Citing data sources in the social sciences: do authors do it?

Hailey Mooney
Data Services and Reference Librarian
Michigan State University Libraries

100 Library
East Lansing, MI 48824
517-884-0857
mooneyh@msu.edu

Abstract

It is expected that authors will provide citations for all papers referenced in their writings. The necessity of providing citations for data is not so widely recognized. Proponents of the data sharing movement have advocated for the citation of datasets in order to recognize contributions and enhance access. This study examines a sample of papers from the Inter-University Consortium for Political and Social Research (ICPSR) Bibliography of Data-Related Literature that are based on secondary analysis of datasets available in the ICPSR data archive to determine the data citation practices of authors. The results indicate that many authors fail to cite the data used in secondary analysis studies. Possible reasons for the dismal state of data citation practices are considered including the recent introduction of data into the scholarly record and its marginalization as an information format. Updating citation practices to include datasets will support data sharing and foster responsible scholarship.

Introduction

Like any system that works well most of the time, we take the availability of bibliographic citations for granted until we encounter a glitch. Complaints of inadequate and incomplete citations have been aired by Henige¹ and echoed by Donovan.² Others have bemoaned the incidence of outright failure to cite sources at all.³,⁴ Most reference librarians can...
recall a time when a patron came to them with an incorrectly transcribed citation that took some creative strategies to track down. In fact, a whole subset of the information science literature investigates the prevalence of inaccurate references. Academic writing would be incomplete without citations, yet despite these occasional calls to attention the citation often falls into the background: a tedious and tiresome technical requirement.

Although technical requirements may seem tedious at times, these considerations are what enable a coherent system of scholarly communication. The citation is a manifestation of an information standard: standards that are adopted by publishers and libraries to provide us with a system of information creation, dissemination, organization, and retrieval. These standards are passed on to authors through style manuals and other documents that communicate the proper format for the creation of references. Furthermore, the practice of citing is essential to the academic and research communities which acknowledge and build upon the intellectual achievements of past and present colleagues.

Traditionally, the medium for contributing to the advancement of knowledge has been the written word. Scholars read, write, and cite books and journal articles. In the early digital age the written word still reigns, but the increased prevalence and importance of digital research datasets (computer files made up of rows and columns of numbers and accompanied by a codebook, or metadata that explains the significance of the numbers) is changing the picture. In particular, computers have fueled the expansion of quantitative analysis through the availability of statistical software programs that read numeric datasets. Combine the highly-regarded tradition of citation to acknowledge intellectual debts with the modern use and re-use of machine-readable data to create new knowledge and presto!—the data citation is born. Unfortunately, it’s not quite that simple. Institutions are slow to change, and the scholarly
publishing establishment is no different. The data citation is alive, but it has yet to reach maturity through consistent use. Acknowledgement of intellectual debts should be repaid by citation regardless of whether those debts originate from the reading of an article or the analysis of a dataset.

**Data Sharing and Citations**

As the importance of digital data to research has increased, a data sharing movement has developed. This is similar to the open access movement, which shares the belief that information should be freely available in order to further the pursuit of knowledge in our society. Using publicly available data and sharing data among researchers is not a new phenomenon (Emile Durkheim used national statistics in his 1897 study of suicide), but the widespread recognition of research data as a valuable commodity to be archived, disseminated, and accessed is more recent. Along with this recognition has come concern about the lack of standards for citation of numeric datasets.

