IASSIST Newsletter, Vol. 2, No. 2 (Spring 1978)

BUILDING A BIBLIOGRAPHIC/MARC DATA BASE FOR SOCIAL SCIENCE DATA FILES IN A NETWORK ENVIRONMENT

by
Sue A. Dodd
University of North Carolina

INTRODUCTION

Social science numerical and textual data files represent a vast amount of valuable and publicly-available information. For example they are widely used by students, faculty, and policy makers engaged in research. Not only have such data files had an unprecedented growth in the last decade, but with the advance of small and relatively inexpensive computer terminals, data analysis and computer simulation models have moved into the classroom as legitimate instructional tools. Specialized files, often referred to as "educational data packages," have been developed to teach students analytical skills, so as to better understand social and economic phenomena. According to Nesvold (1976): "Experience with machine-readable 'laboratory' materials should be as appropriate to the beginning social science student as is the laboratory for the beginning chemistry student."

Unfortunately, many such data resources are not fully utilized because potential users are unaware of the existence and accessibility of social science data files. At the present time, information on usable MRDF is fragmented among varying government agencies, research institutions, and university computing and data centers. Among these various agencies, there is no common format for information on the existence of data files, nor is there any standardized structure that would facilitate retrieval of information from many different sources. Existing information on computerized files is available to some but not to all. What is needed is a central source of information within the public domain that would provide equal access to all interested users. What is needed is some form of bibliographic control and national standards for social science files--not unlike that which is available for printed materials.

DEVELOPMENTAL EFFORTS

The Social Science Data Library of the Institute for Research in Social Science at the University of North Carolina at Chapel Hill is currently engaged in some developmental work to create a bibliographic data base of machine-readable data files that would be available to users within the Triangle Universities Computation Center (TUCC) area. This centralized bibliographic data base would be designed to serve multiple purposes and would be converted to an on-line interactive mode to be accessed within the TUCC network environment. There would be at least three types of potential users of this data base: 1) the academic user who is looking for potential sources of data for scholarly research; 2) user service personnel who act as information brokers to other end-users within the network system; and 3) libraries and data centers who would use the system as a reference tool for clients, and as a records management system for their own data holdings.

The Logical Stucture

The "logical structure" of the bibliographic data base would include three related informational levels. At the highest level would be those bibliographic elements required to uniquely identify the data file. For example: study number, title, author, director or principal investigator, edition, place of production, data producer, date of production, place and name of distributor, abstract, size of file, subject descriptors or headings, and series identification. This would be called the universal level of information, since it would be compatible with existing international standards for bibliographic references and cataloging, plus it would allow for the integration of machine-readable data files into multi media collections. The next level of information would include bibliographic elements required to analyze the file. For example: sample description, number of data units, number of variables, time coverage, etc. This would be called the analytical level of information and would vary depending on the type of file (e.g., text, maps, physical sci-

ence, social science, administrative records, computer software programs, etc.). The lowest level of information would consist of those bibliographic elements which are variant in nature. For example: condition of data, file structure, physical characteristics, contact person, restrictions of use, etc. This might be called the local level of information, since it is dependent on local options, special applications of use, computer compatibility, etc.

Physical Structure

The "physical structure" of the data base would be the MARC II format. By way of explanation, MARC is an acronym derived from the phrase MAchine Readable Catalog. It is a "generic term referring to bibliographic information that has been encoded and transcribed into a machine-readable form to permit its manipulation. . . . ." (Wiesbrod, 1977) The MARC format constitutes the coding conventions under which MARC data may be organized. It was developed by the Library of Congress in the early 1960's and is very quickly becoming the international standard for bibliographic representation. The essential characteristic of MARC-formatted records is that they can accomodate a varying numberof "variable length" data items -- affording considerable generality of use.

The general purpose software programs to be utilized in the processing of the data are ones designed for use by the Carolina Population Center's Technical Information Service Library and collectively are called: Bibliographic/MARC Processing System (BPS). The internal processing capabilities of the BPS file primarily in the areas of information storage, retrieval, and report generation with additional programs allowing for automated thesaurus construction and interactive subject retrieval. Complementing this software is an on-line interactive retrieval program called TOBIAS (Terminal Oriented Bibliographic Informational Analysis System). TOBIAS was designed locally and has been extensively revised by members of the Institute for Research in Social Science programming staff. TOBIAS uses simple english language and appropriate commands, provides prompting and on-line tutorial instruction, incorporates set theory and boolean logic procedures, displays information on-line and prints information off-line which can be received at various destinations.

The current bibliographic data base includes a representative sample of the Social Science Data Library's total holdings numbering to 1500 separate files and including some 500 Harris national public opinion polls. Recently, however, the developmental work with the data base was extended to include other machine-readable data files which were available within the TUCC community.

