Data sharing: Academic libraries and the scholarly enterprise

Abstract

Technological advances have raised expectations for data sharing, and financial exigencies have brought the issue into sharper focus, especially as grant-funding institutions are beginning to require shared access to research results and the data that support them. These data are increasingly linked to publications and related resources, therefore making sharing inexorably linked to scholarship itself. In this study, we offer a descriptive analysis of the state of data sharing in sociology as compared with practices among a representative sample of other academic disciplines. We also consider the implications for the research community of trends in data sharing and data access. Finally, we address the unique role of academic librarians as partners in support of disciplinary researchers, teachers, students and other data users.

Introduction

"Data are outputs of research, inputs to scholarly publications, and inputs to subsequent research and learning. Thus they are the foundation of scholarship," frankly asserts Christine Borgman, Presidential Chair & Professor of Information Studies at the University of California at Los Angeles. [1] These data are increasingly linked to publications and related resources, therefore making sharing inexorably linked to scholarship itself. Attendant to the growth of data intensive scholarship is a growing chorus of voices supporting the nascent Open Data movement. The Open Data movement shares some similar features with the more established and familiar Open Access movement—namely to remove the barriers that limit access to scholarly materials—yet suffers from its newness and lack of definitional rigor.[2] The concept of open data has recently gained prominence, as the United Kingdom's data.gov (http://data.gov.uk/) and the

United States' open governments (http://www.whitehouse.gov/open) initiatives have expressed aims to promote open access to public-sector government data and slowly turn them into linked open data where appropriate. The "data" in these phrases often refer to a diverse set of non-textual materials, primarily statistical or numeric datasets, as well as other non-text data such as genomes, audio files, chemical compounds and biological component materials; yet we include textual materials in our analysis.

While the theoretical wrangling continues around open access movements, it is important to consider this discussion in view of the thorough and ongoing debate around the practical issues regarding the use, reuse and general availability of research datasets in a variety of academic disciplines. The literature is replete with strong support for open data policies,[3] yet equally abounding with the concerns and complications caused by sharing of scholarly data.[4] Perhaps a recent Nature editorial captures it best:

> Does anyone want your data? That's hard to predict, but the easier it becomes to request data and to receive credit for sharing it, the more likely people are to ask. After all, no one ever knocked on your door asking to buy those figurines collecting dust in your cabinet before you listed them on eBay. Your data, too, may simply be awaiting an effective matchmaker.[5]

Given the trajectory of funders requiring data management plans[6] and explicit statements in support of data sharing,[7] it is not surprising that in his 2007 article, Northwestern University sociologist Jeremy Freese makes the case that sociology ought to lead the way in promoting "openness and accountability of professional practice" by providing access to research data at time of publication.[8] He states the premise very clearly:

> Increased replication standards would be beneficial for the credibility of sociological research not only by increasing confidence that work can be verified, but also by making published work more available to elaboration and extension by others and by affording the best opportunity for exemplary work to contribute to teaching other members of the profession.[9]

But where Freese optimistically perceives a mandate for data access, Andrew Abbott offers a litany of worries around “ownership, mechanics, and security.”[10] He relates that it takes weeks to clean a dataset and that the distributed nature of data archives places large hurdles in front of potential (re)users. He also points out the enormous additional burden that the process of data verification would place on editors and reviewers. Yet it is primarily the security and/or confidentiality of the dataset that undergirds his concern.

Prompted by the explicit concerns expressed by Freese and contradicted by Abbott, this study offers a descriptive analysis of the state of data sharing in sociology as compared with practices among a representative sample of other academic disciplines. We use this backdrop to consider the implications for the wider research community of trends in data sharing and data access. Finally, we address the unique role of academic librarians as partners in support of disciplinary researchers, teachers, students and other data users.

Framework and Methodology

Pao-Long Chang and Pao-Nuan Hsieh assert that the doctoral dissertation is the major distinguishing feature of education, and one that, more importantly, "reflects the capabilities of the candidate [and] identif[ies] the student as a future contributor to the field, since the research problem initiated in a dissertation usually constitutes the foundation of research projects subsequently conducted by the student/author after his graduation."[11] Therefore, it may be reasonable to assume that dissertation authors, as emerging scholars, are exemplars in creating rigorous studies that represent the latest trends, theories, and best practices in research and scholarship in their discipline.

