# The Imperative for Data Curation

Joyce L. Ogburn

For additional information about this article
http://muse.jhu.edu/journals/pla/summary/v010/10.2.ogburn.html

# The Imperative for Data Curation

## Joyce L. Ogburn

The processes of creation and expression of our scientific, social, and humanistic inspirations are culminating in a vast corpus of stunning and even life-changing documents, films, recordings, Web sites, and other media, including software. Advances in technology have enabled new kinds of scholarship—the most obvious and profound impact is occurring in the realm of science.

Science is an interwoven system of experimentation, observation, verification, and replication that demands access to durable research results. Investigations into scholarly communication and research practices have brought attention to the evolving conduct of science. Service opportunities have been revealed for supporting the research process, sustaining and capturing the non-published conversations of science, and curating the resulting data.[1] An ARL workshop in 2006 elicited many salient issues regarding data curation, and a new body of literature has been building.[2] Recent developments offer compelling reasons to engage in the future of data.

Since 2008, the National Institutes of Health (NIH) have mandated that researchers deposit their peer-reviewed, NIH-funded research articles in PubMed Central.[3] The NIH already had in place a requirement for researchers to deposit their data with NIH and to do so in prescribed formats.[4] Calls for stronger data management plans by other federal granting agencies are growing. Spurred by the National Science Foundation's (NSF) initiative to build a supportive infrastructure for science,[5] campuses are forming committees and formulating Datanet proposals that involve many segments of the institution, including libraries.[6] In June 2009, Senators Cornyn and Lieberman reintroduced the *Federal Research Public Access Act* that would direct other federal agencies to require the deposit of articles in a certified repository.[7]

> **Leaving digitally based information to languish in personal electronic filing drawers amid a jumble of unrelated information and with no plans for its survival guarantees its disappearance.**

These public investments in science are predicated on the idea that the sharing of research data and publishable results stimulates additional innovation and discoveries. Such an open system of knowledge demands an infrastructure that will endure well

into the future. Leaving digitally based information to languish in personal electronic filing drawers amid a jumble of unrelated information and with no plans for its survival guarantees its disappearance. Unlike the upkeep of our academic buildings, deferred maintenance is not an acceptable strategy for preserving data.

Libraries can make the case for sustaining a role in the future of scientific research beyond the acquisition of published research results. We have been collecting social science and census data in paper and electronic formats for some time. Other data that have found their way into libraries through various channels (in faculty papers, corporate archives, family collections, and such) have gotten to us as much by happenstance as design. Because of our long existence and mandate to manage university historical material, however, quite a bit of scientific information may already reside in our archives and special collections.

Traditional library acquisition and preservation processes and methods were adequate when information was primarily in a tangible form and the responsibility for its stewardship was relatively clear. Digital information, as we know, presents a different challenge; its collection, stewardship, readability, and long-term access cannot be taken for granted, and the responsibility for its care is up for grabs. By the time knowledge in digital form makes its way to a safe and sustainable repository, it may be unreadable, corrupted, erased, or otherwise impossible to recover and use. Scientific data files may be especially endangered due to their sheer size, computational elements, reliance on and integration with software, associated visualizations, few or competing standards, distributed ownership, dispersed storage, inaccessibility, lack of documented provenance, complex and dynamic nature, and the concomitant need for a specialized knowledge base—and experience—to handle data.

Data also may be endangered by the practices of scholars who regard their data as having little value beyond the confines of a small group, a specific project, or a specified period. Data loss may occur due to lack of planning to maintain the research that was shaped or derived from scientific or engineering programs. Research information may be tossed at the completion of a project, may reside in file cabinets that are eventually emptied by retirees, or—if we are lucky—may sit in boxes at a researcher's home with the possibility of being passed to a library or archive in the future. Vast arrays of data are vulnerable to catastrophic loss without standards, systems, and services in place for their long-term support.

Many examples exist to demonstrate the extended impact of data and related information when they have been preserved for future researchers. The year 2009 marks the two hundredth birthday of Charles Darwin and the one hundred fiftieth anniversary of the publication of *On the Origin of Species by Means of Natural Selection*.[8] Present day scholars can still read, validate, and interpret the notes and collections of Darwin, as well as the research results shared with him by people from many disciplines and countries that informed his groundbreaking and paradigmatic research. His findings and theories have stimulated countless research projects and remain the foundation of modern biological science.

