Citation Analysis of Data File Use

HOWARD D. WHITE

Introduction

AN ARGUMENT BY NO means new is that social scientists who work with machine-readable data files (MRDF) should cite them in their writings, with formal references set apart from main text, just as they now do books, papers and reports. Large-scale suppliers of the files urge this so that their de facto role as data publishers will be properly credited—for nonprofit organizations, an important form of reward. Data librarians have urged it on bibliographic grounds: a data file that is properly identified in a citation, and not just vaguely alluded to in the text, is easier to track down. Moreover, since the advent of *Social Sciences Citation Index* (SSCI) in 1972, a few commentators have urged it on grounds that such citations—at least those in the journal literature—would be picked up by SSCI and would constitute a "use-history" of data files of great potential interest to all who perform or promote secondary analysis. This paper addresses the last concern—the current state of affairs with respect to citation indexing of data files in SSCI. To anticipate a bit, the situation is chaotic, but not without possibilities for improvement.

A hypothetical example will show how SSCI works and why the inclusion of data files among the cited documents in it is of importance to researchers. The file to be used for illustration is one that I recently cited in a paper of my own:

Howard D. White is Associate Professor, School of Library and Information Science, Drexel University, Philadelphia.

Temple University, Institute for Survey Research. *Opinion Survey on Current Social Issues* [machine-readable data file]. Philadelphia: The Institute [producer and distributor], 1970.

This is a survey of 2486 adult Americans conducted in early 1970 for the Commission on Obscenity and Pornography; its results were discussed a decade ago in the commission's *Report,* and mentioned as recently as 1979 in Gay Talese's *Thy Neighbor's Wife.* The original analysis of the file was carried out by Response Analysis Corporation of Princeton, New Jersey, and reported in a technical monograph published by the commission as a U.S. government document, *National Survey of Public Attitudes Toward and Experience with Erotic Materials.* But the file has also been used in several secondary analyses reported in the journal literature, and if these papers had cited the file more or less as above (they do not), the Citation Index of SSCI would have routinely picked them up with some such entry as this:

Temple U, I Surv Res
Opin Surv Curr 70

Wilson WC	J Soc Issue	29	19	73

and later, under the same heading:

Wilson WC	J Sex Res	11	46	75
Wilson WC	J Soc Issue	31	69	75
Glassman MB	Pub Opin Q	42	161	78

My paper would recently have been added to the chain:

White HD	Library Q	51	192	81

One translates these highly condensed entries into full bibliographic listings in SSCI's Source Index; for example, the paper in which Glassman uses the "pornography survey" is, in full:

Glassman, Marc B. "Community Standards of Patent Offensiveness: Public Opinion Data and Obscenity Law." *Public Opinion Quarterly* 42(1978):161-70.

But the point is that ideally one can trace the use-history of this file by examining the chain in the various issues of SSCI; or, if one has access to its online version, Social Scisearch, the entire history is available in cumulated form simply by inputting the name of the pornography survey file.

Obviously such cumulated histories should interest data suppliers, like the Roper Center or the Inter-University Consortium for Political and Social Research, since they reveal the use of their offerings in published papers (even if an ill-bred citer fails to mention them as distributors by name). One would expect social scientists to be interested even more, since the use-history registered in SSCI reflects the formation of an identifiable community (of sorts) around a data file. For various reasons (curiosity, rivalry, etc.) geographically dispersed users of the same data want to know of each other's work. Use-histories thus have considerable human interest value, and the value may be intensified by the highly specialized nature of much secondary analysis. Only a relative few will work with a special-topic file like the pornography survey, or with a particular set of questions (such as those on abortion) in an omnibus file like the General Social Survey. One wants to know who and where they are, what technical problems they encountered, and especially what their findings are. Citation indexing can lead directly to answers to these questions in a way that conventional subject indexing does not.