We sought to analyze doctoral dissertations from a representative cross-section of academic disciplines. Working with the typology presented by Tony Becher, refined by Jenny Fry, and tested by Samuelle Carlson and Ben Anderson,[12] we chose to examine dissertations drawn from biology ("hard pure"), sociology ("soft pure"), mechanical engineering ("hard applied"), and education ("soft applied"). We were provided complimentary access to the ProQuest Dissertation and Theses (PQDT) full-text database; and in March 2008 we extracted a large sample of citations from this database.

Once the results were returned we used Microsoft Excel's random number generator to assign numbers to each dissertation; after a basic sort of each subject list (from lowest to highest, according to the assigned random number) we could draw a near-random sample from this set of results. We then for convenience limited our sample to the first twelve dissertations from each of the four sorted results sets, for a total sample N=48.

Each of the 48 dissertations has been content-analyzed and coded for ten descriptive variables. Being mindful of the pitfalls of transforming the raw data of communications into a standardized form, we utilized a mixture of manifest and latent coding, which is described by Earl Babbie in his classic social research text as coding for the visible, surface content (manifest) versus garnering the underlying, or latent meaning.[13] Following a set of inter-rater reliability tests we analyzed each dissertation for its use/creation of primary and secondary data, and for the apparent availability (or non-availability) of such data for reuse and further research.

We defined "data" broadly to include not only numeric datasets, but also resources such as "images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations," as did the National Science Board (NSB) in their informative report entitled Long-Lived Digital Collections.[14] Specifically, again drawing on the work of the NSB to assist us in grounding our analysis, we sought to distinguish the data by their origins (observational, computational and/or experimental) and their functional category (research data, resource data or reference data).[15] We also paid close attention to mentions of data collection and use within each dissertation's table of contents and abstract, acknowledging the assertion by Guy Adams and Jay White that dissertation authors "typically want to represent their research as accurately as possible in an abstract."[16]

Analysis and Discussion

In our coding and subsequent analysis of these dissertations, we sought to address a series of questions concerning both a cross-disciplinary comparison of the creation and use (and availability for reuse) of research data, as well as an examination of the efficacy for librarians and researchers of having a subscription-based commercial provider as the primary means of access to dissertations.

Presence of data revealed in titles, abstracts and tables of contents

Researchers (and the librarians who help them) may only consult first-level descriptive information (title, abstract) from an indexing tool such as PQDT when searching for relevant dissertations to support their research. We therefore sought to determine the extent to which an examination of these first-level descriptors would reveal the presence of data. We then sought to compare these results to those found from examining a more detailed level of dissertation information—the tables of contents.

Although we did not expect to find "data" specifically mentioned within dissertation titles, this was nonetheless our first point of inquiry. As predicted, we quickly ascertained that none of the dissertations in our sample included the word "data" in the title. With our next point of inquiry, we sought to determine whether each dissertation's abstract would reveal the presence of data within the dissertation. As the author-supplied dissertation abstract is often the source of the abstract included in commercial indexes such as PQDT, the information supplied in this abstract is critical for the researcher who may be using PQDT as a means of discovering relevant new research in her field of inquiry. Based on our examination of abstracts, we found that only 31% of the abstracts in our sample stated that data was included in the research; however, if we included abstracts that also seemed to suggest (but not explicitly state) that data may be present, the total increased to 54% (N=36). Broken down by discipline (see Table 1), Education had the highest percentage of dissertations whose abstracts explicitly cited the presence of data (50%), while Sociology was in close second place with 42% in this category. [Insert Table 1 approximately here.]

An examination of the tables of contents of our sample set of dissertations reveals that the proportion of our sample that contains data is, at 63% (N=30), twice as large as the proportion revealed by an examination of dissertation abstracts (N=15). Only 31% of our sample included tables of contents that offered no explicit mention of data. Considered another way, we found that among all of the dissertations with *abstracts* that did not mention the presence of data (44% of our sample, N=21), a subsequent examination of their *tables of contents* revealed that 38% of these (N=8) actually do offer data-supported research. The disciplines that best represented the presence of data in tables of contents were Education and Sociology. As indicated in Table 2, 100% and 75%, respectively, of the tables of contents from dissertations in these disciplines clearly indicated the presence of data. [Insert Table 2 approximately here.]

