Sayeed Choudury

Data curation

An ecological perspective

Editor's note: The last several Scholarly Communication columns have focused on strategies for campus engagement in the here-and-now. This issue's column looks ahead to consider how the library's role within the system of scholarly communication may be changing in a more computationally driven research environment. Sayeed Choudhury draws inspiration from the natural world to explore the need for a diversity of library roles and services that support the curation and scholarly use of digital data of all forms.

The library community has shown a great deal of interest regarding potential roles to support new forms of scholarship often called "eScience."[1] Scientific research has indeed become increasingly data-intensive, but the "eScience" label omits the humanities and social sciences, where scholars from a diverse range of disciplines are also exploring new modes of research and teaching using data. For example, social scientists are accessing data from fields such as the health sciences and environmental sciences and using tools such as geographical information systems to study the connection between health and personal relationships or environmental conditions. The National Endowment for Humanities (NEH) recent solicitation "Digging into Data" represents important acknowledgment of such developments within the humanities. Created in part to "promote the development and deployment of innovative research techniques in large-scale data analysis," this program follows others in adopting a broad definition of data to include almost any information that can exist in digital form.[2] While NEH administered this solicitation, three other agencies—the National Science Foundation (NSF), the UK Joint Information Systems Committee (JISC), and the Canadian Social Sciences and Humanities Research Council (SSHRC)—provided funding to support Digging into Data. This diverse combination of funding agencies from three countries provides evidence of widespread, growing interest for data-driven scholarship within the humanities. Fundamentally, there is a shift from a document-centric view of scholarship to a data-centric view of scholarship, which has promoted recent developments of cyber-infrastructure.

Data curation

One of the most important opportunities for libraries to partner in cyber-infrastructure development relates to data curation. The Sheridan Libraries at Johns Hopkins University have collaborated for years with astronomers to better understand data curation requirements.[3] This track record of research and development has culminated in the Sheridan Libraries leading one of two existing awards through the NSF DataNet program.[4] The goal of the DataNet program is to develop "a set of exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning."

Sayeed Choudury is associate dean and director of the Hodson Digital Research and Curation Center at Johns Hopkins University, e-mail: sayeed@jhu.edu

Contact Mike Furlough—series editor, assistant dean for scholarly communications, and codirector of the Office of Digital Scholarly Publishing at Penn State University—with article ideas, e-mail: mfurlough@psu.edu

The award, known as the Data Conservancy, outlines a shared vision: *data curation is a means to collect, organize, validate, and preserve data so that scientists can find new ways to address the grand research challenges that face society.* The overarching goal of the Data Conservancy is to support new forms of inquiry and learning to meet these challenges through the creation, implementation, and sustained management of an integrated and comprehensive data curation strategy. The Data Conservancy will accomplish this by combining user-centered design methodology to guide the immediate development process, with innovative longer-term information science research to identify and fully understand data practices and curation needs across our initial scientific domains of astronomy, earth sciences, life sciences, and social sciences. The partner institutions of the Data Conservancy feature a diverse array of expertise and experience related to domain science, information science, computer science, user-centered design, digital libraries, and digital preservation.[5] From the perspective of the library community, one of the major goals of the Data Conservancy is to provide a potential blueprint for research libraries in the data age.[6]

At this point, we are using the Open Archives Initiative–Object Reuse and Exchange (OAI-ORE) modeling protocol to organize our work. OAI-ORE "defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video."[7] OAI-ORE explicitly recognizes that new forms of scholarly communication comprise multiple objects that are distributed across multiple locations. While journal articles have always referenced other work through citations, readers access these articles as self-contained resources on the page or screen. However, future "publications" may be more expansive and include many types of information, such as datasets managed within the Data Conservancy, plain text documents within an institutional repository, a spreadsheet within a disciplinary repository, and a blog supported by a publishing platform hosted by a professional society or even a commercial publisher. The relationships among these disparate resources would be critical to maintain. Without any one of these particular components, the overall intellectual coherence of the work would be compromised. Preservation activities must explicitly account for not only the content, but also the context, especially as it relates to connections between objects. I would submit that the greater integration and cohesion that occurs seamlessly within such a network, the greater the prospects for preservation and access in the long-term.

