Toward a National Information System for Social Science Data Files*

Despite a breakthrough in library cataloging of files of machine-readable social science data (directions are expected in the next *Anglo-American Cataloging Rules*), the U.S. system for delivering information about data files still needs strengthening. While descriptive cataloging of locally held files by research libraries should prove helpful, suppliers' restrictions on "rediffusion" of data reduce the value of a national union catalog. Subject cataloging is also a problem, although a new form heading for data files would improve the usefulness of the *Library of Congress Subject Headings*. The greatest improvement would be inclusion of abstracts and indexing of the files in a national bibliographic service with on-line search capability, especially one that could provide the codebooks required by those wanting to reuse others' data. With NTIS furnishing a prototype, ERIC is here proposed as a model of what is needed. ERIC has many desirable features as a "data information system" that would not disrupt present economic arrangements for delivering the files themselves.

Howard D. White
Graduate School of Library Science
Drexel University
Philadelphia, PA 19104

• Introduction

Social science data files are an important type of machine-readable research product not yet under adequate bibliographic control. These files bear the raw numeric data of, e.g., surveys, polls, censuses, and tests. Because the data can be used and reused in research and teaching, they have a permanent value. Several major suppliers have emerged to acquire reusable data and make them available nationwide—among them, the Interuniversity Consortium for Political and Social Research, the Roper Public Opinion Research Center, and the National Technical Information Service. Their institutional customers, mostly campus-based, are local data archives. Suppliers and archives jointly have done impressive work over the past 15 years to publicize their holdings, but still have not matched systems currently providing bibliographic information about other media. Two possible innovations to make social science data files more widely known and used will be discussed here—library cataloging, resulting in a national union catalog; and a national service combining abstracting and indexing, on-line search capability, and document delivery.

Data files have two parts. The document on which any bibliographic system will rest is the human-readable part called the codebook. It complements the machine-readable part, which is a large body of numerals stored on cards, disk, or tape. The numerals are values of variables. The codebook identifies each variable, translates its values (often literally codes) into intelligible language, and tells where each variable is stored, relative to others (see Fig. 1). Obviously the codebook is essential to anyone who wants to use the file in a computer, since variables cannot be defined or interpreted without it. Fortunately, it can be provided for consultation anywhere, including libraries, as a document separate from and much cheaper than the machine-readable part of the file. In file description, the codebook will be a primary

*Revised version of a paper presented before the Special Interest Group on Behavioral and Social Sciences (SIG/BSS) with the title "A National Information System for Social Science Data Files," at the annual meeting of the American Society for Information Science, October 5, 1976.

DECK 1
COLUMNS 1-5
Q. 1

NATIONAL DATA PROGRAM FOR THE SOCIAL SCIENCES

CODES FOR THE SPRING 1972 SURVEY

The codes give the location of all questions by deck and column number. Given for each question is the original question wording or variant as indicated, interviewer directions contained in the questionnaire, the response categories, the punch designations, and the number of responses ("N").
Added comments are either bracketed "[ ]" or indicated under "REMARKS."

BEGIN DECK 1
IDENTIFICATION NUMBER,
COLS. 1-4

First, I have a few factual questions about yourself.

1. Which of the categories on this card comes closest to the type of place you were living in when you were 16 years old?

HAND CARD A

COL. 5

RESPONSE	PUNCH	N
In open country but not on a farm	1	147
On a farm	2	340
In a small city or town (under 50,000)	3	500
In a medium-size city (50,000-250,000)	4	198
In a suburb near a large city	5	91
In a large city (over 250,000)	6	334
Don't know	8	1
No answer	9	2

REMARKS: Card A contained responses for punches 1 to 6 only.

Fig. 1. Sample page from a codebook. The codes are the numerals under the heading PUNCH. For example, the numeral 5 punched in column five of the first 80-column card per respondent ("Deck 1") means that at age 16 the respondent lived "in a suburb near a large city"–a response given by 91 persons. The logic of all other codes and columns is the same. This particular codebook is a model of its kind. Data files should not be publicly released without similarly good documentation.

source of the abstract; it will be what is subject-indexed (either on the level of the entire file or on the level of individual variables); and it will often be what the cataloger uses for library cataloging.

The prospect of library involvement with data files is not farfetched. Under forthcoming Anglo-American rules, catalogers will be able to treat these files as one more "nonbook" medium for which entries can be made. If that is done, not only local data archives but local research libraries will provide intellectual access to data – a desirable outcome, as will be shown. Library cataloging will also make possible a recurrent dream of the data archivists: a national union catalog of their holdings, similar to those published by the Library of Congress for other media. However, there is a question whether a national union catalog as such would best serve the interests of its intended users. For reasons noted below, many data files are not freely diffusible through interlibrary loan, and this restriction undercuts a major purpose of the union catalog.

