for a Reapproachment?" *Review of Public Data Use* 3(Jan. 1975):1-2.

3. Luedke, James A., Jr., et al. "Numeric Data Bases and Systems." In *Annual Review of Information Science and Technology,* vol. 12, edited by Martha E. Williams, pp. 119-81. White Plains, N.Y.: Knowledge Industry Publications, 1977.

4. Herman, Elizabeth, and Byrum, John, eds. "Final Report of the Catalog Code Revision Committee Subcommittee on Rules for Cataloging Machine-Readable Data Files." Mimeographed. Chicago: ALA, p. 19. (ED 119 727)

5. American Library Association, *AACR2,* p. 203.

6. Lucci, York, et al. *A Library Center of Survey Research Data.* New York: Columbia University School of Library Service, 1967. (Available on microfilm from Columbia University.); and Nasatir, David. *Data Archives for the Social Sciences: Purposes, Operations, and Problems* (Reports and Papers in the Social Sciences, no. 26). Paris: Unesco, 1973, pp. 10-14.

7. Byrum, John D., Jr. "Toward a Standard Bibliographic Description for Machine-Readable Statistical Data Sets." Paper presented at the Workshop on Documentation of Large Machine-Readable Statistical Data Sets, 19 April 1974, New York University, New York.

8. Herman and Byrum, "Final Report of Catalog Code."

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

10. Rowe, Judith S., and Ryan, Mary. "Library Service from Numerical Data Bases: The 1970 Census as a Paradigm." *College & Research Libraries* 35(Jan. 1974):7-15.

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

12. Data Use and Access Laboratories. *Report on the Conference on Cataloging and Information Services for Machine-Readable Data Files.* Arlington, Va.: DUALabs, 1978, pp. 204-05.

13. U.S. Bureau of the Census. *Directory of Data Files.* Washington, D.C.: U.S. Bureau of the Census, 1979.

14. U.S. Office of Federal Statistical Policy and Standards. *Technical Paper No. 3: Procedures for Preparation of Abstracts of Public Use.* Washington, D.C.: USGPO, 1980.

15. Network Development Office. *Machine-Readable Data Files: A MARC Format,* final draft. Washington, D.C.: Library of Congress, 1981, p. 4.

16. DUALabs, *Report on the Conference,* pp. 189-203.

17. Clark, William A.V. "Some Comments on the Role of Users in Accessing Machine-Readable Data Files." In *DUALabs, Report on the Conference* pp. 201-03.

18. Loebl, Andrew S. "User Needs and Requirements Related to Cataloging and Information Services for Machine-Readable Numeric Data Files." In DUALabs, *Report on the Conference,* 1978, pp. 42-43.

19. These revised elements were first introduced *in* Sue A. Dodd and Judith S. Rowe. "A Model Bibliographic Information System for Machine-Readable Data Files in the Humanities and the Social Sciences." In *Data Bases in the Humanities and Social Sciences,* edited by J. Raben and G. Marks, pp. 275-79. Amsterdam: North Holland Publishing, 1980.

20. Taeuber, Cynthia M. "The Role of the Librarian in Organizing Social Science Machine-Readable Data." In *Proceedings of the Eighth Annual Conference Association for Population/Family Planning Libraries and Information Centers,* edited by Judith M. Wilkinson, p. 82. Washington, D.C.: Association for Population/Family Planning Libraries and Information Centers-International,1975.

This Page Intentionally Left Blank