What, then, is the case: are use-histories of data files available at all? The answer, not widely known, is that citations to data files do appear in SSCI and have for some years. This is not to say that large numbers of social scientists are taking to heart the counsel of such writers as Rowe[1] and White,[2] and citing files in the style recommended by Dodd.[3] They are not. But a fair number of researchers have in one way or another acknowledged use of files—particularly codebooks—in their footnotes or endnotes, and this has been sufficient to leave at least partial use-histories in SSCI (and in Social Scisearch). The rub lies in the phrase "in one way or another." The lack of a consistent citing style, in combination with editorial practices of SSCI's publisher, the Institute for Scientific Information (ISI), has resulted in a rather spectacular scattering of the citations to any given file, and only the most determined labors of reassembly—i.e., checking SSCI at many different points—will produce a coherent use-history such as the one above.

The Causes of Fragmentation

There are several ways in which this scattering, or fragmentation, comes about. Basic to the problem is that entries in SSCI are keyboarded directly from the texts of papers in journals, with little or no editorial intervention to correct discrepant citing practices. Then the entries are automatically filed by computer, which is not programmed to reconcile two citations to the same work if they include different elements or begin in different ways.

A major unreconciled difference occurs when researcher A cites a data file with the author as the first element, and researcher B omits the author and cites it with title first. This has the effect of throwing the two references into wholly different sections of SSCI's Citation Index. A file cited with title first goes into the section reserved for anonymous works, which is something of a bibliographic slum (cf. the opinion of Garfield, ISI's president[4]). If author is put first, the reference of course goes into an author section, but there are two of these—one for personal authors and another for corporate authors. Unfortunately, many MRDF can be cited by either type of author, and this is where another unreconciled difference occurs. If the file was created in a project with a principal investigator (PI), and if the citer puts the PI—a person—first, the reference will be placed with all the other personally authored works (papers, monographs, etc.) that make up the bulk of the SSCI Citation Index. But if the citer omits the PI and puts a producing or distributing corporate body in author position (as I did with the pornography survey), the reference will go into the corporate author section. (Occasionally, too, a corporate author entry is shunted to the anonymous section of SSCI by mistake.) Thus three different researchers who had worked on the same file, perhaps even on the same set of variables, could find their identically intended citations in three separate sections of SSCI, depending on their choice of first element in citing. Their citations would also be placed in three separate parts of the Cited Reference index to online Social Scisearch.

Further fragmentation occurs within each of the sections. Citers who choose the same first element in their citations often differ in the ways they record titles, or personal or corporate authors. One very common type of fragmentation in SSCI occurs with cited personal authors: researcher A cites by surname and first name (or first initial); researcher B cites by surname, first name and *middle* name (or initial). As transcribed and computer-filed in SSCI, this causes citations to the same work—say, by James N. Morgan—to be entered in two different places, as the arrows show:

Morgan J
←
Morgan JA
Morgan JB
.
.
.
Morgan JN
←

Thus, even in the relatively simple case of a principal investigator as first element, one must look in two places to avoid missing all citations.

With corporate authors as first element, the fragmentation is much worse. Librarians over the years have devised elaborate rules for dealing with corporate authors in card catalogs, but even they have had trouble in achieving consistency, and have on occasion changed the rules. Pity then the citers and journal editors: they may follow style guides, but the overall result is bibliographic anarchy. Take, for example, this nonexhaustive list of ways to render the U.S. Census Bureau as author of a file:

Bureau of the Census
Census Bureau
Department of Commerce, Bureau of the Census
Department of Commerce, Census Bureau
U.S. Bureau of the Census
U.S. Census Bureau
U.S. Department of Commerce, Bureau of the Census
U.S. Department of Commerce, Census Bureau

The same data file—e.g., the *County and City Data Book, 1972,* produced from Census Bureau tapes—could be cited under any of these, with the consequence that a thorough searcher must check at least eight different positions in SSCI's alphabet of corporate authors for possible entries.