The other innovation would be a national system for data files combining features of, e.g., the National Technical Information Service (NTIS) or the Education Resources Information Center (ERIC). This system should provide standard bibliographic descriptions, abstracts of the content of the files, and copies of codebooks (in hard copy or microfiche) on request. Moreover, it should provide "deep" indexing of the abstract (ten or more descriptors) and should be machine-searchable. No data supplier except NTIS now offers a system this complete. Yet the copy needed to implement it – abstracts and codebooks – already exists in abundance, and unlike full data files, *is* freely diffusible. A case for this system will be made below, after a sketch of several issues in data file cataloging.

• A Base for Policy

The fundamental paper on bibliographic policy for machine-readable social data appeared in 1972, about a decade after the movement to save the data of polls, censuses, and surveys for new users can be fairly said to have begun. In its authorship, that paper symbolized the convergence of two groups that had been separate through most of the decade: it was by a librarian, John Byrum, and a data archivist, Judith Rowe, and it was the first detailed account of a bibliographic information system for data files to appear in the library press (*1*). It was also one of the earliest writings to give librarians a specific task in documenting data files. In the 1960's, social scientists in the data archive movement had sometimes speculated on the possibility of finding a stable home for their archives in conventional research libraries, and they had written of what librarians would immediately perceive as matters of bibliographic control: the indexing of statistics or of response data from polls and surveys, and the cataloging and classification of whole data files (*2*). But the social scientists in the movement published their writings in their own press, and generally seemed unable to relate their concerns to actual library practices and tools. Byrum and Rowe's paper thus represented a fresh start.

Their major contribution, influential here, was to define the levels of documentation that are needed to serve different user requirements. Persons who want to know whether a specific data file (or a file with certain subject matter) is locally available need *catalog entries* by author, title, and subject in local libraries. Persons who want data on specific variables need both *abstracts* and *codebooks.* (Abstracts, of course, provide condensed versions of study content, while codebooks reveal it in full.) Finally, persons who want actually to use the data in a computer need *physical descriptions* of the medium on which the data are stored – for example, the characteristics of a particular magnetic tape. [See also

Conger (*3*) on these levels, particularly her lucid discussion of the last.]

• Breakthrough in Cataloging

Since Byrum and Rowe wrote, a great change has occurred in librarians' capability to describe machine-readable data files for their users. Working since 1970, a subcommittee of the American Library Association's Catalog Code Revision Committee has produced standards for writing a new set of Anglo-American cataloging rules specifically for this medium (*4*). With the next edition of the *Rules,* librarians will have directions for cataloging data files – directions based on the same principles as those for other sorts of materials; and this will almost certainly cause catalogers at major research libraries, including the Library of Congress, to view such files as part of their domain. (Major data suppliers could confirm this view by submitting codebooks to the Library of Congress for cataloging before they are released, and reproducing the LC entry as part of the codebook, in the mode of Cataloging in Publication.) Almost no libraries now record the presence of, say, Almond and Verba's machine-readable "Five Nation Study" on campus; that is left to the local data archive or computing center (if it is done at all), while the conventional library contents itself with producing a catalog card on the book that emanated from the data file–in Almond and Verba's case, *The Civic Culture* (*5*). But this state of affairs will change, on some campuses at least, so that we may expect to find catalog entries for locally held machine-readable materials interfiled with entries for human-readable items – journals and books.

There are small but clear advantages to having entries for data files in the library catalog even if a local data archive also maintains records of holdings. The conventional library will very likely be open longer hours and more days of the week than the local data archive, and will serve to introduce data holdings to a wider clientele. Most campuses, of course, lack a data archive altogether, but even where such archives exist, they are typically used by persons who are already knowledgeable about machine-readable materials. The conventional library, in contrast, may permit unknowledgeable persons to discover such files by lucky accident. Furthermore, by catalog filing rules, entries for items with the same author, title, or subject headings are brought together. Thus entries for printed books will for the first time be complemented with entries for associated source data. For many persons, this will reveal the full magnitude, hitherto unsuspected, of certain social science projects. The second *World Handbook of Political and Social Indicators,* for example, might be revealed to exist both as a reference book on nations and as a family of data files from which tables not in the book can be generated (*6*). Occasionally someone might even be saved from a laborious keypunching job by discovering in the library that serviceable data already exist in machine-readable form.