The river metaphor

I was invited to last year's ALA Annual Conference to speak about digital preservation, especially with the context of data curation. I accepted the invitation because I believe it is critical to raise awareness and understanding about the libraries' potential role in this regard. However, I also realized that conveying a message about potential roles for libraries in an accessible manner was a challenging task, particularly because I believe there is no single, correct strategy or approach. I found inspiration for my presentation from an unlikely source: BBC's documentary *Planet Earth*. In particular, it was the episode of *Planet Earth* that focused on fresh water rivers that helped form and coalesce this metaphor: *Libraries should preserve content in the same manner that the Earth preserves its rivers.*

Upon initial reflection, this might seem like an odd metaphor since one might assume the Earth does not preserve its rivers. It is reasonable to assert that the Earth does not explicitly preserve its rivers, but through its natural course of its overall "framework," rivers are indeed preserved in a seamless manner that supports a wide variety of uses. Fresh rivers originate within the deep freeze snow packs of mountains. This literal cold storage provides the foundation from which all rivers flow. The snow packs are necessary, but they are not sufficient to support the river system.

As snow melts, the water enclosed within begins to flow along the mountainsides into an incredibly rich diverse set of environments. Once downstream, the Earth's rivers go through fast and rapid flows; waterfalls; deep, slow meandering areas; and eventually discharge into oceans. Evaporation from the oceans enters the clouds and eventually snowfall replenishes the snow on the mountains restarting and recharging the cycle. The richness and diversity of life—sustained and nourished by access to water in different states—would not be possible without all of the ecosystems. We might compare different scholarly communities and uses of content to the different ecosystems of the river network. None of these ecosystems is inherently more important or superior to any other; each of them represent an integral part of a greater whole.

With this metaphor in mind, the Sheridan Libraries' data curation infrastructure development through the Data Conservancy may be thought of as mountain building. We are focusing on developing the "deep, cold" storage environments that represent the foundation for preservation. It is tempting to assume that we need only this type of infrastructure to support data-intensive scholarship. While we will need multiple, varied instances of this type of deep infrastructure, it is important to remember the model of the Earth's river network. Just as our planet would not exist in its current form with only mountains, the library community might consider the diverse and complementary roles that individual libraries might contribute to an overall data curation network. It is possible that certain large research libraries might undertake the type of infrastructure development starting at the Sheridan Libraries or focus on specific components, such as distributed storage systems. Other libraries might wish to focus on the "downstream" components by supporting different types of repositories (e.g., institutional or disciplinary) or applications such as publishing platforms. Other libraries may elect to focus on the human elements of infrastructure by training current librarians to become "data scientists" (or "data humanists") who possess both domain science and data management expertise. Libraries may choose to focus on developing metadata standards or preservation policies. The Data Conservancy's strategy includes the development of a data framework that might support this type of diverse involvement by the library community. Ideally, libraries would make such decisions about their roles based on local assessments of requirements with an understanding of the global context. Perhaps the mantra for such decisions might be "Act locally, participate globally."

It is my hope that libraries might find inspiration in the work of the Data Conservancy, but also analyze critically and deeply how to adopt, adapt, or modify the environments in which our libraries operate. The Data Conservancy will continue to disseminate its findings through various channels, but it is important that we realize its work constitutes one stream in an overall network of rivers that need to merge together seamlessly.

Notes

1. For examples, see the Association of Resource Libraries' E-Science Survey Resource Page at www.arl.org/rtl/eresearch/escien/esciensurvey/index.shtml.

2. www.diggingintodata.org/Home/tabid/149/Default.aspx

3. I have discussed this effort in more detail in "Case Study in Data Curation at Johns Hopkins University," *Library Trends,* Volume 57, Number 2, Fall 2008, 211–20

4. www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI

5. Funded partners in the Data Conservancy include: Cornell University, DuraSpace, Marine Biological Laboratory, National Center for Atmospheric Research, National Snow and Ice Data Center, Portico, Tessella, Inc., University of California-Los Angeles, and the University of Illinois at Urbana-Champaign.

6. A video presenting a more detailed discussion of the Data Convervancy can be found at bit.ly/1fLQmQ.

7. Open Archives Initiative. Object reuse and exchange. Retrieved March 10, 2010, from www.openarchives.org/ore/.