In *The Mismeasure of Man*, geologist, biologist, and historian of science Stephen Jay Gould exposed the biases of research on race to illuminate the resulting scientific, historical, sociological, and political ramifications.[9] His analysis and critique of cranial

measurements in the 1800s, twin studies in the 1950s, and the rise of IQ testing were possible because the data were still available for scrutiny and replication.

By combining information from multiple sources with creative approaches, existing data can be utilized more powerfully. Because both scientific and personal observations exist from the San Francisco earthquake of 1906, we have a longitudinal understanding of the movements of the earth, their probable recurrence, and the human consequences in the Bay Area. This knowledge is applicable far beyond one geographic location or era. In the medical realm, centuries of genealogical information gathered for the purpose of compiling family histories is now used in combination with genomics to make new breakthroughs on the genetic origins and inheritance of disease. In addition to professional scientists, knowledgeable and committed amateurs make new astronomical discoveries by analyzing information drawn from many sources, including their own observations and historical records.

The examples above are all compelling cases, but the preservation of this information was far from assured; it may have survived because much of it was sustainable in a tangible form. The survival of scientific data may be more certain when they are published and have a permanent home, widely recognized importance, and extensive use, especially when that use crosses disciplines. The human genome may serve as an archetype of information that is so important that it demands shared ownership and wide access; however, not all data can meet this high standard. Climate information encompasses an enormous array of disparate but interrelated data, collected by and derived from studies and records that are both current and decades, even centuries, old. At present, these data are in demand by researchers and policy makers. However obvious their importance is to us today, worldwide climate data were not always considered essential to collect and maintain; we are fortunate that so much has survived to inform present day research. The value and future of other data being produced are less clear. With our present limited knowledge of data generation and utilization, it is difficult to predict with certainty how data will be used over time and what will deserve preservation.

Moreover, the digital technologies and methodologies that have catalyzed rapid changes in science are infiltrating other fields. Numerous digital humanities initiatives have sprung up over the last decade. Researchers are also swapping, sharing, and co-opting data, software, tools, instruments, and technological architectures across disciplines. Consider this recent example of the convergence of science and humanities. In 2009, the Andrew W. Mellon Foundation awarded funding to a group of university presses to develop a model for publishing digital monographs in archaeology that will include the text and illustrations of a printed volume with its related data.[10] University presses lack the infrastructure to pull off this endeavor without partners, either on their campus or with a third party. At least some of the companion libraries will be looking to engage with their presses on this project.

With much new knowledge now being derived from creative interdisciplinary research and collaboration—dependent on data and produced and presented in evolving forms and formats—librarians will have to embrace the role of data curator to remain relevant and vital to our scholars. At present, we do not have many models to guide us in constructing appropriate services. The aforementioned NSF DataNet program will provide more grounding in data management, but practices are still in development.

Despite numerous challenges, the imperative for data curation demands that we keep these valuable assets viable and shareable across fields and for generations yet to come. The accomplishment of this monumental task will require a community effort with many aspects.

> **The funding and planning for the care and retention of data must be built into the front end, not the back end, of the research process.**

First of all, the funding and planning for the care and retention of data must be built into the front end, not the back end, of the research process. Data files must be attended to while they are compiled and analyzed in order to keep them available for a reasonable life span. This will require librarians to be conversant with the language and methods of science, at the table for campus cyberinfrastructure planning, and working with researchers at the beginning stages of grant planning. Librarians may take the lead in supporting cyberinfrastructure adoption by educating the campus and advocating for policies and actions that will propel the development and sustainability of open science.

Because data management requires intense cooperation, librarians will have to invigorate existing skills in formulating, implementing, and sustaining collaborations and partnerships that may transcend disciplines, departments, service units, and institutions. Working with data and other digital information, all being created in huge quantities, will require librarians to reimagine our roles as knowledge managers, stretch our existing skills, recruit new kinds of staff, and assume new responsibilities possibly without additional funding.

To prepare the next generation of scholars, the knowledge and skills for managing data should become part of an education process that includes opportunities for students to contribute to the creation and the preservation of research in their fields. Librarians may support this effort by developing assistantships for their data curation services and through literacy programs for scientists and nonscientists that

> **In order to grow effective future librarians, we must urge our professional graduate programs to incorporate data management into their curricula.**

teach the interpretation of data and visual representations of research findings. In order to grow effective future librarians, we must urge our professional graduate programs to incorporate data management into their curricula.

Research can advance our data management strategies and may shed light on the theoretical and practical aspects of data management that encompass use, description, standards, classification, and other knowledge domains of librarians. We will need to understand the requirements of taking care of the associated software and other information that accompanies data. Methods may be developed for determining and demarking ownership, as well as assessing the utility of data to different audiences and how this unfolds over time. Developing predictive models about the life of data that incorporates how data are and have been used will require imaginative thinking.

