Issues concerning the bibliograhic citation of machine-readable data files

by Richard Hankinson[1]
Editor, Population Index
21 Prospect Avenue
Princeton, NJ 08540

I would like to take the opportunity provided by this IASSIST meeting not to present a formal, academic paper on a substantive topic, but to set the background for a dialogue between a community that is broadly represented by the IASSIST membership, and a service, represented in this case by Population Index. My understanding of the IASSIST community is that it represents an international group of individuals and institutions concerned with the management, operation, and use of machine-readable data archives. Population Index, in contrast, is an annotated bibliographic journal that covers the world's population and demographic literature. The area of common interest I would like to explore and discuss with you has arisen from a decision that we took some seven years ago to cite not only books, journal articles, and other published materials, but also machine-readable data files (MRDF). This decision was made, primarily, because we felt that many of the demographers we serve are not only computer-oriented but are more used to working with data in

[1]Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19-22, 1987.

machine–readable form than in traditional printed form, and because we felt that, for a variety of reasons, a growing amount of the relevant demographic data — particularly that produced by national statistical offices concerning vital statistics or census –is only available in machine–readable form, and therefore that we are obliged to cite it in such form if our coverage is to be as complete as we wish it to be.

Before coming to specifics, I would like to give you a little background information. Population Index has been providing bibliographic coverage of the population literature for just over 50 years. It is based at Princeton University's Office of Population Research, the first university–based center for demographic study set up in the United States, founded in the 1930s. Population Index consists primarily of an annotated bibliography of the population literature, which includes literature in all languages; the emphasis, however, is on Western and Slavic languages, in which most of the materials intended for general distribution are published. Approximately 3400 citations, complete with abstracts, are produced each year. These are made available in two forms: firstly, in a quarterly journal, with a global circulation of some 4700, sent primarily to members of the international and U.S. professional associations of demographers and population experts and to institutions; and secondly, as a computerized bibliographic data base, POPLINE, available through the MEDLARS system at the National Library of Medicine in Bethesda, Maryland. This is a cooperative venture involving four U.S.–based population centers at Columbia, Johns Hopkins, the University of North Carolina, and Princeton. Population Index is a fully computerized operation working with the University's mainframe computer, an IBM 3081. It is the product of three editorial/bibliographer professionals (one of whom is a data base specialist), with an administrative/clerical/data entry support staff. Funding is provided partly from federal sources (including NICHD and USAID), partly from subscriptions provided to professional associations (the Population Association of America and the International Union for the Scientific Study of Population), and the remainder from paid subscriptions to the quarterly journal.

Tabel 1: Mrdf Citations 1980-1986 by year and country

	1980	1981	1982	1983	1984	1985	1986	sub-total	total
World					1	6	2		9
Australia	3	1	-	3	-	4	10		21
U.S. Census	5	5	6	28	27*	11	17	99	
U.S.NCHS	-	8	3	-	2	-	10	23	141
U.S. Other	2	1	1	1		10	4	19	
Malaysia	-	-	2	-					2
Canada					2				2
France					2				2
Brazil						1			1
Scandinavia							2		2
Israel							6		6
TOTAL	10	15	12	32	34	32	51	141	186

The subject area covered concerns population and demography, which includes such concepts as population size and growth, spatial distribution, mortality, fertility and family planning, nuptiality and the family, migration, historical demography, population characteristics, population policy, population statistics (including censuses, surveys, and vital statistics), and the relationships among demographic factors and socioeconomic development, natural resources, and the environment.

Over the seven-year period since 1980, Population Index has cited 186 MRDFs (see Table 1). Of these, just over 75% have concerned the United States, some 53% from the U.S. Bureau of the Census alone. While it is reasonable to expect a preponderance of U.S. MRDFs in a listing of this kind, the geographic breakdown indicates either a surprising lack of MRDF containing population data from other countries or a failure of the existing information services — including Population Index — to identify such files and provide information on them to the user public.

