E-Content

By Donald J. Waters

"Doing Much More Than We Have So Far Attempted"

The broadening deployment of computer-based data-conversion and data-capture instruments and sensors has greatly expanded the scale of humanistic, social, and scientific data for scholars to digest. The Sloan Digital Sky Survey is one example of this expansion; the Google library project is another. Although the systematic exploration of large quantities of information is not a new scholarly practice, what does seem new is the formalization of the very traditional interpretive activities of data-mining, pattern-matching, and simulation in powerful algorithms that represent large and complex sets of data in terms of multiple features and variables that can be analyzed, tested, replicated, and changed at the scale and speed afforded by advanced computation.

The promise of these new automated capabilities is that new knowledge can be created in ways that were not previously possible. To achieve this promise in relation to the vast arrays of electronic content that are appearing in so many different fields of scholarly pursuit, institutions of higher education face a key strategic challenge. Individually and collectively, they must mobilize their resources to create a dependable, deeply scaled, and flexible infrastructure to help faculty and students interact with the electronic content in all the ways associated with rigorous scholarship, including discovering evidence, aggregating it, arranging and editing it for use, analyzing and synthesizing it, and disseminating the results through reports and teaching. To meet this challenge, colleges and universities must address three broad areas of need.

Aggregation for Discovery and Research

One of the fundamental building blocks of scholarly activity is search—both for basic discovery and for more complex analysis and synthesis. Over the last fifteen years, search over a growing body of content available electronically on the Internet has been a growth industry, and search for discovery has become almost a commodity item, having been the subject of intense investment and development by Google, Amazon, Yahoo!, Microsoft, and their predecessors and competitors. But search is effective as a discovery tool only insofar as a sufficiently rich body of sources is comprehensively aggregated to be worth searching. The unsung hero of the success of the search engine industry is the successful aggregation of sources at scale, which includes gathering content in disparate formats from multiple sources operating under a variety of business models and intellectual property regimes, and then organizing and indexing those data for rapid delivery.

But search for discovery is only the beginning of the scholarly process. Scholars then must zero in on the subsets they find—the primary and secondary source objects of interest to their work. They need to pull together these selected subsets for deeper analysis. The process of aggregation at this stage is more difficult and complicated because data need to be reviewed for anomalies, normalized, and prepared in a more rigorous fashion than is likely to be necessary or affordable for the commodity search engines. Provenance and authenticity of the information need to be established; rights need to be cleared; databases and database schemas need to be created; textual objects may need to be translated and marked up for grammatical and structural features, as well as semantically according to certain knowledge structures; numeric data may need conversion to common measures; assumptions and guesswork need to be carefully documented; and provision needs to be made to ensure that the data are maintained and can be reliably cited over time. The maintenance and preservation functions compose what is coming to be known as *data curation*, but the broader set of computationally-based research practices define the domain of *informatics*, which transformed the field of biology beginning in the 1980s and which is gradually being applied in other fields of study today.

Discipline-Based Informatics

Given the broad need for data curation and informatics to be applied to scholarly aggregations of digital and digitized information, a second key strategic issue for institutions is to recognize that these practices cannot be deployed uniformly and in one fell swoop but instead require close attention to disciplinary use and requirements. As numerous studies have shown, investments in automation and computational methods are highly uneven, with some fields bursting with energy and creativity and others operating within relatively static paradigms. For example, under the auspices of Berkeley's Center for Studies in Higher Education, Jud King, Diane Harley, and several colleagues explored the role of tenure and promotion in the spread of innovative computationally-based forms of scholarship and publication. They found that recognition of such practices in promotion and tenure processes occurs slowly in general but that there is significant variation at the discipline level.[1]

Moreover, and perhaps more important, this study made the useful distinction between formal and informal modes of communication and observed that the formal modes, such as publication in peer-reviewed books and journals, tend to be most deeply resistant to change. After all, the formal means of establishing scholarly credentials are the basis on which institutional position, rank, and salary are determined, and few scholars are prepared to take significant action that would disrupt their means of livelihood. The authors also observed, however, that the informal realm is where scholars work with each other on a daily basis, consulting with one another and letting each other know which technique worked and which did not and what new discoveries they have made. In this informal realm—at the edge of the reputational and promotional system, where credentials are being formed rather than fixed—innovation is easier and more likely to occur. And it is here that scholars tend to develop the new and specialized discipline-based methodologies—the informatics of standards and practices—that are needed to identify, mark up, manage, preserve, and develop the algorithms for exploring the large volumes of digital information with which they need to work: economists with tabular data in government publications; literature scholars with literary texts from various genres; social historians with contemporary accounts of various aspects of social life; ethicists with case studies of ethical dilemmas; art historians with evidence about the context of artists and their art; and so on. Colleges and universities can perhaps best assist scholars in these disciplines by deploying librarians and information technology specialists to help experiment with and implement these new data-curation and informatics practices.