As librarians undertake data curation, the concepts, definitions, and scope of library collections will change. Acquiring published work will not suffice; the associated files,

and possibly even the software, that illuminate and amplify the publication will become essential components of our collections. Knowing that we should not and cannot save everything, librarians should apply archival strategies, principles, and practices to selection and curation of data. We may start by identifying at-risk data that require urgent attention. A survey of our special and general collections may discover and expose data we already own amid our other holdings. Our current methods of cataloging and metadata application will likely need enhancement.

It will be distinctly challenging to discover the sweet spot, lying somewhere between discipline specific practices and disciplinary commonalities, that is necessary to achieving scalable and sustainable services. The robustness of our research will influence the scope and type of our services for metadata, tools, discovery, and access. The results may also spawn services that facilitate linkages among researchers who work in disparate disciplines but share similar needs and that may benefit from common solutions. In order to implement this kind of service, we may have to find or create new tools for ourselves.

Data are integral to the knowledge base that underpins scholarship, provides insight into our complex world, and informs decisions about our present and our future. For data to remain a viable component of our intellectual heritage, we have our work cut out for us—let's get started.

*Joyce L. Ogburn is university librarian and director, J. Willard Marriott Library, University of Utah, Salt Lake City, UT; she may be contacted via e-mail at: Joyce.ogburn@utah.edu.*

## Notes

1. See for example: Association of Research Libraries, Joint Task Force on Library Support for E-Science, *Agenda for Developing E-Science in Research Libraries: Final Report and Recommendations to the Scholarly Communication Steering Committee, the Public Policies Affecting Research Libraries Steering Committee, and the Research, Teaching, and Learning Steering Committee* (November 2007), http://www.arl.org/bm~doc/ARL_EScience_final.pdf (accessed December 31, 2009); Christine L. Borgman. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge: MIT Press, 2007); C. Judson King et al., *Scholarly Communication: Academic Values and Sustainable Model*, paper CSHE 16.06 (Berkeley, CA: Center for Studies in Higher Education, University of California, Berkeley, 2006), http://cshe.berkeley.edu/publications/docs/scholarlycomm_report.pdf (accessed December 31, 2009); and The University of California, Office of Scholarly Communication and the California Digital Library eScholarship Program, *Faculty Attitudes and Behaviors Regarding Scholarly Communication: Survey Findings from the University of California* (August 2007), http://osc.universityofcalifornia.edu/responses/materials/OSC-survey-full-20070828.pdf (accessed December 31, 2009).
2. Amy Friedlander and Prue Adler, *To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe* (Arlington, VA, September 26–27, 2006), http://www.arl.org/bm~doc/digdatarpt.pdf (accessed December 31, 2009).
3. U.S. Department of Health and Human Services, National Institutes of Health Public Access, "NIH Public Access Policy," U.S. Department of Health, publicaccess.nih.gov/ (accessed December 31, 2009).
4. Office of Extramural Research, NIH Data Sharing Policy, U.S. Department of Health and Human Services, http://grants.nih.gov/grants/policy/data_sharing/ (accessed December 31, 2009).

5. The National Science Foundation Cyberinfrastructure Council, *Cyberinfrastructure Vision for 21st Century Discovery* (March 2007), http://www.nsf.gov/pubs/2007/nsf0728/nsf0728. pdf (accessed December 31, 2009).

6. The National Science Foundation Office of Cyberinfrastructure, *Sustainable Digital Data Preservation and Access Network Partners DataNet Program Solicitation NSF 07–601*, http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm (accessed December 31, 2009).

7. Open Access News: News from the Open Access Movement (June 25, 2009), "Sens. Cornyn & Lieberman Team Up To Increase Public Access To Taxpayer Funded Research," Peter Suber, editor, http://www.earlham.edu/~peters/fos/2009/06/frpaa-public-access-mandate-re.html (accessed December 31, 2009).

8. Charles Darwin, *On the Origin of Species by Means of Natural Selection* (London: John Murray, 1859).

9. Stephen Jay Gould, *The Mismeasure of Man* (New York: Norton, 1981).

10. University of Colorado, "Mellon Grant to Fund 'Archaeology of the Americas' Project" (April 6, 2009), https://www.cu.edu/content/mellonfoundationgrantfund%E2%80%98archaeologyamericas%E2%80%99ebookproject (accessed December 31, 2009).