Challenges and Opportunities in Social Science Research Data Management

Stefan Kramer
stefan.kramer@cornell.edu
Research Data Management Librarian
Cornell Institute for Social and Economic Research (CISER)
391 Pine Tree Road
Ithaca, NY 14850
USA

Abstract: With the necessity for developing better methods for the discovery and access of available research data, mounting pressure from funding agencies to make and keep research data accessible for the long term, and the complexity of the relationships of different types and formats of files involved in many studies, social science research data management has emerged as an area of multiple challenges and opportunities for information professionals.

The *end* results of scholarly research have a long-standing tradition of being captured and disseminated through well-established publishing methods for journals and books. That's the case in the social sciences - in which I include economics and behavioral sciences in this discussion - as much as in other disciplines. Academic libraries have, in turn, taken on the responsibility of making these published results available to their faculty and students and preserving them for long term access. Within the last decade or so, the new problem has arisen there of how to preserve e-journals, e-books, and other texts that are increasingly born digital, and these problems have begun to be addressed with services such as LOCKSS ("Lots of Copies Keep Stuff Safe")[1] and Portico[2].

While the long-term preservation of published research results continues as a vital activity, universities and other research institutions are now increasingly focusing their efforts also on how to deal with the data collected *during* the course of research projects, and not necessarily dependent on whether those data collections lead to published results. On July 21, 2010, in the session on "The Future of the Academic Library: Space, Digitization, Access, and Curation in the New World of Information" during the EDUCAUSE/Cornell Institute for Computer Policy and Law, Clifford Lynch stated (1 hour, 4 mins. into the streaming video recording[3]) that the "whole matter of data management and data curation" will be a "massive game changer" for particularly research libraries in the next five to ten years.

[1] LOCKSS: How it works. lockss.stanford.edu/lockss/How_It_Works.
[2] Portico: Our Organization. www.portico.org/digital-preservation/about-us/our-organization/.
[3] Cornell University School of Continuing Education and Summer Sessions: video streams from the Institute for Computer Policy and Law, July 19-22, 2010. www.sce.cornell.edu/exec/programs.php?v=CPL&s=Live+Streaming.

In the social sciences, research data can include statistical datasets, interview transcripts, and video or audio recordings, among other formats. Numeric or quantitative data has a history of being archived in some central or institutional archives, such as ICPSR (the Interuniversity Consortium for Political and Social Research), which goes back to the 1960s, or the data archives of universities such as Yale or Cornell, which were begun in the 1970s or 1980s. Those university data archives' focus, however, was traditionally – not unlike university libraries for periodicals and books – the acquisition of data from *external* sources for use by their universities' researchers. Qualitative data, such as gathered in field observations or text content analyses, is a more recent area of concern, with efforts such as Qualidata by the Economic and Social Data Service in the United Kingdom[4].

All of these research data pose a number of challenges for long-term access and preservation, including location of authoritative copy, provenance, file and software formats, ownership and distribution rights, and knowledge of the data gathering, data transformation, and data coding processes. Confidentiality and privacy of data collected from human subjects is a problem found in many of the social sciences that are not a concern in the natural and physical sciences – those have their problems more in the sheer volume of data now being collected by increasingly powerful measurement and simulation instruments and computing resources. In the social sciences, data ownership and dissemination rights become an especially interesting problem in cases where data from different sources are merged together; for example, when a researcher's own data gets joined with data from the public domain (say, the U.S. Census Bureau) and data from a proprietary source that charges for use of their data, such as a marketing firm.

The management of research data from any single project also poses these questions: Should every version of a dataset be kept, from the first one with the originally collected information to the next one where errors were cleaned up to the next one where new variables were computed and others recoded? Should the different file formats of the same data be kept – for instance, if a Microsoft Access database was created to fill in responses during a series of interviews, but then that was transformed to be statistically analyzed with software like SPSS, Stata or SAS? Should only these aforementioned data files be preserved for the long term, or also the command, script, or syntax files that were written to run statistical analyses processes against these data files, as well as the output files that resulted from those analyses? How should the relationships of these data, command, and output files to each other be expressed to remain comprehensible, particularly to future users of those files who are unfamiliar with how they came about? These are questions, among others, that a data management plan might address.

