iassist NEWSLETTER vol. 1, no. 4

DISCUSSION PAPER / sue dodd

Titles: The Emerging Priority in Bringing Bibliographic Control to Social Science Machine-Readable Data Files [MRDF]

by

Sue Dodd
Institute for Research in Social Science
University of North Carolina
Chapel Hill, NC

Two recent attempts to compile "catalogs" of social science data have encountered the lack of consistency among titles for the same data set. One attempt has been the recent cataloging efforts at the Universities of North Carolina, Wisconsin, Princeton, and Yale, whereby traditional library cataloging records are created for social science data generated by academic research. The other has been the efforts of the Association of Public Data Users (APDU) to compile a directory of publicly available data files which represent primarily government produced data. Both groups have experienced the same problem: variance of titles for the same data file. Yet, without some control over titles and some mutually agreed upon primary source of title information, there can be no bibliographic control of social science data and none of the related products such as a union list of machine-readable data files. This paper will attempt to offer some suggestions for remedying the situation, including guidelines for transcribing titles; for creating a "title page"; for compiling a bibliographic reference; and for establishing an "authority list" for titles.

Origins of titles for social science data files

Unlike a book a social science data file may exist for a long time without a title. Until it has been properly titled, it may be known only by a study number (e.g., Study #5063), or by the name of the principal investigator (e.g., The Stouffer Study), or by the source of production (e.g., The RAND Survey). If a data file survives the time period between data collection, data analysis, and data publication sans title, it is likely to assume the title of the primary publication (e.g., Communism, Conformity, and Civil Liberties).

The first appearance of a title usually occurs with the generation of early sources of documentation. Documentation may include a questionnaire, coding instructions, codebook, manual, or project report. Given the nature of MRDF, some type of accompanying documentation is required in order to "read" the data. Titles recorded on documentation are also the most visible because "containers" (protective canisters) of MRDF have no identifying titles; or if they do, it is usually a shortened title given the space constraints of the container. However, an initially applied title of a MRDF is not necessarily the only or lasting one.

During the life cycle of a social science data file, a title is frequently changed or modified as responsibility changes for file creation, processing, analysis and reporting. For example, one group of persons may be responsible for actual data collection plus the conversion to an "automated" format, while another group may be responsible for the data analysis and data reporting. Such diversification of labor often leads to different titles for the same data. After the primary analysis, reporting, and possible publication by the principal parties, a data file may be deposited with a data archive, center, or library for the purpose of secondary analysis. At this point, the data and documentation may go through further processing, including a new codebook and a new title. At about the same time, but not necessarily by the same person, a data abstract, study description or some type of informational notice is written to publicize its availability to the general public.

If there are dual or multiple distributors of the same data file, titles could easily vary from distributor to distributor. For example, there are at least three known distributors for a particular Harris survey with the following titles:

Violence in America
The American Public Looks at Violence
Harris 1968 Violence Survey, #1887
Harris Poll: "The American Public Looks at Violence"

In the case of the APDU directory, the overlap of mutually held and accessable public data files was impossible to determine, since members had listed the same data file under various titles. For example:

City and County Data Book
County and City Data Book
1972 County and City Data Book
County and City Data Book Tape

The cataloging experience at the University of North Carolina has revealed that out of approximately 500 separate social science data files from many different sources, close to 80 of these files had variant titles. In most cases, the title in the codebook varied from informational listings provided by the distributor of these data.

As the life cycle of a data file continues, popularized titles begin to evolve and grow organically and are usually a modified version of the primary title. For example:

Modified title:	French and German Elite - Arms Control Data
Primary title:	Arms Control in the European Political Environment: French and German Elite Responses, 1964

Other titles are compressed into acronyms:

Modified title:	The CSEP Study
Primary title:	The Comparative State Election Project

Others take on the name of the principal investigator(s):

Modified title:	The Matthews-Prothro Study
Primary title:	The Negro Political Participation Study

Finally, variant titles may appear in "notes" or in bibliographic references in the various scholarly journals. Without any guidelines on how to cite numerical data files and without any control over the proliferation of titles, title information will vary among scholars. Often, the fault rests not so much with the person citing the data as it does with the distributing agency which has failed to provide proper bibliographic information on a particular data file.