These results serve to refute, in at least one important part, the claim by Adams and White that abstracts are an accurate reflection of an author's work. The implication of this finding is serious: while abstracts are generally included in dissertation-finding tools (i.e. online indexes to dissertations, such as PQDT), tables of contents are not included. Therefore, the typical researcher's use of only an online index to identify dissertations that offer data-supported research—and that may include access to datasets—would fail to serve up a significant proportion of relevant dissertations. These could only be discovered by a researcher who is willing to click through to the full text of the dissertation in order to view its table of contents. Worse, if the researcher were using an online index that does not include full texts of dissertations, then he would have to complete the laborious process of using interlibrary loan to obtain the full text of a dissertation in order to examine the table of contents. Even under both of these scenarios, the table of contents may not explicitly reveal the presence of data within a dissertation.

Functional categories and data origins

In evaluating adherence to principles of the Open Data movement, it is important to recognize the wide variation in the types of data created when conducting original research, and in understanding the likelihood that such data would be valuable to other researchers for subsequent reuse and analysis. The NSB classifies data collections into three distinct functional categories, which represent distinct gradations of desirability for preservation and reuse.[17]

Research data collections support a specific project and may have no applicability beyond the focused research for which they were created. In our analysis, fully 90% (N=43) of the datasets generated for dissertation research fall into this category (see Table 3). [Insert Table

3 approximately here.] From this, it may be reasonable to conclude that significant nonconformity with the tenets of the Open Data movement does not indicate an unwillingness to make datasets available for further use or reuse, but rather merely acknowledges that other researchers are unlikely to request such access.

Resource data collections—also referred to as community data collections—serve not just a single research project, but also a community of researchers with a collaborative and overlapping research agenda. Last on this continuum are *reference* data collections, characterized by their broad scope and applicability to researchers across a wide range of disciplines and research communities. While only 10% (N=5) of the data collections discovered in our sample could be categorized as resource or reference data collections, we observed that all of these data collections appeared to be accessible to researchers and available for reuse. We also noted that three of these five resource or reference data collections were created by sociology researchers.

Decisions about data preservation (and availability for reuse) are also dependent on the origins of data collections, which, again according to the NSB, can be described in three distinct ways. *Observational* data, which are specific to a time and place and cannot be precisely replicated, may be better candidates for preservation (if they have applicability beyond a single research project) than *computational* or *experimental* data, which are, in theory, more easily replicated. Not surprisingly, within our sample, most of the dissertations in Education and Sociology contained observational data (83% and 75%, respectively). In contrast, 100% of the data collections in the field of Mechanical Engineering were computational, or a combination of computational and experimental; similarly, in Biology, 83% of the data collections were experimental, or a combination of experimental/computational or experimental/observational (see Table 4). [Insert Table 4 approximately here.]

As the discussions continue among sociologists concerning the need for policies and practices that encourage data sharing, an understanding of the correlation between data categories or data origins and the likelihood that data collections will be reused could serve to clarify or simplify the debate. However, while data-category or data-origin labels may offer convenient descriptive distinctions—and these distinctions may drive decisions about data preservation and access—such distinctions can easily become blurred. For example, a data collection that starts out as a *research* collection could be the kernel for a collection that, over time, gains status as an essential *resource* or *reference* data collection. This is especially likely in the case of dissertation-related data sets, which are the creation of early-career researchers, whose full promise and significance to the research community has yet to be determined.

Access to data collections and the role of academic libraries

By employing a very broad interpretation of the concept of "availability" of data collections within dissertations—including the presence of some summary data in tables or appendices—we were able to assert that 67% (N=32) of the dissertations in our sample made some portion of the data collection available. However, as noted above, 90% of the data collections could be categorized as research data—relevant only to the specific project treated in the dissertation—and therefore of minimal interest to subsequent researchers. Also, what we were more interested in finding—in adherence to the spirit of the Open Data movement—was access to the entire underlying dataset (raw data) from which the summary data presented in the dissertation was derived. Such access could arise if the dissertation's author had deposited her dataset in an institutional repository, on a personal website or in one of the growing number of domain specific archives such as the Association of Religion Data Archive or the Protein Databank.[18] Another viable deposit option is within a subscription-level data archive such as the Interuniversity Consortium for Political and Social Research (ICPSR). The ICPSR encourages and welcomes deposit from all social scientists and through its website offers researchers detailed guidelines for data deposit.[19] Interestingly, none of the dissertations in our sample pointed to this level of access.