Early publications from the data sharing movement recognize citation as an issue. For example, Fienberg, Martin, & Straf recommend that “journals should require full credit and appropriate citations to original data collections in reports based on secondary analyses” in order to promote the sharing of research data. However, current discourse asserts as common knowledge that there is no consistent standard for the citation of numeric datasets. Take for example, the observations that “Many researchers…are not aware that published data deserves citation just like published articles, perhaps in part because so many articles presently use data without citation”, that “Data are sometimes listed in the references, sometimes in the text, and only occasionally with enough information to guarantee future access to the identical data set”,

---

*Page 3  
Citations for data*
and finally that “Researchers will cite well known datasets within their subject area (though there is not always an accepted format for doing so), but for less recognised datasets citing articles based on them is more typical”. The gap in citation standards and practices for datasets is acknowledged as a major detriment to further growth towards widespread data sharing. A recent article from Nature on the slow pace of movement towards open data and use of institutional repositories notes that a contributing factor towards scientist’s unwillingness to share data is the need for journals and data banks to “ensure proper citation for data sets”, without which there is no way to guarantee attribution. Lack of credit for data sharing is a clear deterrent for researchers.

**Data Citation Standards**

The response from the scholarly community to the citation worries emanating from the data sharing movement is largely concerned with providing guidelines for accurate and robust dataset citations. The need for data citation standards has been expressed in the literature for over 30 years. In 1979, Dodd published a prescriptive article suggesting guidelines for citation of numeric data files. These guidelines resulted from the work of the Classification Action Group of the International Association of Social Science Information Services and Technology (IASSIST). IASSIST, a group concerned with social science data services, has continued to discuss issues surrounding data citation over the years. The topic arose again in a discussion paper on bibliographic references for computer files based on discussions among representatives of the Inter-University Consortium for Political and Social Research (ICPSR) data archive. Recent years have seen an increase of attention given to the matter, with discussions arising again at ICPSR and IASSIST. These discussions eventually gave way to a new proposed standard that has been adopted by the Dataverse Network Project, a virtual...
archive application used by institutions such as the Institution for Quantitative Social Science at Harvard University. The National Information Standards Organization (NISO) has also taken up consideration of the matter, sponsoring a NISO Thought Leader Meeting on Research Data\textsuperscript{20} which recommended the development of guidelines for data citation. The Organization for Economic Cooperation and Development (OECD) recently weighed in on the issue by providing an anecdotal overview of the current state of citation practices and proposing their own standard for the citation of datasets and statistical tables.\textsuperscript{21} Most recently, in late 2009, the DataCite\textsuperscript{22} initiative was founded to facilitate the citation of data through the registering of Digital Object Identifiers for datasets.\textsuperscript{23}

Undoubtedly as a result of these efforts, researchers seeking to cite data can now find examples from a variety of sources. Two of the major social science data archives, ICPSR and Roper Center for Public Opinion Research, both provide general directions on how to cite datasets from their archives\textsuperscript{24,25} and include a recommended citation along with the documentation for individual studies. For example, here is a citation provided by ICPSR:


ICPSR has been providing standardized citations for datasets in its archive since 1989.\textsuperscript{26} Older documentation recommended that authors insert a note giving credit to the creator and distributor.
Some domain-specific style guides provide clear examples for dataset citations, such as the American Sociological Association Style Guide\textsuperscript{27} and the APSA Style Manual for Political Science.\textsuperscript{28} One of the most popular style manuals for the social sciences, the *Publication Manual of the American Psychological Association*, includes a recommended format for citing both published and unpublished datasets.\textsuperscript{29} The APA manual has provided examples for the citation of datasets since the third edition was published in 1983.\textsuperscript{30}

**Data Citation Practices**

So although best practices for dataset citations are still evolving, there have been readily available citation examples available to data users since 1989 in ICPSR and since 1983 in the APA style manual. Why then are commentators on the data sharing movement still bemoaning the lack of citations? Hasn’t the availability of citation examples informed authors’ writing practices?