Summarizing, the history of a data file usually reveals the various levels of title changes and modification. However, it is unlikely that a cataloger or a scholar will have access to this history. Instead, he will be confronted with the problem of choosing or citing one title from among many for his respective uses.

Guidelines for transcribing titles

In our attempt to offer suggestions on how this situation could be remedied, this section describes the basic components of a title and suggests guidelines on how to transcribe a title for social science data.

Components of a title for social science data files would include the following: 1) descriptive words indicating content; 2) geographic focus or unit; 3) chronological year(s) of data target or data collection; 4) source of data (e.g., court records); 5) producer, contributor or sponsor of data; and 6) study or series number (if important for ordering or for distinguishing one data file from another).

Guidelines include the following:

I. Make the title as descriptive and as complete as possible. A good title should be descriptive of the contents of the document or data it is describing and should include as many of the components described above as are applicable. If there is one major theme or focus, then this should be mentioned in the title. If the data contain information on many different topics, none of which appears to dominate, then a broader or more general subject approach may be taken (e.g., Harris 1972 Public Opinion Survey; or the National Opinion Research Center 1974 General Social Survey; or the Survey Research Center 1976 Social Indicator Survey). When transcribing a title, be aware that the descriptive words contained in a title take on added significance with the existing technology for keyword or full-text retrieval. For example, the only subject approach to SOCIAL SCIENCE CITATION INDEX (whether it be by the printed reference work or by the on-line search capability) is via the descriptive words contained in the respective titles.

II. If at all possible, DO NOT take a title from a publication based on the data file, as this may cause copyright violations and problems with international coding schemes such as ISBN (International Standard Book Number). If this cannot be avoided, then a qualifier should be attached at the end of the title. For example:

Civic Culture (Machine-readable data file)
Communism, Conformity and Civil Liberties (MRDF Source Documentation)

III. For any data that are part of a predictable series (occurring at definite time intervals, such as the Census or election surveys), titles should be consistent throughout the life of the series.

IV. For data that are part of an on-going collection or series (collected at non-predictable intervals and with varying subject focus), one may consider the following sequential title arrangement: 1) organizational name of producer; 2) chronological date of data target or data collection; 3) geographic focus (if unique); 4) descriptive content (including subtitles); and 5) study or series number.

An on-going series of data (such as public opinion polls) tends to be associated with the originating source or producer of these data. Therefore, it is recommended that the organizational name of the producer come first. This arrangement also allows for a large collection of data from the same source to be grouped alphabetically for easy reference. For example:

Harris 1969 Morals and Values Survey, No. 1933
Harris 1969 Science, Sex and Morality Survey, No. 1927

In cases where there is more than one data collection per year on a given topic, the month or season could follow the year in parenthesis. For example:

Survey Research Center 1957 (Fall) Consumer Attitudes and Behavior Survey, No. 3631

If the geographic focus is unique, it is recommended that it be included in the title. For example:

American Institute of Public Opinion 1975 Japanese Election Survey, No. 7811
Harris 1965 Dallas Sports Survey, No. 1545

To indicate that data in a continuing series may have a varying subject focus, it is recommended that sub-titles be used. For example:

Survey Research Center 1963 Detroit Area Study: A Study of Family-School Relationships
Survey Research Center 1964 Detroit Area Study: The Measurement and Validation of International Attitudes

Study numbers should be included in the title if they are part of an on-going collection of data and are consequently helpful in distinguishing one data file from another. For example:

Harris 1967 Public Opinion Survey, No. 1702
Harris 1967 Public Opinion Survey, No. 1718

National Opinion Research Center 1963 (January) Amalgam Survey, SRS-100
National Opinion Research Center 1973 (December) Amalgam Survey, SRS-4179

V. Avoid beginning a title with articles (such as a, an, the, etc.).

VI. Avoid beginning a title with numerics (e.g., 1972 County and City Data Book). With most computerized alphabetic listings, those titles beginning with numerics are placed either at the very beginning of a listing or at the very end. Such placement may cause certain data files to be overlooked.