• Subject Headings: A Problem

It must not be forgotten, however, that the new Anglo-American rules will be rules for descriptive cataloging only. They are essentially directions to copy certain elements (such as title or author) from the file documentation and to order them in a certain way in an entry. Where rules for copying scarcely exist, as is the case with assignment of subject headings, no breakthrough comparable to the one in descriptive cataloging has occurred. Use of the recent eighth edition of the *Library of Congress Subject Headings* according to Haykin's 1951 guidelines (*7*) can still cause difficulty. For example, many machine-readable data files derive from "omnibus" social surveys, in which respondents are asked questions on a great variety of topics. One suspects that subject catalogers (at the Library of Congress or elsewhere) will follow Haykin on "specificity" and assign the heading SOCIAL SURVEYS (perhaps with a geographic qualifier like "United States") to characterize these. (See Fig. 2, the text of "specimen copy" on the study in Fig. 1.) Yet the LC scope note on SOCIAL SURVEYS shows that this heading is already used in an ambiguous way, to cover both works on the general methodology of sample survey research and the finished reports of particular surveys. To apply the heading also to data files – that is, to codebooks and magnetic tapes (or punched cards) – merely confuses things further.

The main problem, however, is not the ambiguity of terms like SOCIAL SURVEYS; it is their overgenerality. This is indeed an intractable problem, because as long as we use an entire data file as our unit of analysis in subject indexing, we are forced to seek highly general terms to cover the multiplicity of variables we find. Yet terms like UNITED STATES–SOCIAL CONDITIONS are so broad as to be nearly useless to the searcher who wants data on a particular variable – say, average annual earnings of persons in various religious denominations. The data may actually be in the file, but the heading UNITED STATES–SOCIAL CONDITIONS hardly brings that fact out.

Almost always the person looking for data will want indexing that leads to particular variables. Two or three broad subject headings applied to the whole data file, in the style of LC subject cataloging for books, usually will

```
Davis, James Allen.
   National data program for the social sciences:  Spring, 1972
general social survey [Machine readable data file]  Chicago,
National Opinion Research Center.  Distributed by Roper Public
Opinion Research Center, 1972.
   1613 logical records.

   Title from accompanying codebook.
   Called also 1972 NORC general social survey.
   Size of file not verified.
   Summary:  Survey of national cross-section of 1613 adults who
answered 61 questions covering such topics as social stratification,
the family, race relations, social control, civil liberties, and
morale.

   1.  Social surveys - United States.  2.  United States - Social
conditions.  I. National Opinion Research Center.  II. Title.
III. Title:  1972 NORC general social survey.
```

Fig. 2. Sample catalog copy for a machine-readable data file, the survey shown in Fig. 1. Adapted from Jean Riddle Weihs et al. *Nonbook Materials; the Organization of Integrated Collections.* Ottawa: Canadian Library Association; 1973.

convey little more than some librarian's rough sense of the "intended audience" for the data – sociologists, economists, or whoever. It is therefore extremely important that any library which maintains a subject catalog on data files should also keep codebooks accessible. If the searcher cannot move from the subject catalog to a codebook with little delay, the whole point of the catalog will have been lost.

Conventional research libraries will also help data searchers if they provide codebooks for files *whose machine-readable part is not held on campus.* In a study I made in 1974, codebooks without the associated tapes or cards accounted for roughly half the sales in transaction samples drawn at the Roper Center and the International Data Library and Reference Service, both of which have national clienteles (*8*). Researchers evidently do want codebooks on hand for reference, regarding them like any other reference work that points to resources elsewhere. Certainly libraries already have many of those.

• A New Form Class

On the national level, the Library of Congress could refine the precision of terms available to subject indexers by authorizing a change in the LC Subject Headings. What is needed is a new *form* subdivision, such as –CODEBOOKS that could be appended to subject headings proper, as we now can append –BIBLIOGRAPHIES or –PERIODICALS or many other designators of form as opposed to content. There is, in the current Library of Congress list, a form subdivision called –COMPUTER PROGRAMS, but none for data files or codebooks, which are form classes highly different from both substantive writings and statistical compilations. The effect of using a subdivision like –CODEBOOKS would be that entries for files of raw numeric data would be inserted after entries for substantive writings on a given topic, in a separate group of their own.

I do not argue that failure to do this will result in not being able to tell data files and substantive writings apart, because under the new rules for descriptive cataloging, the designator "Machine Readable Data File" will appear in brackets after titles whenever it is appropriate. It is simply that, without a form subdivision like –CODEBOOKS, entries on data files will be scattered through scores or even hundreds of entries on substantive works given headings such as SOCIAL SURVEYS or UNITED STATES–SOCIAL CONDITIONS. A person manually searching for data files under the present Library of Congress system would have to search entire blocks of entries to make sure of not missing any, whereas by the simple expedient of making data files a separate and explicit form class, they would be "broken out" in either a subject catalog in book form or a library card catalog. This, of course, would make subject searching more efficient.

• A National Union Catalog?