In the United States, the typical process for faculty members to either gain tenure or have to leave their university within a certain number of years has been called "publish or perish," and that publishing has typically focused on articles in peer-reviewed journals as the final outcome – *data* as a published product of research that should also be regarded as an equally recognized contribution to scholarship is a fairly new proposition. On top of that, especially in times of budgetary shortfalls, researchers are often pushed to seek

[4] About ESDS Qualidata. www.esds.ac.uk/qualidata/about/introduction.asp.

external funding for their projects. As a result of these priorities, researchers themselves are typically focused on the data gathering efforts for the current or next project, rather than on preserving the data from the last or prior projects. Information professionals with the appropriate training or experience have a major opportunity here to fill a rapidly growing, unmet need in research data management, in the social sciences and in many other academic fields.

The anticipation (as of July 2010) that the National Science Foundation (NSF) in the USA would require plans for the management of research data with any grant application[5] within a matter of months has suddenly brought a lot of attention to this area, including from university administrators. Serious amounts of money are at stake here, as the NSF makes over 11000 new yearly funding awards and has an annual budget (in Fiscal Year 2010) of just under seven billion (7×10^9) U.S. Dollars[6]; and it can be expected that other government agencies and private foundations financing research may follow in the NSF's footsteps before too long.

As at the same time interdisciplinary and trans-institutional research becomes more frequent, universities and other research-oriented organizations face new challenges and opportunities for providing data management support to their researchers. One possible effort to address these challenges would enable collaboration among researchers around data *during* the ongoing research process, with the option of pushing *final* research data out to different, possibly multiple, long-term preservation repositories with a *single* workflow. Cornell University's Mann Library has undertaken a project for developing a Data Staging Repository (DataStaR) for this purpose.[7] Information professionals also have a role to play in developing institutional and community cultures in support of research data sharing, in the spirit of the Open Access movement that began with a focus on scholarly journal articles.

Closely aligned with the prior work of information professionals for *other* types of content than research data, enhancing the preservation, discoverability, and access of/to research data would aid research replicability, avoid duplication of data gathering efforts, and of course expose available data resources that could be of value to future research projects. However, data searches often follow an expressed information need from a user that is quite different from searches for text or images or video. In the social sciences, a typical quest for suitable data sources for a research project might include:

- Of course, the subject of the data – for example, public opinion polls about potential or declared presidential election candidates
- The time span that the data need to cover – for example, from 1960 to 2008
- The time frequency at which the data need to have been collected – for example, at least annually

[5] Mervis, Jeffrey: NSF to Ask Every Grant Applicant for Data Management Plan. news.sciencemag.org/scienceinsider/2010/05/nsf-to-ask-every-grant-applicant.html. [5.5.2010].
[6] National Science Foundation Fact Sheet. www.nsf.gov/news/news_summ.jsp?cntn_id=100595. [21.1.2010].
[7] DataStaR, a Data Staging Repository hosted by Albert R. Mann Library, at Cornell University. datastar.mannlib.cornell.edu/about.

- The geographic extent for which data need to be available – for example, for all of the United States, not just one region or state of it
- The geographic granularity for which data need to be available – for example, at the county level

At present, catalogs and search engines are generally not geared towards allowing structured search inputs and outputs with such parameters.

However, a metadata standard for describing social science research data has been developed that has the potential for making such structured queries possible by developing search tools that exploit this metadata structure: the Data Documentation Initiative (DDI).[8] Version 3 of the DDI is designed to support processes throughout the social science research data lifecycle, from study planning and instrument creation to data collection, preparation, and analysis to publication, sharing, reuse, long-term management, to discovery. Explorations are now beginning whether DDI should also be expressed in formats related to the Semantic Web, such as RDFa[9].

The first Data Summit[10] that is being hosted by the company behind the Wolfram|Alpha "computational knowledge engine" in September of 2010 may foreshadow the needed collaborations between developers of widely used search and discovery tools (with which most students and many faculty now begin their academic research) and the organizations that collect, document and archive numerous datasets that have great secondary research potential.