VII. Avoid using acronyms in titles. The full meanings of acronyms should be spelled out and if used at all, should follow full meanings enclosed in parenthesis. For example:

World Event/Interaction Survey (WEIS)

VIII. When applying titles to sub-sets of data files, indicate both the original data title and the fact that it is derived from a larger file. For example:

Comparative State Election Project: Federal District Sub-File

Title control and an "authority list" of titles

Title control for social science data must be applied at one of two stages in the life of a data file: either at the production level or at the distribution level. Ideally, the creator or producer of a MRDF should apply the "authoritative" title. However, in those cases where this responsibility has (for whatever reasons) defaulted to the distributor of the data, then he should provide the singular title. All other references to this MRDF should carry this title.

One way to bring some order to the existing chaos among titles, is to establish an "authority list" of titles for social science data files. Such an effort is being undertaken by members of APDU; and a similar "union list" of catalog records would have the same effect. Again, the primary responsibility for establishing or determining the authoritative title rests first with the producer and then with the distributor. If the producer has abdicated that responsibility when depositing a data file with an archive or data center, then responsibility lies with the distributor.

In those cases where there are multiple distributors of the same data, then a determination has to be made as to the one with the most "authority," or official status, or national prestige, etc.

For data files that have been changed through major processing techniques or reformatted for a more efficient "reading," or have been changed in terms of content or observations, then the title remains the same but the data become a new edition. Thus, an "authority list" of titles would include the various editions of MRDF, just as the National Union Catalog (NUC) carries the various editions of books.

Sub-files taken from larger data files should carry a distinctive title, and if not, the producer or distributor should modify the title with some type of qualifier (e.g., sub-file A; selected sub-files, etc.).

The major data producers and distributors of social science data would be responsible for publishing their respective lists of authoritative titles. These "authority lists" of titles could then be published in some appropriate newsletter or publication such as SSdata or the IASSIST Newsletter.

In establishing "authority lists" of titles there is also the need for establishing a concensus on the primary source of title information. If there is a title on the codebook and another title in a directory, which is the correct title? Without having access to the history of a particular data file, how can the judgement be made as to the proper title? One answer would be to rate, in order of importance, the various sources of documentation. For example, the codebook or its equivalent would be the primary source of title information; the data directory or study description would be the secondary source; the reporting source or publication would be the third, etc. Some discipline has to be applied to social science documentation in general, but specifically to the bibliographic aspects of such documentation. Such responsiblity should not end with titles, obviously, but should be extended to include all the components of a bibliographic citation. For example, the information required for compiling a bibliographic reference should come from the documentation accompanying a data file, and the most obvious place to derive this information would be from the "title page" of that documentation.

Title page for social science MRDF

In the past, very few data producers or distributors have taken the style or content of the title page of documentation seriously. However, if social science data files are to be readily accepted into the mainstream of bibliographic control, then more attention has to be given to these title pages. For example, the title page of a book becomes for the cataloger, the principal source of information. It is so respected by catalogers that the information contained on the title page becomes as "dogma" and cannot be deviated from in the transfer of information to the catalog record. However, the quality and amount of information provided on most title pages of social science documentation cannot be taken seriously by a cataloger. In truth, many sources of documentation for MRDF *have no title page* equivalent.

Information contained on a title page of MRDF documentation should consist of the basic bibliographic components including authorship; title; medium designator; edition; imprint; and series statement.

The "medium designator" is a term used to denote the generic form or type of material listed or referenced. It is used to distinguish one type of medium from another and to provide clarity. The most universally accepted term for this medium is "machine-readable data file".

The "imprint" includes the place of production; name of the producer; date of production; place of distribution; and name of distributor. The producer is defined as that party responsible for the collection, compilation, and physical production of the data (*i.e.*, the mechanized process of transforming information into the format known as "machine-readable") and the distributor as that party responsible for disseminating the data to others upon request.

The "series statement" would provide the reader with relevant information about an on-going collection of data (*e.g.*, SRC/CPS 1958 American National Election Study, No. 4).

As mentioned earlier, an "editon" occurs when data files have been modified through major processing techniques or reformatted for a more efficient "reading" or when the data have been changed in terms of content or observations. Edition statements appear in an abbreviated format on a title page (*e.g.*, DUALabs ed., NORC rev. ed., or 1st ICPSR ed.).