The last issue to be raised in this section is the value of a national union catalog of data files–a likely product as soon as a number of libraries put the new rules for descriptive cataloging to use. If research libraries make original catalog entries for data files locally held, or report ownership of data files for which entries have

already appeared, we have the essential input for the present *National Union Catalog* published by the Library of Congress, and also for such union catalogs as that available on-line from OCLC, the Ohio College Library Center. Initially, entries for data files would probably be interfiled with others in the present printed version of the *National Union Catalog,* which now covers books, maps and atlases, and various serial titles. But in time it is possible that enough reusable data files will be produced and cataloged to warrant a separate national catalog of their own, just as printed music, phonorecords, manuscripts, microform masters, and film products do now. And in fact some people in the data archive movement are already discussing such a catalog (*9*), which would fulfill a dream of members of the now-defunct Council of Social Science Data Archives, who wanted, but never got, a "national inventory" of data files.

Let us be clear on what a national or regional union catalog of machine-readable materials would do. Some library would make an entry on, say, the "Five Nation Study," based on its codebook; the entry would give authors and title, imprint, number of logical records, and so on. This original cataloging would be reproduced in the union catalog, and no other library would have to do original cataloging; it could simply obtain copies of the entry as needed. Presumably it would also report to union catalog headquarters that it, too, had the study locally available, and this fact would be publicly recorded—either as an appendage to the original cataloging or in the national *Register of Additional Locations.* Thus a determined searcher who was unable to find the study on his own campus could find other campuses where it was held.

But here the value of the union catalog for data files becomes doubtful. Besides allowing one library's original cataloging to be shared, a union catalog is intended to facilitate interlibrary loans. If some searcher wants a book or an article that is not locally available, a librarian will often be able to borrow a copy from a holder listed in a union catalog. With many important social science data files, however, that is not the case. The two major academic suppliers of data files, the Interuniversity Consortium for Political and Social Research (ICPSR) and the Roper Public Opinion Research Center, ask considerable annual subscription fees for their services, and they prohibit their customers from "rediffusing" data files to nonsubscribers. ICPSR and Roper in effect claim copyright on a fair number of the most interesting files, so as to protect their large financial investments in making the files available in the first place. Thus if I am at Temple University in Philadelphia, and I find in a union catalog that a certain machine-readable Gallup Poll that I want is held by the University of Pennsylvania in the same city, that does not mean I can arrange to have a copy supplied through interlibrary loan – not if the wishes of Roper, the purveyor of Gallup Polls, are carried out. The same holds true for the "Five Nation Study," sold by ICPSR.

• Further Doubts

The picture is complicated because not all attractive data files are controlled by Roper and ICPSR: some are available from the originators or other data suppliers, and another large group is produced and distributed by the US government. (The National Technical Information Service is the chief distributor.) But even assuming these are diffusible from campus to campus, in the manner of interlibrary loan, the process still is not as cheap or as straightforward as slipping a book in a bookbag or mailing a Xeroxed article. It involves the creation of administrative liaisons, as yet only imaginary, between conventional libraries, local data archives, and campus computing centers, and a way of passing on to customers charges that will exceed anything they are used to in the present interlibrary loan system. Also, in some cases the originator may impose conditions on diffusion of a file that the loaning library will have to meet.

There is yet another doubt about the value of a union catalog of data files. For the searcher who already knows the content of a particular file (or who can easily get a codebook), such a tool might occasionally be useful.* But for the searcher not familiar with a file and without a codebook, the information on a standard library catalog entry will almost certainly prove too meager to permit a decision on retrieval (or purchase). If we are designing a tool to help searchers decide what data files to obtain, we might as well design one that is adequately informative. By itself, a union catalog would not be.

One might argue that at least a union catalog would serve to standardize and publicize descriptions of data files for a wide clientele, and that such "consciousness-raising" is a subtle good in itself. But this argument does not seem compelling enough to give priority to the creation of a union catalog. The same advantage, and many others, could be realized with a national current abstracting service for data files, and it is this form of "national inventory," rather than what the union catalogs of OCLC or the Library of Congress imply, that should receive first priority.

*For instance, a person who already had a copy of a file might use a union catalog to find whether some other institution in the area had the file also. If so, it might be reasonably easy to find someone who could discuss problems in processing the file, give advice on it, supply additional documentation, and so on. (I am indebted to Judith Rowe for this example.)

• The State of Abstracting

What follows is not an argument for change in the economic arrangements for distributing data files themselves. I assume that ICPSR, Roper, US government agencies, and other suppliers will go on selling them to customers as they do now, and that in general we should not try to bring data files under the interlibrary loan system. My concern is with documents that, unlike relatively costly machine-readable materials, *are* freely diffusible—namely, abstracts and codebooks. It seems probable that organizations like ICPSR and Roper would welcome wider diffusion of their abstracts and codebooks than they now have, since it is in their long-term interest to advertise their wares, to attract more customers for machine-readable items. My remarks, therefore, can be taken as a critique of the present system for distributing abstracts and codebooks, the relatively "circulable" items that data suppliers have to offer. For a similar, though less specific critique, see Peters (*10*).

