MANAGING TECHNOLOGY

- Managing the Data Deluge

by Geoffrey Little

Available online 29 August 2012

In September 2010, Professor Christine Borgman of UCLA's Department of Information Studies declared that "The data deluge has arrived," and the international conglomerate IBM now contends that we are in the era of "big data" where 90% of the world's data has only been produced within the last two years.1,2 In everyday terms, more and more academic librarians are hearing about—and having to deal with—data. For the unininitiated, data can be a source of anxiety and stress. What is "data?" Why do my faculty and students want it? Where do I find it? How do I archive it? What tools or software do I need to analyze it? Is it a plural or singular noun? Until fairly recently, the scholarly outputs of researchers in the sciences, social sciences, and humanities were confined to paper documents and publications like journal articles, conference and technical papers, and monographs. Now, however, to paraphrase Alan Liu at UC Santa Barbara, everyone is producing files and librarians are being called on to help manage, curate, preserve, and facilitate access to them.3 While many of us work in libraries where a colleague is responsible for social science data, including statistics, census information, and large analog or computer datasets, the rest of us may be fielding requests for data for the first time or we may find ourselves dealing with increasing expectations from our faculty and students in terms of what the library can offer in terms of data services, specifically in how we can help interpret funding agency regulations and rules for data gathering, archiving, and digital preservation, and facilitating future access to research data through what Purdue University librarian Michael Witt calls “data literacy.”4

The Dictionary of Sociology defines data as “records of observations...from which inferences may be drawn, via analysis.”5 Another definition from the University of Toronto Data Library Service gives data as “the raw material from which statistics are created, and more often used to answer questions that ask ‘how’ or ‘why’. Data are raw, unsummarized characteristics as originally collected.”6 Data can take many forms and can be found in many formats: numbers, images, e-mails, fieldwork diaries, test scores, text notes, transcribed interviews, photographs, magnetic tape, CD ROMs, Excel files, and even punch cards. Managing research data can, to put it mildly, be a challenge and requires us to think about data curation, a function that the Data Curation Centre (DCC) in the United Kingdom describes as “maintaining, preserving and adding value to digital research data ... [Data] curation enhances the long-term value of existing data by making it available for further high quality research.”7

Why are access to data and data curation important? Making research data available and preserving it means that, aside from allowing students and faculty to conduct their research, published experiments and tests can be recreated to verify and confirm findings and results; existing data on one topic, including data gathered by national, state, or local statistical agencies, can be re-used now and in the future by researchers and policy makers in other areas; and many governmental and private funding organizations like the Howard Hughes Medical Institute, the Joint Information Systems Committee in the UK, the National Institutes of Health, and the World Bank require researchers to deposit their data in freely accessible repositories. The expectation here is that data gathered using public or philanthropic funds should be available to the public and to the worldwide research community. As I also noted in my column on the digital humanities (JAL 37:4), interest in data is not just confined to scientists or social scientists, but scholars working the arts as well. Text mining of digitized books, analysis of archival audio-visual materials, and the use of GIS to create maps or to chart historical patterns and trends are just some of the ways that humanists and artists are using data. The Digging Into Data Challenge (http://www.diggingintodata.org/), funded by a number of international agencies, encourages scholars across disciplines to think of new ways of using data in their research. Many of us work in institutions where we are only now starting to work with data or where a data program is in its infancy, or we may be starting from zero. Responsibility for data may also be unofficially or informally distributed among liaison librarians, in which case we may all have to know a little something about this kind of work. Where to start? Thankfully, there are many articles within the professional and scholarly literature that describe how libraries have set up data archiving services, workflows, and toolkits or the ways by which librarians have inserted themselves into the process of gathering and preserving data and creating programs for data instruction and reference. There are also online resources available for the data novice and expert alike. Google (www.google.com) has as suite of tools for utilizing and manipulating data like Google Fusion Tables (http://www.google.com/fusiontables/Home/), Google Earth (www.google.com/earth/index.html) and Google Maps (https://maps.google.com).