Although, it is highly recommended that only the basic bibliographic information be placed on a title page, there may be situations where additional information would be helpful. Examples would include: date of data focus, if not part of the title; source of funding, if appropriate; scope of documentation, if documentation consists of more than one volume; study number, if necessary for identification or ordering purposes; etc.

Information pertaining to unique classification schemes such as the International Standard Book Number (ISBN); the Library of Congress Card Number; and the catalog card facsimile should appear on the verso of the title page. For an example of a title page of a machine-readable data file codebook, see appendix.

Placement of the information is flexible. However, placement of a study number behind or immediately under a title will be construed as being part of that title. For example:

The SRC 1952 Election Study (S400)
or
The 1972 German Election Panel Study
(Zentralarchiv Nos. 635,636,637 -- ICPR No. 7102)

Bibliographic references for social science data files

The title page, in addition to providing the basic information required for the catalog record, should also provide the information required to compile a proper bibliographic citation. The Classification Action Group of IASSIST has been working to define the necessary components and structure for a proper bibliographic citation and has followed the guidelines provided in the forthcoming publication entitled: American National Standard for Bibliographic References. However, these guidelines have been modified slightly to represent the particular needs of social science numerical data and to conform with the forthcoming AACR II cataloging rules.

The basic components of the bibliographic reference would include authorship; title; medium designator; sub-title; edition; imprint (place of production, name of producer, date of production; place of distributor, and name of distributor); extent of file; notes; and series statement. The examples that follow were provided by the Classification Action Group:

Title first: Mexico's naturalized citizens, 1828-1931 [Machine-readable data file]. Principal investigators: Harold Sims, Susan Sanderson, and Philip Sidel. Pittsburgh, PA : University of Pittsburgh, 1975-76 [producer and distributor]. 1 data file (8,066 logical records).

Author first: Shanas, Ethel. The health of older people [Machine-readable data file] : a social survey : public attitudes of older people. Norc rev. ed. Chicago : National Opinion Research Center, 1957 [producer and distributor]. 1 data file (2567 logical records) and accompanying codebook (166p.).

Swidzinski, Susan. Syllabication [Machine-readable data file] : a drill and practice lesson. Bloomington, MN : Control Data Corporation, 1976. On-line program lesson available only via the Plato System.

Henry, Neil. MAXCLS.BAS [Machine-readable data file] : a program for maxium likelihood estimation of paramaters of unrestricted latent class models. Lafayette, IN : Gary Income Maintenance Experiment, 1974 : Pittsburgh, PA : Social Science Computer Research Institute [distributor]. 1 program file (95 statements, BASIC) and accompanying manual (53p.).

The information contained in the brackets, while highly recommended by the Classification Action Group, are optional according to the ANSI standards and the forthcoming AACR II rules. The extent of file (logical records, program statements, etc.) and "notes" are also optional. Many distributors of data already provide the user with some "data acknowledgement" information. Guidelines or examples of how to cite these data in the literature should be part of this information.

Conclusion

In conclusion, we have discovered that it is not unusual for social science data files to receive many different titles in their lifetime. Some titles are modified through data processing efforts; others grow or evolve from popular usage; and even others are erroneously recorded from one source to another. This lack of title control has proven to be detrimental to the efforts of those who are attempting to compile any authoritative listing or "catalog" of social science data. It is also apparent that there is an immediate need for an "authority list" of titles, and that some decision has to be made regarding the primary source of title information.

It is hoped that this paper will bring these problems and needs to the attention of those parties who have both the responsibility for and the control over the situation. Critical attention and immediate action, on the part of the major producers and distributors of data, is necessary to bring about true title control and better bibliographic documentation. Without it, information on social science data files will remain in "elite obscurity".

Appendix

MACHINE-READABLE DATA FILE COOEBOOK

BERKELEY RADICALS FIVE YEARS LATER: A FOLLOW-UP
SURVEY OF STUOENTS WHO WERE ARRESTED IN THE 1964
FREE SPEECH MOVEMENT AT THE UNIVERSITY OF CALIFORNIA

Conducted by
The Oetroit Free Press of Knight-Ridder Newspapers, Inc.
under the direction
of
Philip Meyer and Michael Maidenberg

July 1969

SSDL 1972 rev. ed.

SOCIAL SCIENCE OATA LIBRARY - UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA
27514