Inconsistent renderings of titles are no less a problem. There are two major reasons why title citations to the same work may turn up in widely different positions in the anonymous section of SSCI. One is that many data files are actually known by several titles—a source of confusion documented by Dodd[5]—and researchers reflect this diversity when they cite. The other is that even researchers who use the same title do not always record it in the same way. For example, one person may write "General Social Survey 1972" and another, "1972 General Social Survey." The latter style, with year first, will almost surely cause citations to be lost to some users of SSCI or Social Scisearch: the computer puts entries starting with digits wholly outside the alphabetical sequence, and the person searching alphabetically for a title may never think to look in the numeric positions following "Z." *Many* MRDF can be cited with either a word or a string of digits (such as a year) coming first and determining where the entry will be computer-filed. So, again, the potential for scattering is great. It should also be noted that these problems with titles persist when the title is the *second* element in citing, after personal or corporate author, since both first and second elements are used in computer sorts.

It should now be clear that anyone who wants to examine the use-history of a data file in SSCI has a time-consuming task ahead. One needs to look in many places to achieve both positive success, which is finding entries, and negative success, which is ascertaining that there is nothing in a particular place to be found. The essential problem with SSCI, whether we want use-histories of data files or anything else, is insufficient vocabulary control. This is a classic problem in creating large and growing bibliographic files: its ramifications were recognized long ago by library catalogers, whose response was to create: (1) authority lists that standardized personal and corporate author names; and (2) uniform titles that conveyed the fundamental identity of works, despite the multiplicity of editions, versions, translations, etc., of these works appearing under diverse names. (Thus, the uniform title *Arabian Nights* in card catalogs unites all editions of this work under a single heading, whether they are titled *Arabian Nights, The Thousand and One Nights,* or *Tales of Scheherazade.*) Lubetzky and Hayes in 1969 directed attention to the fact that *Science Citation Index* was failing to unify references to specific intellectual works because it merely transcribed citers' references to editions of these works.[6] (They use a paper by the American physicist F. Willard Gibbs as an example.) The problem remains, in both SCI and its newer companion SSCI. Griffith recently noted that a computer search of SSCI tapes failed to show *Das Kapital* as heavily cited.[7] The reason is not that this most influential of writings is not heavily cited, but that the citations it receives are scattered among many different editions and many different citing styles. To a computer, it seems that many different works are being cited only a few times each: one by Marx, another by Marx K; one named *Das Kapital,* a second named *Kapital,* a third named *Capital,* a fourth a volume in Marx's *Collected Works,* and so on. Exactly the same thing has happened with MRDF.

The Institute for Scientific Information is aware of the varieties of fragmentation recorded here. ISI's problem is economic: it is prohibitively expensive to make the copy of thousands of citers conform to authority lists of authors' names and uniform titles. If this work is to be done, it will very likely have to be done by outsiders—a point to which I shall return in closing.

An Experiment in Finding Citations to Data Files

Three major data files—or rather sets of data files—were chosen for an experiment in citation retrieval in SSCI: the General Social Survey, conducted annually since 1972 (except 1979) by the National Opinion Research Center (NORC) in Chicago; the Panel Study of Income Dynamics, conducted annually 1968-78 by the University of Michigan's Institute for Social Research; and the same institute's American National Election Studies, conducted biennially in election years since 1948. Copies of these files are held by hundreds of colleges and universities, and are known to be used by social scientists and their students. One would expect at least some of this use to be registered in references in published papers; and in fact some is. A trial manual search of SSCI, 1973-79, produced 110 citations to the General Social Survey, 47 to the Panel Study of Income Dynamics, and 23 to the Election Studies, as of August 1980. The search was "uncritical" in that all of these files are actually multiyear serials with heterogeneous content, and a citation to any part in any year was counted as a hit. On the other hand, many of the questions in these surveys are repeated over the years, and persons citing files issued in different years may be using the same questions. In any case, the point of the trial search was to find as many citations to the three files as possible, without worrying about refinement by year or by subject.

Table 1 sets forth the various author and title headings under which citations to the three files were found. To keep the table from being unmanageably complex, not all variants in entries have been listed. Even so, the dominant impression from the table is one of complexity and high fragmentation, in sharp contrast to the earlier, idealized example in which a single entry named the pornography survey. Anyone compiling a use-history of the data files in table 1 (or of any others) must in fact search along lines suggested there, and earlier in this paper, if near-completeness is to be attained. Note not only the divergent forms of the same heading, but also the several wholly different headings under which one finds entries in the sections of SSCI.

