Academic Libraries and Collection Development of Nonbibliographic Machine Readable Data Files

by Daniel C. Tsang[1]
Social Sciences Librarian/Bibliographer
University of California

It has been over two decades since Phil Converse (1964) and Ithiel de Sola Pool (1965) called on librarians to take the initiative in providing service for machine-readable data files (mrdfs) arguing that such materials logically belong in libraries. Since that clarion call, others have echoed their vision, as librarian Howard White (1974) discusses in his pathbreaking dissertation on social science datasets.

One of the pioneers that White documents, Ralph Bisco (1967, 17) in a speech delivered at the opening of the University of Florida's graduate research library, noted that two librarians (one of them Herbert Ahn, a colleague of mine, and then Systems Librarian at the University of California, Irvine) were members of a subcommittee of the newly established Council of Social Science Data Archives, a federation of data archives, that sought to improve access to social science datasets.

Another pioneer was Karl Pearson, whose 1968 University of California, Los Angeles library science thesis was described by White as "the fullest statement to date on libraries and numerical data sets" (White, 1974, 30-31). Yet another visionary was Jack Dennis, who envisioned the eventual assimilation of the data archive by the library after an "initial period of cooperation" (cited in White, 1974, 36; see also Adams and Dennis, 1970, 57-58).

Dennis, Linton Freeman and Robert Hayes were members of a Council on Social Science Data Archives committee that visited Northwestern University's Intersocietal Information Center in 1968, and subequently issued a report which recommended placing Northwestern's data archive in the university library as the "best place to have such a facility because of its central location, its interest in computer-oriented approaches, its knowledgeable personnel, its general policy of serving all people in the university, and the apparent availability of space

[1]Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Washington, D.C., May 26-29, 1988

necessary to house such a facility" (cited in White, 1974, note 65, 213). But that did not happen.

According to another observer, "[s]ince libraries did not regard such data files as within their collection and service parameters, data archives, to a large extent were operated independently of libraries; libraries often neither managed these archives nor referred their clients to them... By 1984, more libraries were accepting responsibilities for the collection, preservation, and dissemination of nonbibliographic machine-readable data bases as a legitimate activity" (Hernon, 1986, 341), although most of these efforts dealt with online databases.

In the last twenty years or so, the library world has devoted several special issues of scholarly journals to the topic of nonbibliographic data files (White, 1977; Claydon and Soergel, 1982; Heim, 1982). In IASSIST itself, more and more members are traditional librarians beginning to provide mrdf service. At this conference, there is even a workshop on integrating mrdfs into traditional library service.

Traditional libraries struggling with what to do with mrdfs have in the past been rather conservative in their move to integrate mrdfs into the library. For instance, until recently, most have avoided cataloging such materials, except perhaps codebooks.

There has been an absence of any overall policy that would state clearly the role mrdfs play in library collections.

In part, that is perhaps due to the traditional library's mrdf 'phobia', even as more and more libraries are overcoming 'computer-phobia'. For example, many libraries are actively acquiring CD-ROM products, or subscribing to online services, and patrons are enthusiastic about searching Infotrac or BRS After-Dark. But mrdfs — by which I mean data files used on mainframes for computerized statistical analysis — are still considered for the most part a bothersome format.

In this paper, I want to focus on the collection development implications of having mrdfs in a traditional library setting. I shall limit my definition of mrdfs to nonbibliographic files accessible through mainframes, and not deal with the acquisition or collection of CD-ROMs and the like.

Traditional library literature is not much help in this area; in Library Literature (February 1988 issue), under "Collection development," the researcher is referred to "Libraries — Book collections!" More diligent research will locate a few essays on or brief references to the topic, mostly on the process of how to locate data, i.e., data acquisition. Again, Bisco (1964), is a pioneer, writing twenty-four years ago about the "complex" process involved in acquiring new data. Robert Mitchell (1964, 90-91) also is an early essayist on the acquisition of Third World datasets. David Nasatir (1977), in a brilliant essay, expounds on the joys and tribulations of "Stalking the Wild Data Set" at home and abroad. He also provides the most extensive discussion on the "Data Acquisition Function," in a UNESCO report entitled Data Archives for the Social Sciences (1973).

In his Ph.D thesis White (1974), focuses on an analysis of purchase orders at a number of social science data suppliers, and concludes with a call for libraries to actively purchase codebooks, but not datasets.