The present situation is one in which a fair amount of abstracting copy is written, but in different styles and lengths, and then scattered through various publications that are usually undiscoverable through regular bibliographic channels. (Using tools in a library, try to find ICPSR's *Guide to Resources and Services,* or the Project TALENT *Handbook,* or *ss data*, or the Roper *Newsletter*, even when you know they are present on campus.) If these items are in fact present on campus, they will typically be in a small data archive that is underpublicized and unknown outside a particular circle of researchers and students. The same is true also of codebooks, the document one wants to see after reading an interesting abstract (or catalog card). But codebooks, too, are rarely discoverable through formal bibliographic channels (*11*). The present system strongly favors insiders in particular research coteries.

• National and On-Line

The most promising means for improving the situation is to put abstracts of data files, and instructions for ordering their codebooks, into an existing, national, on-line bibliographic service. Both ERIC and NTIS should be considered, and perhaps others as well. An existing service is preferable because data files are not so numerous or so in demand as to justify the creation of a new service just for them. Nor should we have to develop new software if existing packages are readily adaptable to our documentary aims. NTIS in fact already provides a complete prototype of the service advocated here. It publishes abstracts of data files both in *Government Reports Announcements* and in its on-line search service; it also sells codebooks (in hard copy or fiche) and the data-bearing tapes themselves. However, its mission is to sell the files made public by governmental agencies. I shall therefore use ERIC as my model in what follows because of its receptiveness to materials from nongovernmental sources (e.g., academic or commercial researchers). While some other service might eventually prove better, ERIC is specifically intended to publicize "educational resources," which many data files are, and its *Thesaurus* of indexing terms is already rich in social science topics. Its *Thesaurus* is also open to fairly rapid growth as new terms become needed. The tapes produced by ERIC, moreover, are obtainable (if one chooses) in MARC II format, which means that bibliographic information originally gathered for ERIC can be put to any use to which MARC records are put.

The data archive movement and the automated bibliographic searching movement grew up apart, and it is perhaps not surprising that the data archivists still are cut off from the services that deliver both abstracts and copies of the report literature in various fields. It is time to recognize, however, that a codebook is a document that can be delivered exactly like a report. And it is time to recognize that an abstract of the study that the codebook reflects is deliverable like thousands of other abstracts published in ERIC and elsewhere. A strong case can be made that data files are as deserving of national publicity as unrefereed reports; in many instances, more so. Means of discovering their existence and nature should not be confined to a small group of initiates on each campus.

I am, of course, recommending that abstracts such as those in *ss data,* the Roper *Newsletter, Computers and the Humanities,* and the ICPSR *Guide,* be disseminated by ERIC (or a comparable system). It would be desirable, too, to have codebooks available in inexpensive microfiche through ERIC: that would mean not only that individuals could get them cheaply, but that libraries which have standing orders for ERIC fiche would get codebooks automatically. However, if organizations like ICPSR and Roper want to retain codebook selling privileges to themselves, it still would be possible for them to publish abstracts of their studies through ERIC. Many ERIC abstracts carry the message, "This document not available through ERIC Document Reproduction Service," and then give the address of the supplier—which could be, for example, that of Roper or ICPSR. If these and other data suppliers wanted to continue to issue their own catalogs, abstract bulletins, and so on, they could do so; the ERIC dissemination would merely complement and augment their present arrangements.

• Main Gains Considered

What would be the main advantages for data archivists in going to something like ERIC? The first is that they would gain a bibliographic information system that has far more outlets than their own, and that reaches a large and diversified group of people who now are unaware that reusable data exist. Both Roper and ICPSR have organized their sales so that annual subscriptions by institutions are their chief financial support. Within departments of these institutions there will be persons knowledgeable about reusable data; in other departments on the same campus, ignorance usually reigns. In institutions not now subscribing to either Roper or ICPSR, the ignorance may be total. ERIC could occasionally tell persons in nonsubscribing institutions or uninformed departments that data files are available. It could bring descriptions of these files to students doing literature searches in preparation for writing dissertations, to educational resource persons in public school systems, to teachers in junior or community colleges, to researchers in professional schools, to librarians, to information specialists in business and industry, and so on through the ranks of those who are not now purchasers of data. The objection that such persons lack sufficient computer skills to reanalyze others' data is no longer persuasive, in view of the widespread emergence of SPSS, the Statistical Package for the Social Sciences, as a "people's package" of computer routines that anyone can learn to use in an afternoon.