The search for citations to the General Social Survey (GSS) yielded use data that could be compared to those in NORC's 254-item *Annotated Bibliography of Papers Using the General Social Surveys* of April 1979. The NORC compilers, while acknowledging that their list is far from complete, state that they included "a computer-assisted check of *Sociological Abstracts, Dissertation Abstracts,* and the *Social Science Index*..." in doing their search.[8] It is not clear whether the latter is H.W. Wilson's *Social Sciences Index,* which cannot be searched by computer, or the *Social Sciences Citation Index,* which can. Interestingly, however, the manual search of SSCI for the present article turned up fully sixty papers citing one or more annual issues of the GSS that are not

TABLE 1
CITATIONS TO THREE DATA FILES IN SSCI'S CITATION INDEX

Data File	*Section and Headings**	*Hits*
General Social Survey (various years)	Personal Author —Davis J —Davis JA	42
	Corporate Author —Nat Op Res Ctr —Nat Opin Res Ctr —NORC —Rop Publ Op Res C	46
	Anonymous —Codebook Spring 197- —Codebook 197- General —General Social Surve 197- —National Data Progra 197-	22
Panel Study of Income Dynamics, 1968-1978	Personal Author —Morgan JN	11
	Corporate Author —I Soc Res —I Soc Res Surv Re —Mi I Soc Res —Mi U Surv Res Ctr —Surv Res Ctr —U Mi Surv Res Ctr —U Mich Surv Res Ctr —U Mich I Soc Res —U Min Surv Res Ctr [*sic*]	28
	Anonymous —Panel Study Income D —Pan Stud Inc Dyn —Pan Study Inc —Pan Surv Inc Dyn [*sic*]	8
American National Election Study (various years)	Personal Author†	0
	Corporate Author —Ctr Pol Stud —U Mich —U Mich Ctr Pol St —U Mich Interu Con —U Mich Pol Beh —U Mich Surv Res C	19
	Anonymous —CPR 197- Am National [*sic*] —CPS 197- Am National —CPS Am National Elec 197- —SRC 197- Am National	4

*Author and title headings for this search were derived from title pages of codebooks and Inter-University Consortium for Political and Social Research. *Guide to Resources and Services* 1979-1980. Ann Arbor: University of Michigan, 1981. Except where noted, hits were found under all headings listed.

†No hits were found under Angus Campbell or Philip E. Converse, both of whom have been principal investigators for this survey.

recorded in the NORC bibliography. A few of these are discussions of the GSS as a resource in the librarian's or data archivist's sense. But the great majority are substantive research papers from the same period as those contained in the bibliography. Some are from "major social science journals" of the sort the compilers say they searched (e.g., *Social Problems, Human Relations, Social Forces, Sociological Quarterly, Political Science Quarterly*); others are from "unpredictable" specialty journals (e.g., *Law and Contemporary Problems, Archives of Sexual Behavior, Journal of Communication, Review of Religious Research, Personnel Psychology, Curriculum Inquiry, Journal of Homosexuality*), indicating quite vividly the cross-specialty diffusion of GSS data.

These remarks are not intended to derogate the NORC bibliography, which is a valuable work, especially its notes on specific GSS variables employed by researchers. Rather, they are intended to show that SSCI can reveal use of data files even to persons, like those at NORC, who are well placed to know about such uses. However, SSCI's full potential can only be brought out by searchers who know its peculiarities and are willing to look in many different places for citations.

Prospects for Use-Histories

Over the next decade, it may be that both researchers and editors of journals in which they publish will settle on a few more or less standard ways of citing MRDF. The goal is not just proper credit for a file's originator, producer and distributor, but its retrievability as an intellectual work. Citers need to learn to see citation as a contribution to document retrieval—no less so when the "document" is an MRDF than when it is another author's monograph or paper. Toward this end, it would greatly help if journal editors published model citations to data files in their instructions to contributors, just as they now do for works of other kinds. Such model citations should also be incorporated as soon as possible in widely used style guides, such as the Modern Language Association's and Turabian's.