We recognize that there are many informal channels by which researchers who are involved in convergent lines of scholarly inquiry may communicate and collaborate, and that such collaboration or collegial correspondence could lead to satisfactory sharing of research data collections. While informal paths to data access and data sharing may serve some members of a research community, it seems that maintaining such a haphazard and idiosyncratic system of data sharing—especially when better methods are available—is contradictory to fundamental principles of scholarly practices supported by sound and consistent methodologies. Sometimes, librarians are able to assist researchers in their hunt for elusive data collections, but in the absence of consistent mechanisms for collection, preservation, curation and dissemination of research data, there is no expectation of consistent results, so future researchers are not well served. As they do with the creation of other types of information that support the research process, researchers serve the research community through good analytic practices, which include a record of how data are stored and what procedures were used to derive results from data analysis. In the absence of such practices, it cannot be known if the data sets and data codes are available.

While Freese does not specifically cite the role of libraries in his recommendation for systematic standards for data curation, he does acknowledge that publications-related data archives such as ICPSR are well positioned to partner with researchers in sociology as they strive for uniform processes for data sharing.[20] Although this still results in certain barriers to access for subsequent researchers (access to ICPSR data collections is only available via a paid subscription), it also creates a consistently understood method for data curation, thus allowing librarians to more easily assist researchers.

Conclusions

Data are integral to the scholarly process, yet there is little indication that researchers in sociology and other disciplines are embracing the principles of the Open Data movement and moving toward a consistent method for providing access to data collections. The lack of available data for reuse by other scholars prevents replication and enhancement of existing research.

Within sociology, the debate over the need for open access to research data continues to play out, as evidenced by the nascent discussion among scholars, such as Freese and Abbott.[21] While this places sociology ahead of many other disciplines, the results of our examination of recent dissertations reveals that implementation has yet to ensue.

The need to leverage investments in research by maximizing the impact and usefulness of research results is well understood. Technological advances have raised expectations for data sharing, and financial exigencies have brought the issue into sharper focus, as grant-funding institutions are beginning to require shared access to research results and the data that support them. And yet, policy changes seem not to have gone far enough to encourage (or mandate) routine and consistent digital depositing of data. The National Science Board's *Long-Lived Digital Data Collections* provides some guidance around the requirements for maintaining access to data for replication, reuse, and enhanced research and educational needs.[22] And more recently, the Interagency Working Group on Digital Data reinforced these requirements by issuing a report that offers "a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society."[23] However, our observations indicate that new researchers neither provide nor have access to the full range of research data collections, thus impeding data reuse and replication opportunities for the advancement of knowledge.

Librarians can address some of the limitations of current practices in data sharing by incorporating into their workflow activities that promote data storage and data sharing. For example, reference librarians can work with researchers to probe for data sets that may reside in unexpected places; departmental liaison librarians can promote awareness and use of domain specific data archives; and collection management librarians can encourage commercial publishers to strive to make underlying datasets available in conjunction with the articles that reference these datasets. As librarians' roles continue to evolve in ways that cause these traditional functions to blur, there will be other opportunities to promote better and more uniform data sharing behavior, perhaps by coordinating initiatives for development of data management guidelines; or by serving as a bridge between researchers and institutional repositories in order to include data deposit as a routine function for the repository. Recognition that these functions fall naturally into the domain of librarianship is evidenced by the establishment of data curation specializations within graduate schools of library science such as those at the University of Illinois,[24] and the development of an international data curation curriculum at the library school at the University of North Carolina.[25] Lastly, by leveraging the work of organizations such as SPARC,[26] Librarians can educate and advocate for more consistent data sharing practices and standards as espoused by the Open Data movement.

Limitations and Future Research

For this study, we set forth an initial inquiry that focused on recent research in only four disciplines, and we limited our sample to a small number of doctoral dissertations within these disciplines. By gathering a more expansive sample, we could subject our results to more extensive statistical analysis. Also, it could be more informative and enlightening to compare data sharing practices in sociology against a wider selection of disciplines, and to move beyond dissertations to include journal articles, books, and other recent scholarly works among the sample of works examined. We also acknowledge that the discussion among sociologists about the need for improvement in data sharing practices is still fairly recent (Freese and Abbott both voiced their concerns in 2007), and that adoption of recommended best practices is not instantaneous. Future studies could look for evidence of improvement in data sharing practices.

Further research might also focus on the nature, extent and consequences of practices within sociology—and other social science disciplines—of withholding access to data, as has recently been examined in the life sciences. While Vogeli et al.[27] offer compelling evidence for a strong correlation between withholding of data from life sciences doctoral students and impediments to their research progress, a diminished quality of their relationship with other research professionals, and perceived negative educational experiences, it would be interesting to explore the measurable effects of similar data-withholding behavior in the social sciences.