Despite all of the anecdotal confirmations, there is a lack of empirical evidence in the current literature affirming that authors often fail to cite datasets. Only one such study surfaced: an examination by Sieber and Trumbo\textsuperscript{31} of publications based on the General Social Survey and published from 1976 to 1988 paired with a survey of author attitudes towards dataset citations. The findings showed that only a small percentage (19\%) of authors included the survey name in the reference section of their papers and 9\% of authors omitted any mention the name of the survey that their publication was based on. The survey of author attitudes found that 62\% of respondents were unaware of any existing standards for the citation of data. Overall, the study concluded that “researchers’ behavior, attitudes, and knowledge concerning the citation of data
sets fall short of the ideal that would foster openness, fairness and economy in the pursuit of scientific knowledge.” This certainly confirms the assumption that data is not consistently cited.

**An Exploratory Investigation**

Unable to identify any more recent studies similar to Sieber and Trumbo’s analysis of publications using the General Social Survey dataset, I initiated my own small-scale empirical investigation into data citation practices. My exploratory study looks at the data citation practices of social science faculty at a large research university. Although this is a view into a single community of researchers, it functions as a microcosm of the broader scholarly landscape, and can be seen as an indicator of the wider academic community.

**Methods**

In order to determine the extent to which authors include dataset citations in their published research articles, a content analysis of journal articles was performed. The sample of articles was limited to those published by social sciences faculty at a large research university, with which the author of this study is familiar, from the departments of criminal justice, political science, psychology, and sociology. Specifically, only empirical research articles published in the scholarly journal literature based on secondary analysis of datasets deposited in the ICPSR data archive were included in the sample.

The ICPSR Bibliography of Data-Related Literature, which indexes citations of works based on analysis of datasets that are freely available in the ICPSR archive to member institutions, was used to identify the articles in the sample. Articles included in the Bibliography have either been self-submitted by authors or identified by ICPSR staff as using or describing datasets available through ICPSR, although some datasets may also be available from other
sources. This indicates that articles in the Bibliography are likely to acknowledge data sources at least enough for identification and inclusion. Additionally, as users of ICPSR data are provided with a suggested citation, a sample drawn from the Bibliography will represent a best case scenario for data citation practices.

The Bibliography’s author index was searched for names of current departmental faculty to develop an initial list of articles in the sample. These could be works that were either authored or co-authored by a faculty member and must have been published from 2000-2009. Non-journal article works (such as agency reports, working papers, book chapters, etc.) and review articles were removed from the sample. Articles based on the original data collection activity of the author were also removed to keep the sample limited to secondary analysis research only. Given these criteria, a total sample of 49 journal articles was included in the study.

Each article was analyzed to determine whether or not a reference to the dataset was given anywhere outside the main body of the text that supplied key elements of citation information such as title, producer, and distributor. The location of the reference was then noted to allow for a distinction to be made between a formal standardized citation appearing in the reference list and an informal citation made in a note (such as an author’s note, footnote, or endnote). Since a common practice is to cite a publication related to the data, rather than the data itself, any related publications that were cited when explaining the origin of the data were noted. Finally, the website of the journal in which each article appeared was visited in order to ascertain the citation style format specified in the author guidelines.

What counts as a citation?
Note that there are two measures of citation in this study. Moving beyond the mere mention of the dataset name used in a secondary analysis article, a citation must contain the information necessary to easily find and retrieve the dataset and appear as supplemental documentation outside of the main body of the text. Therefore, notes that acknowledged the original creator of the data and directed the reader to an access point (i.e., the distributor—such as the ICPSR data archive or a website maintained by the survey organization), were counted as a citation. The additional measure of location differentiates between these reference notes and citations appearing in the reference list that adhere to bibliographic standards enforced by style manuals.

Results

Sample Characteristics

The sample of 49 journal articles represented 25 individual authors publishing in 39 different journals and using 33 unique dataset series. Table 1 shows the distribution of the sample across the social science departments. Both the criminal justice and political science departments had 28% of authors each, with the largest number of articles from the criminal justice department at 39%. The smallest group was the psychology department with 20% of the authors and 10% of the articles. The sample skews slightly towards the first half of the decade, with about 60% of articles published from 2000-2004 and 40% from 2005-2009.