There are examples on which to draw. For the GSS and certain other files, a standard citation now appears in the front matter of the codebook. These are influenced by the style developed for cataloging data files in the *Anglo-American Cataloging Rules*, 2d ed. (AACR2). Under AACR2 (which is superior as a guide to the ANSI standard for bibliographic references to data files), the choice of initial element in a citation may come down to two: principal investigator's name, like that of any other author; or title when the PI is not known. Corporate author entries seem to be falling from favor. With some such standardization in the author and title fields, use-histories of data files would be easier to compile.

However, skeptical wisdom suggests that citers and editors will continue to go their idiosyncratic ways, which jointly yield the fragmented entries now in SSCI. Skeptical wisdom also suggests that the fragmentation will not be corrected by human intervention at ISI (although computer algorithms for resolving some of the differences in table 1 seem possible). Those wanting use-histories of data files can only expect to earn them, in the foreseeable future, by hard digging—i.e., by multiple look-ups in a manual search of SSCI and by much consultation of the Cited Reference index in Social Scisearch online. The present article shows that hard searching, using all the entry points implied by table 1 plus any others that appear plausible, does produce hits, and that is perhaps its most encouraging finding.

Researchers should benefit from this knowledge, since they gain from being able to extract use-histories of files of interest. But one hopes that not all use-histories will remain private documents in the hands of researchers. There is a need for published bibliographies of works based on data files, like that of NORC for GSS. The compilers presumably would treat items they found in SSCI, however diverse, as "raw copy" to be bibliographically standardized (in forms of titles, authors' names, etc.) and newly arranged in some appropriate order. It would seem natural for such bibliographic projects to be sponsored by major data suppliers (e.g., the Inter-University Consortium, the Roper Center), at least for data files they know are widely held and used. It would also seem natural for the federal government, which issues so many files as government documents, and which now actively promotes secondary analysis, to take an interest in use-histories as a newly deserving form of bibliography, and to support compilation of them adequately with funds.

ACKNOWLEDGMENT

The author wishes to thank Carolyn Latta, his graduate assistant in the School of Library and Information Science, Drexel University, for the arduous search of SSCI reported above.

References

1. Rowe, Judith. "Facilitating Information Access: Interaction Between System Components (the Data Library and the Traditional Library)." *SIGSOC Bulletin* 6(Fall/Winter 1974-75):34.

2. White, Howard D. "Libraries and Access to Social Science Data." In ________. *Reader in Machine-Readable Social Data,* p. 182. Englewood, Colo.: Information Handling Services, 1977.

3. Dodd, Sue A. "Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines." *Journal of the ASIS* 30(March 1979):77-82.

4. Garfield, Eugene. "Anonymity in Refereeing? Maybe—But Anonymity in Authorship? No!" In ________. *Essays of an Information Scientist,* vol. 2, pp. 438-40. Philadelphia: ISI Press, 1977.

5. Dodd, Sue A. "Titles: The Emerging Priority in Bringing Bibliographic Control to Social Science Machine-Readable Data Files [MRDF]." *IASSIST Newsletter* 1(Fall 1977):11-18.

6. Lubetzky, Seymour, and Hayes, R.M. "Bibliographic Dimensions in Information Control." *American Documentation* 20(July 1969):247-52.

7. Griffith, Belver C., and Small, Henry G. "A Philadelphia Study of the Structure of Science: The Structure of the Social and Behavioral Sciences Literature." In *Proceedings, First International Conference on Social Studies of Science,* Cornell University, 1976, p. 23; and Griffith to White.

8. Smith, Tom W., et al. *Annotated Bibliography of Papers Using the General Social Surveys.* Chicago: National Opinion Research Center, 1979, p. 1.

This Page Intentionally Left Blank