Apart from the United States, only Australia seems to be adequately represented over the full period. The 15 citations relating to Malaysia, Canada, France, Brazil, Scandinavia, and Israel were largely the result of information originally gathered from the Guide to Resources and Services produced annually by the Inter-University Consortium for Political and Social Research (ICPSR),[2] and it is at least possible that they represent only the tip of the iceberg. We know that the official statistical agencies of many of the European countries are creating MRDFs containing demographic data. If so, where is the information on these products available on a regular basis? We certainly have not been able to provide our users with a steady flow of information on new products through the 1980s. These and other related questions are at the top of our agenda at this time, and we have come to Vancouver to help find some answers. We particularly want to determine to what extent our lack of success in citing such files is our own fault, which we can and will rectify when we know how to do so, and to what extent it is due either to the lack of adequate information from the creators of such files—which is a more difficult situation to change, but may be one where action is needed—or to inadequacies in information services such as those provided by ICPSR. The same questions may well be posed for other parts of the world. Japan is a case in point in the developed world. As for developing countries, the situation is variable according to our limited information. We have extensive information concerning the MRDFs created during the course of the World Fertility Survey and the Contraceptive Prevalence Surveys, primarily during the 1970s.[3] Furthermore, the Dynamic Data Base in Voorburg, the Netherlands, has an on-going program to add new files concerning survey data on fertility.[4] Latin America has several countries generating MRDF with demographic data, thanks in part to efforts by the Centro Latinoamericano de Demografia (CELADE), a U.N. organization in Santiago, Chile, that has worked with countries for many years and built up a data base of census and survey data sets for Latin America. They also have an ongoing information program that publishes information on a regular basis concerning its holdings in this area.[5] It is unlikely that there are extensive demographic MRDFs in Africa at this time, but they probably exist

---

[2]Inter-University Consortium for Political and Social Research (ICPSR), Guide to Resources and Services, 1984-1985. Ann Arbor, Michigan. 525 pp.

[3]International Statistical Institute (ISI) and International Union for the Scientific Study of Population (IUSSP). Dynamic Data Base. Catalogue of Survey Data Files. 153 pp. Voorburg, Netherlands. January 1987.

[4]Cleland, J.G. "A new service for demographic analysis: the Dynamic Data Base. Population Index (Princeton, N.J.) 52(4), pp. 540-7. Winter 1986.

[5]United Nations. Centro Latinoamericano de Demografia (CELADE). Bulletin of the Data Bank (Boletin del Banco de Datos), Santiago, Chile. No. 11, LC/DEM/G.39, April 1986. 52 pp.

in several Asian countries if we could track down the available information.

The next question I would like to raise concerns what elements are necessary to a complete bibliographic citation of an MRDF. The policy followed by Population Index is based on guidelines suggested by Judith Rowe,[6] which were in turn based on standards developed by the American Library Association's Subcommittee on Rules for Cataloging Machine Readable Data Files, and the work of Sue A. Dodd within the IASSIST framework. They are consistent with the ANSI standard and its application for social science data files as enunciated by Dodd.[7] The elements of a citation we have been trying to include are as follows:

1. Authorship: full name of author(s) or corporate body responsible for the intellectual content of the file (e.g., principal investigator, project director, or sponsoring agency).

2. Title: full title (no acronyms) containing descriptive words or phrases and dates.

3. Subtitle: secondary title to amplify or restrict main title as in serial or multi-part works.

4. General material designator: denotes the generic form or type of material cited, e.g. MRDF.

5. Statement of authorship: indicates the relationship of the work to the person(s) or corporate body named in the author heading or to other significant parties such as principal investigator, sponsor, or even funding agency.

6. Edition: provides users with valuable information on the source, date, or revisions; new editions indicate addition or deletion of data elements, variables, or fields; recoding or restructuring of the file; addition or deletion of logical records (cases, observations, etc.) or, in the case of programs, changes in the programming language.