Betty Yantis (1980) and John Nixon (1980) both write about the acquisition of data at the University of Nevada's Center for Business and Economics Research. Robbin (1977), writes about the "pre-acquisition process." Ray Jones (1982) describes the variety of governmental and academic datasets acquired at the University of Florida Libraries, and Pope (1984) details the committee process involved in the acquisition of new datasets at that institution. Ann Gerken

(1984, 9), then at the Cornell Institute for Social and Economic Research, notes the variety of sources used for data collection, and adds that "[d]etailed collection development policies are developed in collaboration with faculty, librarians, and members of the Institute." Then U.S. Archivist Bob Warner and Francis Blouin (1980) as well as Charles Dollar (1980) address the complex issue of appraising mrdfs. Chiang (1986, 68) reports that Cornell Library collects "both microcomputer and mainframe accessible data."

A data archive's implicit collection development guidelines may evolve into something more explicit. Laine Ruus (1982a, 399–400), then at the University of British Columbia Data Library (created as a joint library–computing facility venture), concedes that as is commonly the case, "the collections policy is rather vague and ad hoc. The original mandate made only one stipulation regarding collections -- that the Data Library 'develop collections...in accordance with the academic requirements of the University, in parallel with the policies of the Computing Centre and the Library'." Nonetheless, a "policy has evolved over the years" which can be summarized as follows: "the Data library will collect automatically all significant Canadian data files such as census data, election studies, and other major social surveys... All other MRDF are acquired on request, tempered by considered need, potential for furture use, and of course, budgetary constraints. In addition, the library will function as a data archives in the sense that an attempt is made to acquire any original MRDF produced by local researchers, or offered for deposit by outside researchers (depository MRDF) and every effort is made to ensure that these are maintained for prosterity."

Ruus (1982a, 400) has not found it necessary to limit datasets to particular disciplines, but she will not acquire datasets that cannot be made available to all academic users, or where individual privacy is violated. Nor will she maintain mrdfs that "lack adequate documentation or are so 'dirty' as to be useless for secondary analysis." UBC's Data Library's collection development policy is well delineated in the "Data Library Procedures Manual," which offers a section on the "Care and Feeding of the MRDF Collection," and includes within that an acquisitions policy (Ruus, 1982b; Henderson, 1988). Its policy may well be the most explicit of all such collections, even to the extent of considering previous use patterns as shown by tape mount statistics.

A 1984 survey by the Association of Research Libraries found that of 34 responding libraries, only four had drafted a collection development policy statement for machine–readable data bases (Association of Research Libraries, 1984, 3). The four are not further identified.

From my own informal, admittedly small sampling of a handful of data archives, it appears that most do not have written collection development policies. At Simon Fraser University, Walter Piovesan (1988) has "found no real need to actually formalize a policy... Not having a fixed budget makes it hard to develop a collection policy. If you have a budget, then you will need to 'prioritize." Ann Janda (1988) at Northwestern University data archive, which is operated by the Academic Computing and Network Services rather than the library, also does not have a "formal collection development policy for MRDF... at least not yet," with orders "need and demand driven." According to Janda, "This has worked fine as we have been working largely in the realm of ICPSR data requests." For users requesting non–ICPSR data, she does invoke certain principles: The data requested must be of a "fairly general nature" and is potentially going to be used by more than one project or user; and the user should share the cost of the purchase." At Yale, JoAnn Dionne (1988) reports that the Social Science Data Archive "doesn't have a formal collection policy for mrdfs." She generally "buys only when a user

requests a data set and only then only within certain dollar amounts -- but that depends on anticipated use." She will not spend more than $200 unless more than one person will use the data.

A 1983 report by a task force on nonbibliographic databases at the State University of New York, Albany calls for drafting a collection development policy that at first "should be limited to acquiring data sets on demand. The collection should also be limited to locally developed and/or locally used data" (State University of New York, 1983, 22).

In the University of California library system, work has begun on drafting local mrdf collection development policy statements. A draft policy, "Collection Policy Governing Machine Readable Data Files," dated November 11, 1986, for the Berkeley library, applies to all mrdfs, including software. In its present version, which may not survive in final form, it recommends substituting mrdfs for printed information only "with extreme caution," given the volatility of the information industry, the limited number of simultaneous users, and the need for staff assistance. It also cautions against acquiring mrdfs purely as a depository function, and urges that all collection be evaluated against other potential acquisitions and weighed against other uses of book monies.

