The "Digital" Scholarship Disconnect

By Clifford A. Lynch

EDUCAUSE recently kicked off a discussion with Clifford A. Lynch, Executive Director of the Coalition for Networked Information, by asking: "How would you define ***digital scholarship****?"*

Digital scholarship is an incredibly awkward term that people have come up with to describe a complex group of developments. The phrase is really, at some basic level, nonsensical. After all, scholarship is scholarship. Doing science is doing science. We don't find the Department of Digital Physics arguing with the Department of Non–Digital Physics about who's doing "real" physics.

ILLUSTRATION BY LORENZO PETRANTONI, © 2014

Interestingly, one of the first terms that people used for *digital scholarship* as a large-scale phenomenon was *e-science*; this was popular in the United Kingdom in the very late 1990s and early 2000s. While helpful for funding agencies, the term puzzled scientists, who might say: "We don't do e-biology. We do biology. And in our biological research, we use technologies that are constantly changing and improving."

When e-science came to the United States—as an organized effort within the federal funding agencies and especially the National Science Foundation—the term used instead was *cyberinfrastructure,* further adding to the confusion. *Cyberinfrastructure* really was an umbrella term for the high-performance computing and networking and scientific data-management programs that supported these new technology- and data-intensive scholarly practices and sometimes, by extension, the scholarly practices themselves.[1]

Soon, people wanted to start talking more broadly about newly technology-enabled scholarly work, not just in science; in part this was because of some very dramatic and high-visibility developments in using digital technology in various humanistic investigations. To do so, they came up with the neologisms we enjoy today—awful phrases like *e-scholarship* and *digital scholarship*.

Having said that, I do view the term *digital scholarship* basically as shorthand for the entire body of changing scholarly practice, a reminder and recognition of the fact that most areas of scholarly work today have been transformed, to a lesser or greater extent, by a series of information technologies:

- High-performance computing, which allows us to build simulation models and to conduct very-large-scale data analysis
- Visualization technologies, including interactive visualizations
- Technologies for creating, curating, and sharing large databases and large collections of data
- High-performance networking, which allows us to share resources across the network and to gain access to experimental or observational equipment and which allows geographically dispersed individuals to communicate and collaborate; implicit here are ideas such as the rise of lightweight challenge-focused virtual organizations

All of these information technologies have contributed to changing the practices of scholarship, with an increased emphasis on the good management of research data (either the evidence that has been collected or the research outputs that are reflected in data sets) and with a growing number of discussions about best practices for sharing and reusing and recombining data in various ways. Particularly in the humanistic disciplines but legitimately beyond them, there's a new conversation emerging about the evidence base (the "cultural record") and what that looks like today: we need to manage this to support future scholarship.[2]

The argument has been made that the incorporation of these technologies led to the establishment of a new, fourth paradigm for science—data-intensive science—to accompany the long-standing traditional observational and experimental approaches and also the computational approaches (e.g., approximation, simulation) that have been emerging since the middle of the twentieth century.[3] From this perspective, one might make at least a weak case for speaking about a "digital" physicist in contrast to a theoretical or experimental physicist, but I don't find the distinction very helpful. If anything, I suspect that computation and data-intensive approaches have led to a blurring of the old observational/experimental dichotomy.

Particularly in the humanistic disciplines but legitimately beyond them, there's a new conversation emerging about the evidence base and what that looks like today.

In addition, numerous other digital-information-related technologies continue to be essential to progress in various scholarly disciplines, even though these digital developments are not often listed as part of the traditional core of cyberinfrastructure. Consider CERN's Large Hadron Collider (http://home.web.cern.ch/topics/large-hadron-collider), which incorporates extensive digital technology in data capture and reduction, as well as in the downstream data analysis conducted through traditional cyberinfrastructure facilities. Certainly before digital times, many fields of scholarship relied heavily on the ability to capture, share, and analyze various kinds of images taken at various wavelengths. The whole world of imaging has been transformed by sensors such as charge-coupled arrays and all of the digital signal-processing that accompanies them, creating significant new opportunities in all kinds of scholarly work, from biomedicine to the humanities.

We now have enormous curated databases serving various disciplines: GenBank (http://www.ncbi.nlm.nih.gov/genbank/) for gene sequences; the Worldwide Protein Data Bank (http://www.wwpdb.org/) for protein structures; and the Sloan Digital Sky Survey (http://www.sdss.org/) and planned successors for (synoptic) astronomical observations. All of these are relied upon by large numbers of working scientists. Yet the people who compiled these databases are often not regarded by their colleagues as "real" scientists but, rather, as "once-scientists" who got off-track

and started doing resource-building for the community. And it's true: many resource-builders don't have the time to be actively doing science (i.e., analysis and discovery); instead, they are building and enabling the tools that will advance the collective scientific enterprise in other, less traditional ways. The academic and research community faces a fundamental challenge in developing norms and practices that recognize and reward these essential contributions.