Some more opportunities for developing services around research data that would seem to fall squarely into the domain of library and information science professionals:

1. Providing guidance and instruction on proper citation of data sources in their papers to researchers anywhere, but particularly university students, because it can be notoriously difficult to unambiguously identify a dataset from a citation in many research papers and journal articles.
2. Improving ways of making research data discoverable, and differentiating it from other types of material, in the catalogs of libraries. Yale University, for instance, has begun work on utilizing the “Numeric data”-type of computer file code that is already defined in the MARC record format[11] for this purpose; in particular, this has the potential to help searchers distinguish between the “raw data” and statistical publications or research papers based on the data - which often have very similar, if not identical, titles, and are increasingly both in digital format.
3. Improving the bidirectional linkages between published research results and the data they are based on, so that finding a journal article can help find the research data used in the described findings, or that finding research data can help find the published results that utilized that data or part of it. An example of a high-quality

[8] DDI Alliance: What is DDI? www.ddialliance.org/what. [27.8.2009].
[9] RDFa Primer: Bridging the Human and Data Webs (W3C Working Group Note). www.w3.org/TR/xhtml-rdfa-primer. [14.10.2008].
[10] Wolfram Data Summit, Sept. 9-10, 2010. www.wolframdatasummit.org.
[11] Library of Congress: MARC 21 Format for Bibliographic Data – Computer Files. www.loc.gov/marc/bibliographic/bd008c.html. [Feb. 2010].

resource that has pursued this in the social sciences is the ICPSR Bibliography of Data-related Literature[12], focused on data held by that archive.

Lastly, there is an opportunity for making social science research data generated within an institution more easily accessible to users who may not be comfortable with the use of statistical data analysis software (yet) – those would include university students in their first years, policy makers and administrators and their support staff, and members of the public with a need for or interest in statistics related to demographics, politics, or other social phenomena. This opportunity is in converting and mounting the data via platforms that allow for interactive online exploration directly through a web browser, such as Survey Documentation and Analysis (SDA) from the U. of California Berkeley[13], Google Fusion Tables[14], or Nesstar[15]. Two examples of publicly accessible datasets for interactive online analysis are those in the SDA Archive, which include the General Social Survey and the American National Election Studies, and the *Wohlfahrtssurvey – Online* available through GESIS (Leibniz-Institut für Sozialwissenschaften)[16].

In conclusion, the management of research data in the social sciences provides both considerable challenges and opportunities to researchers, information professionals, data archivists, developers of search-and-discovery tools, and others along with their institutions and communities. Agencies that fund the undertaking of social science studies are beginning to demand sharing of and long-term access to the research data collected in those studies, which changes research data management from an activity that is worthwhile to one that becomes mandatory, and that will likely move it closer to the center of the radar screen of high-level administrations of universities and other research institutions. The way that those in need of social science data search for it is not presently well supported by many search engines or catalogs, which were developed for searching for other types of content such as texts or images. Unlike online texts in most formats, datasets are not "self-describing" in a way that allows external computerized crawling and indexing to make them searchable, so the development of metadata structures and discovery tools that truly utilize them is essential for the rapidly growing body of social science data to become and remain an asset to the scholarly and public policy communities, among others.

Research data is emerging as one the major new thrusts of information management at the beginning of the 21st century, and it will require a considerably higher level of institutional and professional commitment and attention in the coming years, including the training and education for the next generation of "data-centric" information professionals.

---

[12] ICPSR: About the Bibliography of Data-related Literature. www.icpsr.umich.edu/icpsrweb/ICPSR/citations/methodology.jsp. [Feb. 2009].
[13] SDA: Survey Documentation and Analysis. sda.berkeley.edu. [17.5.2010].
[14] Google Fusion Tables Tour. tables.googlelabs.com/public/tour/tour1.html.
[15] Norwegian Social Science Data Services: What is Nesstar? www.nesstar.com/about/background.html
[16] Wohlfahrtssurvey 1978-1998 – Online. www.gesis.org/dienstleistungen/daten/umfragedaten/wohlfahrtssurvey/wohlfahrtssurvey-online/.