The second major advantage is the benefit conferred by an on-line system. Abstracts of data files, and instructions for ordering codebooks (and possibly the complete files), would be deliverable to any terminal linked to the computers of the System Development Corporation, Lockheed, or Bibliographic Retrieval Services (all supply ERIC). This of course means virtually anywhere in the United States, and in some foreign countries as well. Moreover, under ERIC, the abstracts would be indexed, which would make them deliverable as a result of on-line subject searches. It would be possible to specify that one wanted only data files on some subject by means of a form descriptor (e.g., DATA FILE, already a standard NTIS "keyword," as shown in Fig. 3). But it would also be possible, if the form descriptor were not used as a search term, to have abstracts of data files turn up with abstracts of report and journal literature, simply because they carried the same subject indexing. This would be one means whereby persons who were unaware of the existence of data files could come on them serendipitously. In either case, the capability of retrieving abstracts of files on the same subject from more than one archive—Roper, ICPSR, other academic and governmental suppliers—would be a very marked improvement over anything we now have.

• Useful Categories

A third advantage lies in the amount of information provided by the typical ERIC entry—an amount greater than that of a library catalog entry. The present ERIC categories would accommodate author's name, file title, organization that gathered the data, sponsoring organization (if any), supplier's ID number for the data file, publication date, codebook pagination, discretionary notes on such things as related studies (principal publications from the data might be put here), address of source for obtaining the codebook (or full file), and price for obtaining the codebook (not the full file) in hard copy or fiche from ERIC. To these categories could be added another already in use by NTIS: physical characteristics of the tape(s) on which the data are stored (see Fig. 3).

Note that, as here envisioned, the categories orient the customer toward buying the full machine-readable file from the originator or a supplier, rather than expecting to get it by interlibrary loan. Codebooks, on the other hand, would be available by purchase from the data suppliers and perhaps from ERIC, by loan from a local data archive or research library, and by interlibrary loan. (Publications based on data files would also be available from libraries, of course.) This division of responsibilities would probably combine smoothly with existing service and economic arrangements, while

CEN/DF-73/110

National Travel Survey - 1972.

Data file. The file shows the volume and characteristics of travel by residents in the U.S. Data show estimated number of households in which some household member took one or more trips, persons who took at least one trip, person-nights, and person-miles. Data are shown by such travel characteristics as means of transport, purpose of trip, duration of trip, distance, size of party, type of lodgings, characteristics of traveler, etc.
Publications: U.S. Census of Transportation: 1972, National Travel Survey.

Keywords: *Data file, *Travel, *Passenger transportation, *Routes, *Statistics, *Households, *Trip generation.

Geocoding: United States.

Availability: Bureau of the Census, Data User Services Office, Washington, D.C. 20233, 2 reels mag tape, $70.00/reel. IBM compatible, 7 track, 556 or 800bpi, Bcd, odd or even parity; or 9 track, 800bpi, Ebcdic, odd parity.

Fig. 3. Sample entry from National Technical Information Service. *Directory of Computerized Data Files, Software & Related Technical Reports.* Springfield, VA: NTIS; 1976. The format differs somewhat from that of data file entries in *Government Report Announcements* and on-line NTIS.

greatly increasing the availability of codebooks, the key document in reusing others' data.

Under the ERIC system, some 10 to 15 indexing terms would be used to characterize file content, as opposed to two or three (at most) on the typical catalog entry with LC subject headings. ERIC's indexing terms include straight subject descriptors; form class names like BIBLIOGRAPHIES, DIRECTORIES, and GUIDES; and so-called identifiers—terms not in the *Thesaurus*—which could be used (for instance) to show the geographic locus of survey data (e.g., India or Detroit; as shown in Fig. 3, NTIS reveals the geographic locus of data in a special "geocoding" category) or the unit of analysis surveyed (e.g., California cities, OECD nations, African students).

To judge the appropriateness of ERIC subject indexing terms for social science data files, consult the *Thesaurus of ERIC Descriptors* and its updates in monthly issues of *Resources in Education.* It will be seen that ERIC terminology is not confined to the jargon of professional educators, as some might think. Rather, it is a general purpose indexing language, capable of expressing any topic on which education might take place.*

• Indexing Questions and Files

The categories of bibliographic information discussed thus far are followed in ERIC by substantial abstracts. In the abstract of a study involving a sample survey, one could put all those things that data archivists agree are desirable, including number of respondents, sampling techniques, survey methodology, and question content. The Roper Center has an interesting technique of showing the latter by stringing together names or brief designations of every question asked in a survey. A hundred variables can be revealed in surprisingly little space this way. If all abstracts of survey data files were written in this fashion, we would be able to do machine searches not only on the 10 or 15 broad terms characterizing an entire file, but also on every single designation of a variable in the text of the abstract. Full-text searching, in which any term or string of terms in the entire abstract is potentially retrievable, is a reality now in on-line ERIC. Thus, if abstracts were prepared in the right way, we could immediately begin to realize another dream of the data archivists: machine retrieval of variables rather than of file titles alone. As noted above, persons looking for data are usually interested in whether particular questions (or variables) are present in a study, and merely by drawing on existing capabilities, the data archivists could deliver such information wherever ERIC printed or on-line products appear.*

Figure 4(a) is a mock-up of an ERIC entry for the study shown in Figs. 1 and 2. The body of the abstract is given in two different possible styles—Fig. 4(b), a general characterization of study content; Fig. 4(c), a question-by-question analysis in the manner of the Roper Center. The two styles can be contrasted somewhat like those called "indicative" and "informative" for abstracts of papers. There seems little doubt that the question-by-question (or "full-variable") analysis is preferable, although more costly.