This idea—of people not doing "real" research, even though they are building up resources that can enable others to do research—has played out as well in the humanities. The humanists have often tried to make a careful distinction between the work of building a base of evidence and the work of interpreting that evidence to support some particular analysis, thesis, and/or set of conclusions; this is a little easier in the humanities because the scale of collaboration surrounding emerging digital resources and their exploitation for scholarship is smaller (contrast this to the literal "cast of thousands" at CERN) and it's common here to see the leading participants play both roles: resource-builder and "working" scholar. A pathbreaking case study in the humanities is *The Valley of the Shadow* (http://valley.lib.virginia.edu/), a database constructed by Edward L. Ayers and William G. Thomas III at the University of Virginia beginning in 1993. Looking at two American communities, one Northern and one Southern, in the time just before the Civil War, the project has constructed an extensive database comprising everything from tax and census records to newspapers and geospatial surveys. Thomas and Ayers used this enormous database in their teaching and later wrote an article, drawing extensively from the database, on the difference that the institution of slavery had made to the structure of the economy and the social arrangements in the two towns. Whereas footnotes normally point to inaccessible material such as old newspapers and archives, Thomas and Ayers assembled these sources underneath the online version of their article, so that the reader could move back and forth between the evidence and the analysis.[4] That setup was a true breakthrough demonstration not only of how a digital marshaling of evidence could be put to powerful use in the humanities but also of the distinction between resource-building and analysis.

Still, in all of these examples of digital scholarship, a key challenge remains: How can we curate and manage data now that so much of it is being produced and collected in digital form? How can we ensure that it will be discovered, shared, and reused to advance scholarship? We are struggling through the establishment of institutions, funding models, policies and practices, and even new legal requirements and community norms—ranging from cultural changes about who can use data (and when) to economic decisions about who should pay for what. Some disciplines are less contentious than others: for example, astronomy data is technically well-understood and usually not terribly sensitive. Reputation, rather than commercial reward, is wrapped up in astronomical discoveries, and there is no institutional review board to ensure the safety and dignity of astronomical objects. On the other hand, human subjects and their data raise an enormous number of questions about informed consent, privacy, and anonymization; when there are genetic markers or possible treatments to be discovered or validated, serious high-value commercial interests may be at stake. All of these factors tend to work against the free and convenient sharing of data.

More and more scholars, in all fields, are recognizing that they need to be thinking about data as an explicit part of their research outcomes and their research work.

Another, closely related challenge is long-term funding for data resources.[5] The science funding agencies tend to have a fairly short-term view. Since they want to be funding today's breakthrough research, they would like to make grants that run a few years. They don't like to go much further out than that time frame because they need to constantly reassess and fund new opportunities and new work. They will (often reluctantly) fund certain long-term community-wide resources, such as GenBank; they view these resources as somewhat akin to major scientific instruments that are shared by a community. But even large-scale scientific instruments typically have a defined life-cycle, and the repeated "renewal" of funding for these community resources runs counter to the institutional culture among funders. As a result, it is not uncommon for researchers who are writing a proposal involving data that doesn't fit within the scope of an existing community repository to state in their data-management plan: "We guarantee to keep our data available for ten years and will build the cost of doing so into the grant proposal, and at the end of ten years, we will have some kind of reassessment process to figure out if anybody thinks the data is still worth keeping and if anybody is willing to pay to keep it."

I do think we have made substantial progress. One example involves the just-noted data-management plans that are now required as part of grant proposals. In recent years, funding agencies have gradually begun to recognize that data *is* an important asset that comes out of the work they fund. They have thus started to require, through the inclusion of data-management plans, that the research proposals be explicit about data as an asset and make commitments regarding the management of this data; review of

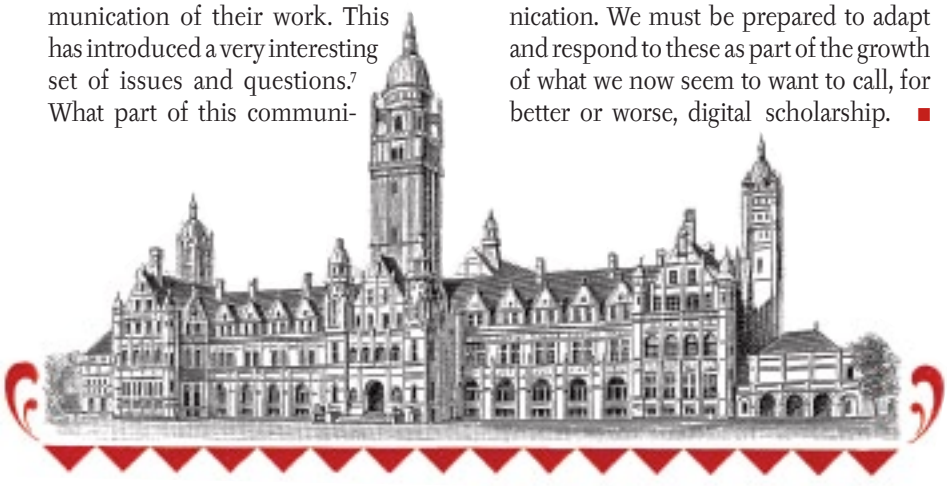
the data-management plan is part of the overall evaluation of the proposal.