7. Imprint: consists of the producer statement and the distributor statement.

8. Extent of file: the MRDF equivalent of number of pages is number of logical records. Other physical characteristics such as tape size, recording density, etc., are excluded since distributors normally offer various recording options. This information is subject to change and it does not affect bibliographic identification, which is determined by the content of the file and not by the container.

9. Accompanying material: includes codebooks, reports, manuals, and other associated material in printed or machine-readable form. Such materials may or may not warrant separate citation.

10. Series statement: collective title(s) and item number or other designation within the series.

---

[6]Rowe, Judith S. "Population Index to cite publicly available machine-readable data files." Population Index (Princeton, N.J.), 45(4), pp. 567-75. Winter, 1979.

[7]Dodd, Sue A. Bibliographic references for numeric social science data files: suggested guidelines. Journal of the American Society for Information Science (Washington, D.C.), 30(2), pp. 77-82. 1979.

In addition, each citation includes an appropriate abstract.

The main problems we have had concerning the preparation of MRDF citations have been as follows:

1. • No. 5. Statement of authorship. Since we do not have a field for this, we have been including this information, the few times we have found it appropriate, in the abstract following the citation.

2. • No. 6. Edition. We have not found it possible to provide the necessary information in this field because we have been unable to obtain it. We suspect that this information is available, but to date when we have asked MRDF distributors or publishers for information about whether the file they are listing is a revision or edition, they have not been able to supply us with the answers. At this stage, we really do not know if there is a problem here or not. How often are the MRDFs of demographic interest revised and updated? Most of the U.S. examples we are aware of, such as the Current Population Survey, can be cited as new files rather than revised files. Certainly, Population Index has not recited any MRDF in the last seven years because we have learned that revised and updated editions are available.

3. • No. 8. Extent of File. We have had problems in this field primarily because we are editors and bibliographers rather than computer experts. We understand that the key facts to include are number of logical records and logical record length, and we add this information in the abstract if we know it. However, the information we receive does not always include this information in a way we can understand it. Furthermore, the ICPSR Bulletin, for example, as far as we can establish does not specify whether it is a question of logical or physical records. The objective should be to provide the necessary information so that a potential acquirer or purchaser of the MRDF being cited knows whether he or she can mount and use the MRDF on the equipment available. The question we have to clarify is what that information needs to be.

4. • No. 10. Series Statement. We don't seem to pick up or include any information in this category.

There is one related item we would like guidance on. It is our custom now to add information for each citation on the source of information (e.g. the ICPSR Bulletin) and on the location of the file in question (e.g. U.S. National Technical Information Service). Would it be useful to add to this information addresses for the location of the agency from which the file could be obtained? Or can we assume that most people communicate with the appropriate government agencies by phone and that these numbers are easily obtained by those who need them?

We are also not sure whether we need to add information about the release status of the data contained in MRDF. Although many such files containing demographic data are publicly available, the data in them is often restricted: in many cases, no data can be released without the specific authority of the government of the country concerned. Should we cite MRDF files with restrictions of this kind -and if so, what information should we provide on such restrictions?

Another issue that has concerned us is the relationship between MRDF and hardcopy or published materials. The general principle we follow is that we will cite both the report and the MRDF if they are complete in themselves. If the hard-copy materials are only codebooks, reports, and manuals that are to be used in conjunction with the MRDF, they are not cited separately but are merely referred to in the abstract following the MRDF citation. In other words, we feel that citing the same material in two different forms, paper and MRDF, is not an unnecessary duplication given the different needs of our users.

In conclusion, we are struggling to expand the service we are offering to the demographic community by preparing citations and abstracts to appropriate MRDF. As this review has indicated, we have had some success but have run into two main obstacles. The first is our lack of information concerning the availability of new MRDFs in this area, and we need to work with the IASSIST community to establish the extent to which the information exists and we are missing it, and the extent to which it does not exist and to which we can encourage those responsible to provide it. The other obstacle is interdisciplinary: with our bibliographic, editorial, and library backgrounds, we have a familiarity and expertise with the printed word. Our efforts to include MRDFs involve a new concept and language that we will be struggling with for some time.□