Librarians can also take advantage of education and training opportunities ranging from CurateCamp (http://curatecamp.org/), courses and sessions via the Digital Curation Centre (http://www.dcc.ac.uk/training), and the Tufts University-based Social Sciences Librarians Boot Camp (http://sites.tufts.edu/sslbc2012/). The big data conference is the International Association for Social Science Information Services & Technology (IASSIST) annual conference, which supports librarians working with information technology and social science data, and the Council for Library Resources (CLR) and the Digital Library Federation (DLF) have also recently started to offer data curation post-doctoral fellowship in academic libraries (http://www.clr.org/fellowships/datacuration). The iSchool at Syracuse University even offers a graduate certificate in data science.8

But what about finding data? Obviously, national statistical agency web sites and tools are immediate go-to sources for data, as well as resources produced by international organizations like the United

Geoffrey Little, Concordia University, Montreal, QC, Canada <Geoffrey.little@concordia.ca>.
Nations, World Bank, and OECD, but many of these powerful databases are only available via expensive subscription. In response, the “open data” movement, which seeks to make data available for free, has been gaining ground in recent years. Countries like Canada, Spain, and France, have launched free national open data catalogs, as have cities, states, and provinces around the world. DataCatalogs.org (http://datacatalogs.org/) is a great free resource curated by data experts listing international open data catalogs. Google Public Data (http://www.google.com/publicdata/directory) also points to publicly available data and allows researchers to visualize it using graphs, charts, and maps, as does Gapminder (http://www.gapminder.org/), another great free resource. IBM’s Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/) is one more popular data visualization tool. As with any resource, though, librarians need to take time to evaluate and experiment and weigh the pros and cons of using them or one versus another. Are you comfortable using a Google product given the terms of service? Will your institutional review board allow you or your researchers to use a free tool if potentially sensitive information is stored on a server you cannot access? Do you have the budget or skills to acquire and use commercial software packages? How much training do you need? Can you pool resources and technologies with other local institutions or make arrangements to share expertise?

Data curation presents another distinct set of challenges. Do you have the staff and budget to set up a full-scale repository? Do you look at proprietary software or do you explore open access options? If you already have an institutional repository (IR), can it ingest and provide access to datasets? Have you thought about metadata standards? How much of a commitment to data curation can you make without infrastructure and financial support from your provost or vice president for information technology? What is the demand for data like at your institution now compared to current and future staff time, education, and training resources? Data management goes hand in hand with data curation. In addition to deposit requirements, many funding agencies, like the National Science Foundation, require funding and grant applicants to submit a data management plan. Here again librarians may be called upon to help come up with this plan or to interpret data management and archiving provisions. Thankfully, there are good examples out there for librarians to look at. DMPTool (or Data Management Plan Tool, https://dmp.cdlib.org/) is an online resource that was developed by a number of research libraries to help streamline and simplify the data management process. Libraries need to become “contributing institutions” in order to get access to the tool, but it may be a worthwhile investment depending on the number of grants generated by your faculty and researchers. Other libraries like those at MIT, Madison-Wisconsin, and the California Digital Library have developed excellent data management guides that can be used as templates or best practices if you decide to build your own institution-specific guide. ICPSR (Inter-University Consortium for Political and Social Science Research, http://www.icpsr.umich.edu/), which is celebrating its fiftieth anniversary in 2012 and is hosted at the University of Michigan, has been a pioneer in data management and curation best practices and has a great many resources available on its web site.

Data management also requires proactive common sense. Data can be lost when faculty retire, move offices, wrap up research projects, or decide to clean their desks or hard drives, and electronic files can easily be corrupted, lost, or discarded. The goal here is to reach out to faculty sooner rather than later so that they become aware of the library’s data services. As well, graduate students may welcome opportunities to use data in their research, so liaison work is also important.

Libraries and librarians should welcome opportunities to become involved in data management and curation. They speak to our efforts to preserve and provide access to information in diverse formats, but also to our abilities to support new kinds of research and scholarship. We should not find too many reasons or excuses for fearing data. It can be overwhelming and confusing at first, and establishing a data program (large or small) involves an investment of time, energy, resources, and goodwill. We should think of data librarianship and the use of data in our daily work along the same lines as Carly Fiorina, the former Hewlett-Packard CEO and unsuccessful senate candidate, who told attendees of Oracle OpenWorld in 2004 that their work was: “about transforming data from passive to active, from static to dynamic—transforming data into insight.”

NOTES AND REFERENCES

3. Alan Liu, “Digital Humanities and Academic Change,” English Language Notes 47 (spring/summer 2009), 27.