In addition, although we may still have some trouble with asset discovery and how to describe data, especially in environments where the data is being reused for purposes distant from the original intent that led to collection of the data, we've made a good deal of progress on the technical systems for storing data. We have also seen the establishment of a number of new data archives on a disciplinary or institutional basis. And we have seen advancements in the standards and practices that allow the linkage of data sets to journal articles; when someone writes about an experiment, he or she can tie the analysis in the article quite tightly to specific data sets that can be made available so that others can reproduce the results or redo the analysis under different assumptions. More and more scholars, in all fields, are recognizing that they need to be thinking about data as an explicit part of their research outcomes and their research work.[6]

On a final note, I have talked above mostly about changes in the practice of scholarship. But changes in the *practice* of scholarship need to go hand-in-hand with changes in the *communication* and *documentation* of scholarship. We're starting to see this phenomenon pick up steam. Increasingly, scholars want to include visualizations and interactive models and the like as part of the communication of their work. This has introduced a very interesting set of issues and questions.[7] What part of this communication of their work fits inside the traditional scholarly publishing framework of journal articles and monographs? What part should go through new digital channels, such as project websites? What part should be considered data, to be archived and handled in a data-management setting, and how should this be woven into the overall exposition of the research? In all of these cases, how—and through what organizations and mechanisms—should the work be organized, preserved, and evaluated?

Over the years, we have found answers to these questions for the traditional journal and book publishing system (though I, at least, am not convinced they are always good answers—or answers that will be very helpful in guiding us into the future). For example, we have well-established and well-understood refereeing systems and other quality-control and reputational mechanisms in the traditional publishing world. As essential milestones to promotion and tenure in many disciplines, scholarly print monographs are reviewed in various venues, leading to a collection of public reviews that augment the review process implicit in the particular publisher's decision to publish the monograph. Libraries preserve the published record.

The scholarly community needs to find new answers to these questions of organization, preservation, and evaluation in the digital environment. Changing scholarly practices are leading to changing behaviors in scholarly communication. We must be prepared to adapt and respond to these as part of the growth of what we now seem to want to call, for better or worse, digital scholarship. ■



Notes

1. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (Daniel E. Atkins, chair), January 2003, http://www.nsf.gov/cise/sci/reports/atkins.pdf.
2. See Clifford Lynch, "Memory Organizations and Evidence to Support Scholarship in the 21st Century," 2012 Windsor Lecture, University of Illinois at Urbana-Champaign, April 17, 2012, http://www.lis.illinois.edu/events/2012/04/17/windsor-lecture-clifford-lynch (audio: http://waterfall.lis.illinois.edu/dl/events/windsor_lecture/windsor_apr17_12.mp3), and also Clifford Lynch, "Challenges of Stewardship at Scale in the Digital Age," lecture given at Indiana University, Bloomington, January 30, 2014, https://www.youtube.com/watch?v=rfvLlQ2nZj0. A paper based largely on these two lectures is in preparation.
3. For example, see Tony Hey, Stewart Tansley, and Kristin Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Redmond, Wash.: Microsoft Research, 2009), http://research.microsoft.com/en-us/collaboration/fourthparadigm/.
4. William G. Thomas III and Edward L. Ayers, "The Differences Slavery Made: A Close Analysis of Two American Communities," *American Historical Review*, vol. 108, no. 5 (December 2003), pp. 1299–1307, online version: http://www2.vcdh.virginia.edu/AHR/.
5. See *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, Final Report from the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (Francine Berman and Brian Lavoie, co-chairs), February 2010, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.
6. On the issue of data citation, see Paul F. Uhlir, *For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*, National Research Council (Washington, D.C.: National Academies Press), 2012, http://www.nap.edu/catalog.php?record_id=13564.
7. See *CTWatch Quarterly*, vol. 3, no. 3 (August 2007), issue on "The Coming Revolution in Scholarly Communications and Cyberinfrastructure," http://www.ctwatch.org/quarterly/archives/august-2007, and also the work coming out of the Force conference and Beyond the PDF meetings of Force11, https://www.force11.org/.

(Photo credit: Cecilia Preston)

Clifford A. Lynch (clifford@cni.org) is Executive Director of the Coalition for Networked Information (CNI).