At University of California, Davis, library staff have been discussing recommendations of a committee on numeric and textual databases; a proposal that "collecting responsibility for all formats belongs to all selectors" received "wide support," as did a recommendation that the position of a coordinator for machine-readable resources be created. At the present time, "programming expertise sufficient to access large datafiles on mainframe computers is not required, although the candidate may need to develop this ability if the need arises" (University of California, Davis, Library, 1987). At University of California, San Diego, Jim Jacobs (1986), also has drafted a statement on collection development, describing the collection there as "being built passively in response to requests for machine readable data from faculty. There is no active program for acquiring data in anticipation of possible future needs."

The Research Libraries Group, comprising the nation's top research collections, has also begun working on a mrdf collection management study (Jones, 1988).

I suspect a major reason for the flurry of activity among collection development librarians is the proliferation of CD-ROM products. A policy is needed before libraries become inundated with new technological products. I suspect mrdfs on tape are the least of most librarians' worries.

Nonetheless, before proceeding, it may be useful to pause and reflect on why a collection development policy would be productive. After all, most data archives appear to have survived without such written policies!

Several reasons come to mind. First, more and more libraries are being run like corporations, and hence, wanting a written policy for every procedure may just be the application of good management pracatices.

Also, we must realize that our individual tenure as data archivists or data librarians -- whatever we may hope -- is finite. At a recent meeting on mrdfs with the technical services librarians in my library, the Head of Acquisitions turned to me and said plaintively, "Dan, you could get run over by a car tomorrow!" In other words, what happens when I am gone? Will my replacement be able to figure out what was done? To be sure, through our daily work, every good librarian becomes a repository of obscure facts and important information. That is unalterable. Not everything can be written down, or passed on easily to the next generation! But a collection development policy

would be one way in which to clarify on paper what past practice has been, and what future practice should be.

Another major reason is so that the collection can develop in an orderly manner, and not be subject entirely to the whims of a particular bibliographer or researcher.

A well-written policy on collecting mrdfs might also protect the collection from any arbitrary change; at least, one could point to the policy to try to forestall any attempt to get rid of the collection!

In addition, such a written policy would be useful in training or orientating new staff in the library, as well as an aid in publicizing or explaining the collection, to users and potential donors.

At Irvine, the librarians who do the collecting are called bibliographers; at other libraries, they could be called "selectors." One immediate question we are confronting at Irvine as we merge mrdfs into the collection, is whose responsibility it is to select mrdfs? In theory, our bibliographers are responsible for all formats of materials. On paper this looks good; but in practice, this has primarily and almost exclusively meant traditional library formats such as books, periodicals, microform, video or audio tapes or films. As the Official Representative for ICPSR and the Social Sciences Bibliographer, I have become the *de facto* mrdf selector. The working solution we have implemented is that all mrdf selections will be passed through me just to see they are compatible with the hardware (or software) researchers use. But thus far, no one else has placed any orders.

To let all bibliographers select mrdfs may sound like heresy to a data archivist; but I would argue that if we are to integrate mrdfs into traditional library services, we must avoid stereotyping mrdfs as some weird format, and thereby perpetuating the segregation by format.

If the problem is lack of awareness or familiarity with mrdfs, then that surely can be remedied. Just as reference librarians are retooling for database searching, I believe that bibliographers can be educated about mrdfs. Instead of seeing this as an attack on one's turf, one might rather see this as an opportunity for others to contribute their expertise. For bibliographers are subject specialists who are responsible for working with faculty, and thus should be aware of the research needs on the campus within a particular discipline. There will soon be no way that one person can know or attempt to meet all the mrdf needs of all the disciplines on a campus.

In traditional libraries, collection policies for books and serial titles at major academic libraries generally are divided by subject and within each subject, by level of collecting. For example, at the University of California, Irvine (1983, 25-26), the levels are comprehensive ("all significant works" within a defined field), research (supporting doctoral and post-doctoral work), advanced study or beginning research (graduate and advanced undergraduate work), teaching or initial study (undergraduate curriculum), and the lowest level, basic information (non-curriculum-related).

Sections on mrdfs, then, could well be included within the individual subject chapters of a collection development manual, where mrdfs are an important part of a research or instruction program on campus. That, I believe, would be a long-term goal as mrdfs become futher integrated into traditional library services.

But it still would be helpful for the library to have an overall collection development policy on mrdfs, if only because of the processing and service implications any acquisitions entail.

Having a written collection policy does not mean it is engraved in stone. A policy must be

flexible and open to revision (Robbins, 1977, 25). No one policy can be written for all data collections. Local conditions will dictate what is important for that collection (Bernard and Jones, 1984, 98).

With that in mind, I would just like to focus on some important elements I believe such a policy should contain. Many of the ideas or categories are taken from a report on "Textual and Numeric Data Files" written by an ad hoc committee of the Librarians Association of the University of California (1983, espcially 13–14).