Frequency of Data Citations

Overall, 30 of the articles (61%) fail to provide any type of citation to the dataset. The remaining 19 articles (39%) do provide a citation under the broadest definition of the variable, meaning that an access point is made available outside the main body of the text. Out of these 19
articles, 14 of them (74%) provide a formal citation of the dataset in the reference list; this is 29% of the entire sample. There is almost a 50/50 split of articles that cite a related publication, with 47% not citing and 53% citing a related publication. Figure 1 illustrates these percentages. A cross-tabulation of dataset citations by related publication citations reveals that 10 articles (20%) have no citation whatsoever for their data (neither data nor a related publication), relying solely on in-text explanations, and that 6 of the articles (12%) cite both the data and a related publication.

**Journal Styles and Data Citation Practices**

An examination of style manuals adopted by the journals in the sample (as shown in Figure 2) reveals that the majority of journals instruct authors to use APA style, followed by generic styles specific to the journal and Chicago style with 14% each. Represented in small numbers was Harvard, ASA, APSA, AMA, and 4% of journals did not specify any particular style in their author instructions. Figure 3 shows the number and percentage of articles within each citation style that include dataset citations.

The majority of journals instruct authors to use the APA style manual, which does provide cursory instructions on the citation of datasets. Nonetheless, articles from journals that have adopted APA style set the overall trend of the approximately 60%-40% split of not citing and citing data. Based on the inclusion of examples for formatting dataset citations in the APSA and ASA style manuals, the expectation is that the articles from journals using these styles would have dataset citations, yet the APSA article in the sample fails on this account and only 3 out of 4 ASA articles provide the citation. This suggests that editors are not consistently enforcing the guidelines used by their journals, or may be unaware of the full extent of the style manual’s
recommendations. None of the journals with generic individual styles published articles with
dataset citations in the reference list, presumably due to a lack of recognition for the importance
of citing data.

Discussion

This content analysis of journal articles using secondary analysis of ICPSR datasets
confirms the assumptions voiced by proponents of the data sharing movement that data are not
consistently cited. Only 29% of the sample articles cite in the reference list the dataset analyzed
in the article. This is seriously lacking considering that citing anything less than 100% of the
articles and books referenced in a paper would be considered a serious omission. Furthermore,
this shows only a minor increase from Sieber and Trumbo’s analysis of papers published in the
1970s and 80s which found 19% of dataset titles cited in the list of references.

Admittedly, this study analyzed a small convenience sample of publications authored by
faculty at just one institution over a period of 10 years. While it may not be statistically valid to
generalize this sample to the entire universe of scholarly writers and researchers engaging in
secondary analysis, it does provide a much needed empirical exploration of actual data citation
practices over the previous decade. There is certainly room for future studies to augment these
findings based on a larger sample of articles published by researchers across multiple institutions
and looking at a single year of the most recent crop of articles. The extent to which disciplinary
style manuals and journal style sheets address citation of data would also benefit from further
examination.

The importance of this investigation lies in the questions it raises rather than the
questions it answers. We can now state with a greater degree of certainty that authors are not
citing datasets and that editors are not enforcing the inclusion of citations for datasets. Why is this? Journals and disciplines vary in the formats that they adopt, but you would be hard-pressed to find an academic publication that doesn’t require citations to all of the resources referenced in a paper. Isn’t it necessary to explain the source of data used in secondary analysis? Shouldn’t this explanation include the information necessary to locate the dataset? Don’t we already have a system of citation that provides readers with the necessary information to locate sources referred to in a paper? Isn’t a dataset (especially those available in archives) just another format of information that we can cite? Unearthing attitudes towards citing data and current knowledge of citation standards is a ripe area for future research.