New Publication Emphases

A third strategic issue needing the attention of higher education institutions is publication. The important distinction made above between formal and informal modes of scholarly communication helps explain why the physics ArXive, to which all high-energy physicists routinely deposit their papers, continues to exist alongside of rather than, as some have promised for almost two decades, instead of the publication system to which they also continue to submit their papers: one is an informal mode of communication, and the other is formal. The innovative automation of the preprint process in the ArXive in the early 1990s was built on a stunning ethnographic insight about the informal scholarly communications process in physics, and it has been usefully extended to other science and social science fields in which there have been informal traditions of circulating preprints and working papers. Little innovation has occurred in this area since the initial breakthrough, and as Paul Ginsparg recently reported,[2] even the code base for the ArXive system has changed little since the mid-1990s. Real innovation in scholarly publication is now occurring elsewhere in the formal and informal systems of communications, and continued attention to the potential interaction between preprints and formal publication threatens to divert resources from other areas where they might be needed and better invested.

One of these other areas involves bringing publication expertise to bear on the construction and curation of electronic data. Some journal and book publishers are beginning to incorporate or refer to original datasets on which new publications are based. However, scholars in some fields are thinking even more innovatively and are trying to build peer-review systems around the data so that they can be judged formally on qualities such as coherence, design, consistency, and reliability of access. With JISC support in the United Kingdom, scientists and professional associations in the field of meteorology have joined to establish a new kind of electronic publication called a *data journal*, to which practitioners would submit data sets for peer review and dissemination.[3] In the field of nineteenth-century literary studies, Jerry McGann at the University of Virginia, with support from the Andrew W. Mellon Foundation, has organized scholarly societies into a federation for the purpose of providing peer review for data in the form of online documentary editions of nineteenth-century authors.[4] And Bernard Frischer, a specialist in online virtual reconstructions of archaeological sites, has received support from the National Science Foundation to plan a journal-like outlet that would provide peer review of virtual reconstructions.[5] More research and experimentation with forms of peer-reviewed data could have significant impact in helping to organize the field of data curation and informatics, provide additional information for promotion and tenure committees, and avoid wasting resources in a frontal assault on a long-established and, by many accounts, still highly valued system of formal publication in books and journals.

■ ■ ■

Not long ago, a Mellon grantee explained the need for features that he wanted to add to a database of images, measurements, virtual reconstructions, and other representations of the features of several hundred churches in medieval France. He said: "The database cannot fully provide what a good teacher can—but it can do much more than we have so far attempted."[6] He is right, and we need to help him and other scholars with similar ambitions.

Notes

1. Diane Harley, Sarah Earl-Novell, Jennifer Arter, Shannon Lawrence, and C. Judson King, "The Influence of Academic Values on Scholarly Publication and Communication Practices," Research and Occasional Paper Series, CSHE.13.06, September 2006 (Berkeley, Calif.: Center for Studies in Higher Education), <http://cshe.berkeley.edu/publications/publications.php?id=232>.
2. Paul Ginsparg, "Read as We May," presentation at the De Lange Conference VI, Rice University, March 6, 2007, webcast available at <http://webcast.rice.edu/webcast.php?action=details&event=985>.
3. Alan Gadian, "The Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) Project," <http://www.see.leeds.ac.uk/research/ias/dynamics/current/ojims.html>.
4. Bethany Nowviskie and Jerome McGann, "NINES: A Federated Model for Integrating Digital Scholarship," September 2005, <http://www.nines.org/about/9swhitepaper.pdf>.
5. The SAVE (Serving and Archiving Virtual Environments) project. See <http://www.iath.virginia.edu/save/>.
6. Stephen Murray to Suzanne Lodato, e-mail, May 4, 2007, quoted by permission.

Donald J. Waters is Program Officer for Scholarly Communications at the Andrew W. Mellon Foundation.