I have also garnered some ideas from an amazing book of abstracts compiled by the staff at the Correlates of War Project at the University of Michigan. Beyond Conjecture in International Politics (Jones and Singer, 1972), is a collection of abtracts of data-based research, systematically analyzed by a set of categories that are useful as we develop a collection development policy. Finally, others come from essays already cited.

Some important elements of a mrdf collection development policy for an academic library are:

1. Selection responsiblity. Who has the responsibility; all bibliographers? Or just the mrdf bibliographer?

2. Budget source: Who pays?

3. Level of collection activity (see above).

4. Subject Scope: Is the subject of relevance to research and instruction at the university?

5. Temporal domain: Is the time period covered of relevance to research and instruction at the university?

6. Spatial domain: Does the mrdf cover a region or location that is of relevance to research and instruction at the university? For example, Is the data collection a regional depository?

7. User need: Does the user need to manipulate data or just use manipulated data?

8. Uniqueness of data: Are the data available in print format? Is it necessary to get it in mrdf format? Are they only available in mrdf format?

9. Currency of data: Are the data from an ongoing study that will be quickly superseded by more recent revisions? Is it important to acquire quarterly updates or just annual cumulations?

10. Confidentiality of data: Is there a need to restrict personal or proprietary information in the dataset? Will acquisition violate privacy?

11. Physical format: Is the medium compatible with available hardware?

12. Software compatibility: Are the data accessible by software currently available? Or are they software dependent?

13. Documentation: Are the data supported by adequate documentation?

14. Data quality: Are the data sufficiently "cleaned" so that the data set can be added to the collection without further processing?

15. Access: Is the data set accessible to all users? Are there any restrictions? Is it accessible online?

16. Producer reliability: Is the distributor/producer of the data reliable? Are its products well regarded?

17. Historical importance: Is the data set worth preserving even if use is limited in the foreseeable future?

18. Ownership: Who retains ownership of the data set?

19. Levels of analysis: Does the data desired level of analysis?

For guidelines to the evaluation of a scientific data set (as opposed to a social science dataset), see Bruce Ewbank's "Comparison Guide to Selection of Databases and Database Services" (1982).

Drafting of a sound, collection development policy is a prerequisite before a library engages in a full-scale effort to acquire data files. Otherwise, bibliographers may well be forfeiting responsibility for selection to database service suppliers and vendors, or to the user community. Collections that are based entirely on demand without any clear policy may be uneven, lack depth or focus, and become unmanageable. As Bisco (1970, 282) pointed out almost twenty years ago, "there are notable gaps in the collections of archives."

At the very least, the process involved in drafting a policy can be used to bring together all bibliographers and other librarians so that it serves an educational and unifying function, and further help integrate mrdfs into a traditional library setting.

The challenge for those of us who straddle both the library and the archival worlds is to develop a collection policy that is not overly restrictive, but flexible enough to permit us as selectors the necessary leeway to develop our collections.¤

References

Adams, Margaret O'Neill and Jack Dennis. 1970. Creating local social science data archives. Social Sciences Information, 9/2 (April), 51-60.

Association of Research Libraries. 1984. Nonbibliographic macine-readable data bases in ARL Libraries (Washington, D.C.: ARL, Office of Management Studies, Systems and Procedures Exchange Center). SPEC Kit 105.

Bernard, H. Russell and Ray Jones. 1984. The use of MRDFs in the social sciences: an anthropologist and a librarian look at the issues. In Association for Research Libraries, Nonbibliographic machine-readable data bases..., 94-99.

Bisco, Ralph L. 1967. The research library and data archives for social research. Speech prepared for the dedication of the Graduate Research Library, University of Florida.

———————. 1970. Summing up. In Ralph L. Bisco, editor, Data bases, computers, and the social sciences (New York: Wiley), 273-285.

Chiang, Katherine S. 1986. Computer accessible material in the academic library: avoiding the kludge. In Danuta A. Nitecki, editor, Energies for transition: proceedings of the Fourth National Conference of the Association of College and Research Libraries, Baltimore, Maryland, April 9-12, 1986 (Chicago: The Association), 67-69.

Claydon, Charles R. and Dagobert Soergel. Numeric databases. Drexel library quarterly, 18/3-4 (Summer-Fall).

Converse, Philip E. 1964. A network of data archives for the behavioral sciences. Public opinion quarterly, 28/2 (Summer), 273-286.

Dennis, Jack. 1977. The relation of social science data archives to libraries and wider information networks. In Howard D. White, editor, Reader in machine-readable social data (Englewood, Colo.: Information

Handling Services), 165–174.