**Looking for answers: Why don’t authors cite data?**

Let’s start with the bigger picture first, why do authors cite at all? Numerous theories (many of which are summarized by Nicolaisen\textsuperscript{33} and Camacho-Miñano & Nuñez-Nickel\textsuperscript{34}) have been proposed to answer this question. The normative theory of citing, based on Robert Merton’s conception of science as a social system with a set of shared values or norms, is among the most compelling. Kaplan\textsuperscript{35} applied Merton’s norms to citation behavior, observing that citations allow for the sharing of ideas while maintaining “individual property rights.” Interestingly, the very complaint that researchers have against sharing data is that it is their “private property”.\textsuperscript{36} Perhaps criticism of the normative theory of citation as overly idealistic\textsuperscript{37} is correct, as personal motivations and rewards have a strong impact on the desire to even provide a resource for other scholars to cite. Yet, the focus of my exploratory investigation was on datasets that already have been made widely available for others to use. The general reluctance towards sharing data may be having an effect on the overall treatment of data sources in research findings.
Regardless of motivations for citing, there is a long and venerable tradition of citation within academia. Uniform citation formats are a more recent development of the twentieth century (for example, Chicago Manual of Style was first published in 1906, the National Information Standards Organization was founded in 1939, and APA style was first published in 1952). Recall from earlier discussion that Sue Dodd’s work in 1979 to bring bibliographic control to datasets forms the basis of what are really very recent and evolving standards for the citation of data. Datasets themselves are also quite young as an information format. The development of machine readable data files dates back to the 1890 Census and the use Herman Hollerith’s punch card machine, but the founding of ICPSR in 1962 is a marker for the beginnings of the widespread adoption of social science data files in academic research. Time is the explanation here. It simply takes time for normative behaviors to develop and for systems to integrate new developments.

Data has historically been a marginalized resource due to its inability to fit into the print-based paradigm of scholarly publication and dissemination systems. Although data may be ubiquitous in the world of social research, as an information format handled by publishers, editors, and librarians it has largely fallen by the wayside. ICPSR was founded by political scientists responding to a need for a data repository that established information institutions were not meeting. As a format, data was originally difficult to handle (think punch cards and magnetic tapes) and required expert knowledge to manage and use. Although technological advances have simplified access and use, the specialty nature of the data format has continued reliance on a network of dedicated archives to provide for the needs of collection, access, and storage. However, unique formats do not negate the validity of datasets as primary “unit of communication” in the scholarly environment. Data and other digital formats are creating the
need for a more flexible scholarly communication system that can incorporate non-traditional
materials. Recognizing data as an essential component of shared research resources along with
journal articles and books, rather than as private property to be exploited for personal gain, will
move it from the margins to the mainstream. If researchers view data as a medium for the
transmission of information just like any other resource, then it will be imminently more citable.

Conclusion

The current state of practice of the citation of datasets is seriously lacking. Acknowledgement of intellectual debts should not be limited to only certain formats of information. Editors and publishers also have a role in updating and upkeeping the scholarly communication infrastructure by embracing full citation of digital formats. Publishers hoping to capitalize on advances in infrastructure that create easily navigable links using DOIs between scholarly content will need datasets to appear in the reference list to facilitate this capability. Scholars analyzing datasets disseminated by ICPSR are provided with citations that can easily be put to use. It is the responsibility of authors as members of a community of scholars to adhere to the social contract of citation. We can foster the data sharing movement by developing a culture of data citation that gives researchers a tool to ensure the recognition of their contributions and provide access for future scholars. And of course, citing is just good scholarship.

Illustrations

Table 1. Distribution of sample across social science departments
<table>
<thead>
<tr>
<th>Department</th>
<th>Authors</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>Criminal Justice</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Political Science</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Psychology</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Sociology</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 1. Percentage of articles with data and related publication citations

Figure 2. Percentage of articles by style manual
Figure 3. Type of data citation by style manual
References


31. Sieber, J.E. and Trumbo, B.E. 1995. (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics* 1(1): 11-20. [http://dx.doi.org/10.1007/BF02628694](http://dx.doi.org/10.1007/BF02628694)