The Network Community

The TUCC network community represents a wide and diverse set of users with varying degrees of sophistication in the areas of education, research, and commercial enterprises. TUCC may be described as a "star network" built around a central computer facility which is owned, operated and shared by North Carolina's three major universities -- University of North Carolina at Chapel Hill, Duke University, and North Carolina State University. Within this environment, there exists a Time Sharing Option (TSO) which allows a number of users to utilize the computer concurrently and in a conversational manner via a terminal (with telephone link-up) which may be remotely located from the system installation. TUCC maintains a small service staff, but most of the informational operations are carried out by the respective Computation Center's user services and by contact persons throughout the network system representing libraries, data centers, academic departments, state government agencies, and research organizations. Users outside the three major universities must purchase computer time from TUCC. Three very important "commercial" users of TUCC are North Carolina Educational Computing Service; Research Triangle Institute; and North Carolina Science and Technology Research Center.

The source of information for the TUCC data holdings is what is officially known as General Information Series Document No. GIR-045-3 "Data Files and Data Banks in the TUCC Community," but what is popularly known as the TUCC Directory. After some preliminary meetings, a cooperative effort was launched to automate the TUCC Directory which would then be processed by the BPS software. Also for the first time, additional information on these data files were collected to provide for cataloging information and

for better access points to the files including descriptors and index terms.

The cooperative effort included representatives from the following institutions: UNC at Chapel Hill, Duke University, North Carolina Educational Computing Center, Triangle University Computation Center, and the School of Library Science (UNC). Four graduate students from the UNC School of Library Science joined this effort as part of a practical learning experience under the direction of Professor Martin Dillon. All representatives provided input via remote terminals and the local QED (Quick Edit) program available through TSO. Information was transferred from the various off-line disks to a master on-line disk and was then processed by the BPS "update" program which created the MARC record data base. In the near future, it is expected that the TUCC bibliographic/MARC data base will be converted to TOBIAS as an on-line interactive file available to all users within the TUCC network community.

In a network environment like TUCC, in which there is no one centralized reference center, the contact personnel at key locations throughout the state become crucial operational and information brokers of the entire network system. Consequently, they would act as "information specialist" or as a buffer between the "end-user" and the system. Having on-line access to information on available data in the TUCC community would greatly enhance the level of service these staff can provide to their users. The automated system would also offer more flexibility and refinement in the retrieval of information pertaining to a particular user's need than hard copy versions of the same information.

Selected staff in libraries and data centers throughout the state would perform the same responsibilities mentioned above, but in addition they would use the Bibliographic/MARC data base for certain in-house purposes such as acquiring, maintaining, classifying, and reporting information on available data files. For example, data centers and libraries could use the data base for the following:

1. Record management -- including the inputting, editing, updating, and various listings and sort arrangements.

2. Cataloging of records -- including shared cataloging (reducing duplication of effort), verification of title, author, series, etc.

3. Classification of records -- including the application of descriptors, construction of a thesaurus, and implementation of various subject classification schemes such as the Library of Congress Subject Headings.

4. Acquisition -- including information on contact person, location of data and file documentation, restrictions (if any), cost, etc.

5. Report generation -- including the extraction from the master file of an inventory of their own local holdings; printed catalog records; authority lists of titles, authors, series, etc.; shelf list of file documentation; etc. -- all with a choice of print formats such as on-line display, magnetic tape, microfiche, and hard copy.

The relationship between on-line technology and library functions, including how on-line technology has contributed to significant changes in quality and amount of services provided by the library is well defined in a recent paper by Miriam A. Drake (1977) of Purdue University Libraries. Drake points out that in addition to on-line technology allowing libraries to communicate with each other, it has also had an "impact" on such library functions as "collections' building, order processing and accounting, cataloging and user service" (Drake, 1977).

At the present time, cataloging information on machine-readable data files is not currently available through any library network system. Consequently, there is a need for a bibliographic data base that would provide librarians with this type of information. The bibliographic/MARC data base can generate such catalog records in many different formats, including the computer generated perforated 3X5 card stock.

CONCLUSION

Success is invariably measured not by what you hope to achieve, but by what you are able to produce. At the same time, the sum total of the lessons learned and the refinements made in the initial stages of any new endeavour becomes the foundation for future successes. The work described here is still developmental but it is designed to be a prototype which could be expanded and implemented by other parties. Present limitations could be overcome with additional resources -- both personal and technological. However, we have moved closer to our objective which is to provide information on available machine-readable data files to a wider audience. Building a multi-purpose bibliographic/MARC data base for a network environment with on-line capabilities opens up information exchanges and data access that have not previously existed for social science data files. By building the data base according to existing international standards for bibliographic representation, the potential audience is extended to include any user of a library resource.

References

Drake, Miriam A. Impact on On-line Systems on Library Functions. Paper presented at the Pittsburgh Conference on the On-line Revolution in Libraries, Pittsburgh, P A, November 14-16, 1977.

Nesvold, Betty A. Instructional Applications of Data Archive Resources. American Behavioral Scientist. 19 (No. 4, March/April 1976), 445-67.

Weisbrod, David L. NUC Reporting and MARC Redistribution: Their Functional Confluence and it Implication for a Redefinition of the MARC Format. Journal of Library Automation. 10 (No. 3, Sept. 1977), 226-7.