Dionne, JoAnn. 1988. Electronic mail message to Dan Tsang, 8 March.

Dollar, Charles M. 1980. Machine-readable records of the federal government and the National Archives. In Carolyn L. Geda et al, editors, Archivists and machine-readable records, 79–88.

Ewbank, W. Bruce. 1982. Comparison guide to selection of databases and database services. Drexel library quarterly, 18/3–4 (Summer–Fall), 189–204.

Fischer, Jayne D., editor. 1980. Readings in business and economic research management: execution and enterprise. Vol. 1. Madison: University of Wisconsin, School of Business, Office of Research Administration.

Geda, Carolyn L. et al, editors. 1980. Archivists and machine-readable records: proceedings of the Conference on Archival Management of Machine-Readable Records, February 7–10, 1979, Ann Arbor, Michigan (Chicago: The Society of American Archivists).

Gerken, Ann E. 1984. Organizing a data library. IASSIST quarterly, 8/4 (Winter), 7–10.

Heim, Kathleen M., issue editor. 1982. Data libraries for the social sciences. Library trends, 30/3 (Winter).

Henderson, Jim. 1988. Electronic mail message to Dan Tsang, 18 May.

Hernon, Peter. 1986. Numeric data bases. Encyclopedia of library and information science, 40/suppl. 5, 339–354.

Jacobs, Jim. 1986. Collection profile. University of California, San Diego.

Janda, Ann. 1988. Electronic mail message to Dan Tsang, 11 March.

Jones, Ray. 1982, The Data Library in the University of Florida Libraries. Library trends, 30/3 (Winter), 383- 396.

———————. 1988. Telephone interview, 6 May.

Jones, Susan D. and J. David Singer. 1972. Beyond conjecture in international politics: abstracts of data-based research (Itasca, Illinois: F.E. Peacock).

Librarians Association of the University of California. 1983. Textual and numeric data files: primary information sources in machine-readable form. Reprinted in Assocation for Research Libraries (1984), Nonbibliographic Machine-Readable Data Bases..., 5–18.

Mitchell, Robert Edward. 1965. A social science data archive for Asia, Africa and Latin America. Social sciences information, 4/3 (September), 85–103.

Nasatir, David. 1973. Data archives for the social sciences: purposes, operations and problems (Paris: UNESCO).

———————. 1977. Stalking the wild data set: the acquisition of machine-readable social science data at home and abroad. Drexel library quarterly, 13/1 (January), 43–47.

Piovesan, Walter. 1988. Electronic mail message to Dan Tsang, 5 April.

Pool, Ithiel de Sola. 1965. Data archives and libraries. In Carl F.J. Overhage and R. Joyce Harman, editors, INTREX: Report of a planning conference on information transfer experiments, September 3, 1965. (Cambridge: MIT Press), 175–181.

Pope, Nolan F. 1984. Providing machine-readable numeric information in the University of Florida Libraries: a case study. In Ching-Chih Chen and Peter Hernon, editors, Numeric databases (Norwood, New Jersey: Ablex), 263–282.

Robbin, Alice. 1972. The pre-acquisition process: a strategy for locating and acquiring machine-readable data. Drexel library quarterly 13/1 (January), 21–42.

Ruus, Laine G.M. 1982a. The University of British Columbia Data Library: an overview. Library trends, 397–406.

———————. 1982b. Data Library Procedures Manual. Excerpted in Association of Research Libraries (1984), Nonbibliographic machine-readable data bases..., 62–97).

State University of New York, Albany. 1983. Final report. Task Force on

Non-bibliographic Databases. Reprinted in Association for Research Libraries (1984), Nonbibliographic machine-readable data bases..., 22-56.

University of California, Davis, Library. 1987. Report of the Committee on Numeric and Textual Databases: summary of comments and recommendations for revision.

University of California, Irvine. 1983. Collection development policy for the University Library.

Warner, Robert M. and Francis X. Blouin, Jr. 1980. Some implications of records in machine-readable form for traditional archival practice. In Carolyn L. Geda et al, editors, Archivists and Machine-Readable Records, 242-248.

White, Howard Dalby. 1974. Social science data sets: a study for librarians. Dissertation (Ph.D, Librarianship). University of California, Berkeley.

————————, editor. 1977. Machine-readable social science data. Drexel library quarterly, 13/1 (January).

Yantis, Betty. 1980. Data acquisition, storage and management in a new research center. In Jayne D. Fischer, editor, Readings in Business and Research Management, 117-119.