Notes

[1] Christine L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (MIT Press, 2007), 115.

[2] "Open Science Data." Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Open_Data (accessed July 14, 2010).

[3] See the following for discussions on the merits of sharing: Stephen Fienberg, *Sharing Research Data* (Washington, D.C., National Academy Press, 1985).; Christine L. Borgman, "Data, Disciplines, and Scholarly Publishing," *Learned Publishing* 21,1 (2008): 29-38.; Eric Campbell, "Data Withholding in Academic Genetics - Evidence from a National Survey," *JAMA-Journal of the American Medical Association* 287,4 (2002): 473-480.; Gary King, "Publication, Publication," *PS: Political Science and Politics* 39,1 (2006): 119-125.

[4] Christine Vogeli et al., "Data Withholding and the Next Generation of Scientists: Results of a National Survey," *Academic Medicine* 81,2 (2006): 128-136.; Eric Campbell and Eran Bendavid, "Data-Sharing and Data-Withholding in Genetics and the Life Sciences: Results of a National Survey of Technology Transfer Officers," *Journal of Health Care Law and Policy* 6 (2002): 241.; James L. Gibson, "Cautious Reflections on a Data-Archiving Policy for Political Science," *PS: Political Science and Politics* 28,3 (1995): 473-476.

[5] "Got Data?" *Nature Neuroscience* 10 (2007): 931.

[6] National Science Foundation, "Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans," *Press Release 10-077*, 10 May 2010, http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF (Accessed June 2, 2010).

[7] National Institutes of Health, "Final NIH Statement on Sharing Research Data," (2003).

[8] Jeremy Freese, "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods & Research* 36,2 (2007): 154.

[9] Ibid., 158.

[10] Andrew Abbott, "Notes on Replication," *Sociological Methods & Research* 36,2 (2007): 212.

[11] Pao-long Chang and Pao-nuan Hsieh, "A Qualitative Review of Doctoral Dissertations on Management in Taiwan," *Higher Education* 33,2 (1997): 119.

[12] Tony Becher, "The Disciplinary Shaping of the Profession," In *The Academic Profession: National, Disciplinary, and Institutional Settings*, edited by Burton R. Clark, (Berkeley: University of California Press, 1987): 271-303; Jenny Fry, "Coordination and Control of Research Practice Across Scientific Fields: Implications for a Differentiated E-Science," In *New Infrastructures for Knowledge Production: Understanding E-Science*, edited by Christine Hine, (London: Information Science Publishing, 2006): 167-87; Samuelle Carlson and Ben Anderson, "What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use," *Journal of Computer-Mediated Communication* 12,2 (2007): 635-51.

[13] Earl Babbie, *The Basics of Social Research*. 4th ed. (Belmont, CA: Thomson/Wadsworth, 2008).

[14] National Science Board. "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," (National Science Foundation, 2005): 9.

[15] Ibid: 19.

[16] Guy Adams and Jay White, "Dissertation Research in Public-Administration and Cognate Fields - an Assessment of Methods and Quality," *Public Administration Review* 54,6 (1994): 566.

[17] National Science Board, 20-21.

18. The Association of Religion Data Archive, http://www.thearda.com (accessed July 14, 2010) and the Protein Databank, http://www.rcsb.org/pdb/home/home.do (accessed July 14, 2010) represent just two of the growing number of domain specific data archives.

19. The Interuniversity Consortium for Political and Social Research encourages and welcomes data deposits. Its website http://www.icpsr.umich.edu/icpsrweb/index.jsp (accessed July 27, 2010) includes an extensive explanation and guidelines for data deposit.

20. Freese, 159.

21. See Freese, 153-72 and also Abbott, 210-19.

22. National Science Board.

23. "Harnessing the Power of Digital Data for Science and Society Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council." (Washington, D.C.: Interagency Working Group on Digital Data), http://www.nitrd.gov/about/Harnessing_Power_Web.pdf (accessed July 21, 2010).

24 Master of Science: Specialization in Data Curation, http://www.lis.illinois.edu/academics/programs/ms/datacuration (accessed August 23, 2010).

25 DigCCurr II: Extending an International Digital Curation Curriculum to Doctoral Students and Practitioners, http://www.ils.unc.edu/digccurr/index.html (accessed August 23, 2010).

26. SPARC®, the Scholarly Publishing and Academic Resources Coalition, is an international alliance of academic and research libraries, http://www.arl.org/sparc/about/index.shtml (accessed July 27, 2010).

27. Vogeli et al., 128-36.