• Current Awareness

Let me mention one last advantage that ERIC would have over present arrangements for delivering abstracts of data files. Its printed *Resources in Education* and its tapes appear in new issues monthly. In contrast, the existing announcement literature for (non-NTIS) data files is made up of quarterlies, annuals, and irregulars.

*The *Thesaurus of ERIC Descriptors* already contains exact equivalents or near synonyms of the great majority of indexing terms used in three subject catalogs of data files against which I experimentally checked it: the Substantive Index of *A Guide to Resources and Services, 1975-1976,* Interuniversity Consortium for Political and Social Research; the Category Index of *Survey Data for Trend Analysis; An Index to Repeated Questions in U.S. National Surveys Held by the Roper Public Opinion Research Center,* 1975; and the Subject Index of the *Stanford University Data File Directory,* 1973. There are exact *Thesaurus* equivalents for more than half the terms in these catalogs, and, depending on one's sense of synonymy, one can approximate roughly another quarter of the terms. An area in which the *Thesaurus* is presently rather weak, however, is the terminology of international and cross national political studies.

*To illustrate, let us suppose that a survey available from ICPSR and another survey available from the National Opinion Research Center both contain questions on the respondent's sex, membership in a union, and attitude toward the computer in the workplace. Someone interested in how union members felt about "automation" and whether male members felt differently from female members, might retrieve the two surveys if the abstracts of each used such terms as "sex," "union membership," and "computers" to designate these particular variables. The searcher would not retrieve full question text, of course—that would be found only in the codebook, and so the searcher would not know immediately whether the seemingly like questions he had retrieved were truly comparable. There are enormous problems to be faced in developing standard designations for comparable questions, and applying them successfully over time and across different data archives. Without minimizing these problems, it is still exciting to consider the prospect of full-text searching of question-designators in abstracts.

ED 123 456 SO 654 321
Davis, James A.
National Data Program for the Social Sciences: Spring 1972 General Social Survey [Machine readable data file]
National Opinion Research Center, Chicago, Ill.
Spons Agency--National Science Foundation, Washington, D.C.
Pub Date July 1972
Note--Codebook, 122p. Magnetic tape, 1 reel, IBM compatible, 7 track, 556 or 800bpi, Bcd, even parity; or 9 track, 800bpi, Ebcdic, odd parity.
Available from--Roper Public Opinion Research Center, Williamstown, Mass. 01267. Codebook ($1.50); magnetic tape ($70).
EDRS PRICE (Codebook only) MF-$0.75 HC-$6.60 plus postage
Descriptors--Abortions, Authoritarianism, Civil Liberties, *Codebook, *Cross-sectional Studies, *Data File, Family Background, Family Structure, Morale, Race Relations, Sex Discrimination, *Social Attitudes, *Social Values, Socioeconomic Status
Identifiers--United States

(a)

1613 respondents, 2 cards of data per respondent. The data were collected by the National Opinion Research Center as the first in a five year series of general social surveys. The survey was administered in February-April 1972 to a national cross-section sample of adults 18 years of age and older. In addition to the standard personal characteristic items, the survey covers items viewed by the NORC staff and an advisory panel of sociologists as "mainstream" interests of modern academic sociology. The interview covers the areas of stratification, the family, race relations, social control, civil liberties, and morale. A major objective of the project was the replication of questions that have appeared in previous national surveys.

(b)

Sample size: N = 1613 (2 records per respondent). Summary: plans for family increase; ideal number of children; married women working; woman for President; abortion; death penalty; gun permits; court treatment of criminals; premarital relations; censorship; race relations; Negro for President; busing; trust in people; personal health; job satisfaction; personal finances; happiness; vote recall; newspaper reading habits; party preference. Demographic data: residence at 16; father's occupation; income; ancestry; brothers and sisters; marital status; employment; work week; spouse's occupation; number of persons in household; family income; number of children; religion; church attendance; region; age; sex; race; education of respondent's parents, spouse, and self; social class.

(c)

Fig. 4(a). Simulated ERIC entry for the 1972 General Social Survey. Some of the data, and their placement in certain fields, are hypothetical. (b) Abstract for the 1972 General Social Survey, adapted from *A Guide to Resources and Services, 1975-1976,* Interuniversity Consortium for Political and Social Research. This abstract is based on the preface to the codebook. (c). "Roper-style" abstract for the 1972 General Social Survey, adapted from *Roper Public Opinion Research Center Newsletter.* 1972 December; 7(2).

Thus ERIC is superior for current awareness.* And even though the *printed* abstracts of data files would be scattered through the monthly issues or semiannual cumulations of *Resources in Education*, the *on-line* ERIC file would hold a total cumulation of abstracts; the whole corpus would be permanently available for searching. If this corpus grew to reflect something like the total collection of reusable data files available to the American public, we would have that "national inventory" that has so long eluded us.

Whether ERIC will disseminate bibliographic information about data files is a matter to be decided by its governmental sponsor, the National Institute of Education. If the decision is favorable, it may be possible to submit codebooks and abstracts to one or more of ERIC's 16 clearinghouses for input into the national distribution system. (A candidate is the Clearinghouse on Social Studies/Social Science Education in Boulder, CO.) Standards for writing codebooks and abstracts, and for indexing the abstracts, are as yet underdeveloped, but even so, much copy already exists and could probably be used as is. The initiative for setting documentation standards and for moving suitable copy into a national on-line system now rests with the information wing of the social science community. In addition to special interest groups in the American Society for Information Science and the Assocation for Computing Machinery, this wing includes the new International Association for Social Science Information Service and Technology. The latter group, comprising many data archivists, seems currently in the best position to seek advances in the bibliographic information system for data files.

References

1. **Byrum, J.D.; Rowe, J.S.** 1972. "An Integrated, User-Oriented System for the Documentation and Control of Machine-Readable Data Files." *Library Resources and Technical Services.* 1972 Summer; 16(3): 338-346. A notable earlier discussion of cataloging "magnetic tape data bases" (though not specifically numeric ones) appeared in **Hayes, R.M.; Becker, J.** 1970. *Handbook of Data Processing for Libraries.* New York: Wiley-Becker-Hayes; 1970. 716-723.
2. Cf. **White, H.D.** 1974. *Social Science Data Sets: A Study for Librarians.* Ph.D. dissertation. Berkeley, CA: University of California; 1974. 2-62.

*ERIC's monthly *institutional* index, by the way, would let us see whether a particular data supplier—for example, the National Opinion Research Center or ICPSR or the Louis Harris Data Center in North Carolina—had recently brought out new data files. Having the output of these suppliers jointly under one cover would ease acquisitions work.

3. **Conger, L.D.** 1976. "Data Reference Work with Machine Readable Data Files in the Social Sciences." *Journal of Academic Librarianship.* 1976 May; 2(2): 60-65. This article is a good introduction to some widely used numeric files and bibliographic tools.
4. **Herman, E.; Byrum, J.,** eds. 1976. *Final Report of the Catalog Code Revision Committee Subcommittee on Rules for Cataloging Machine-Readable Data Files.* Chicago, IL: American Library Association; 1976. (ED 119 727)
5. **Almond, G.; Verba, S.** 1963. *The Civic Culture: Political Attitudes and Democracy in Five Nations.* Princeton, NJ: Princeton University Press; 1963.
6. **Taylor, C.L.; Hudson, M.C.** 1972. *World Handbook of Political and Social Indicators.* 2d ed. New Haven, CT: Yale University Press; 1972. The five data files associated with this book are indexed under Taylor and Hudson in *A Guide to Resources and Services 1975-1976,* Interuniversity Consortium for Political and Social Research. Documentation on the files can also be found in on-line NTIS, e.g., by an author search.
7. **Haykin, D.J.** 1951. *Subject Headings; A Practical Guide.* Washington, DC: Library of Congress; 1951. See the discussion, pp. 9-11, which begins: "The heading should be as specific as the topic it is intended to cover."
8. See (2), pp. 131-136. The latter data archive mentioned is in the Survey Research Center, University of California at Berkeley.
9. E.g., **Dodd, S.A.** 1977. "Cataloging Machine-Readable Data Files—A First Step?" *Drexel Library Quarterly.* 1977 January; 13(1): 48-69. (In particular p. 58.)
10. **Peters, P.** 1974-75. "Describing a Social Science Data Information System, Networks and Components," *SIGSOC Bulletin; A Quarterly Publication of the Special Interest Group on Social and Behavioral Science Computing.* [Association for Computing Machinery] 1974-75 Fall-Winter; 6(2/3): 6-25. This issue of the *Bulletin* has seven articles on documentation of numeric files in the social sciences.
11. An exception is provided by Stanford University, where codebooks are entered in the library catalog and assigned LC call numbers for placement in the library stacks. They are also listed in the *Stanford Data File Directory,* which is available in both printed and on-line versions on campus. See **Ferguson, D.** 1977. "Social Science Data Files, the Research Library and the Computing Center." *Drexel Library Quarterly.* 1977 January; 13(1): 70-81.