On the Threshold of Cyberscholarship

Ronald L. Larsen

Journal of Electronic Publishing

Volume 11, Issue 1, Winter 2008

DOI: http://dx.doi.org/10.3998/3336451.0011.102 [http://dx.doi.org/10.3998/3336451.0011.102]

Abstract

The widespread availability of digital content creates opportunities for new forms of research and scholarship that are qualitatively different from those represented in traditional scholarly literature. To capitalize on these opportunities, digital content must routinely be collected, managed, and preserved in ways that are significantly more rigorous than conventional methods. A new form of infrastructure is required to ensure that digital content, including research products and primary sources, is readily available, accessible, and usable.

The US National Science Foundation (NSF) and the UK Joint Information Systems Committee (JISC) co-sponsored a workshop to consider the implications of digital content on network infrastructure. This paper is based on the results of that workshop, which is more fully elaborated in the final report. [1] [#N1]

The rapid growth and widespread availability of network-accessible scholarly materials in digital form are creating opportunities for new forms of research and scholarship that are qualitatively different from those represented in the traditional scholarly literature. Whereas early digitization projects rarely did more than render traditional analog publications in a digital medium, more recent efforts are beginning to explore the transformative potential of digital scholarship. The success of these projects has depended on a shared infrastructure including not only networks, but also repositories and their related services. In addition to advancing conventional research programs by independent and small teams of investigators, network-accessible digital repositories are opening new opportunities for larger teams of geographically distributed collaborators. Early and visible success among a few of these projects has fostered the perception that scholarship in the 21st century is becoming a truly collaborative endeavor, resulting in communities of scholars routinely sharing data. This perception is enhanced by the widely publicized deployment of institutional repositories, with the expectation that scholars would eagerly contribute their data to the information commons. The reality is that progress has been disappointingly spotty.

The US National Science Foundation (NSF) and the UK Joint Information Systems Committee (JISC) sponsored an invitational workshop in April 2007 to explore the seeming contradiction between the explosive growth of digital materials and the slower-than-expected adoption of digital repositories throughout the scholarly community.

The workshop concluded that a new form of infrastructure, called *content infrastructure*, is needed, situated within a broader cyberinfrastructure, [2] [#N2] to support the emergence of *cyberscholarship*.

Motivation: Emerging Opportunities for Scholarship

Computation is clearly becoming an increasingly vital component of research, particularly as the scale and complexity of datasets exceed the analytical capacity of individuals, and the data itself is becoming increasingly important as a scholarly resource. Gregory Crane, a classics professor at Tufts University, observed in a recent seminar, “when collections get large, only the computer reads every word.” [3] [#N3] Lewis Lancaster (UC Berkeley) and his colleagues demonstrate the creative potential of computation to interpret large and diverse text and data resources in their work on temporally dynamic maps in the Electronic Cultural Atlas Initiative. [4] [#N4] Mining a variety of primary and secondary sources, from original texts to MARC records from the Library of Congress, Lancaster’s team constructed animated maps to visualize geospatial relationships that are otherwise obscure (see Lancaster’s AVI demo at http://www.sis.pitt.edu/~egyptdlw/presentations.html [http://www.sis.pitt.edu/~egyptdlw/presentations.html] for an example of visualizing the spread of the Mongol empire over a period of 300 years). Related examples and possibilities abound in the scientific literature. Peter Murray Rust (University of Cambridge) notes the potential “to discover new science from heterogeneous data sources” through text mining and data mining the chemical literature, for example. [5] [#N5] Across the disciplines, from the humanities to the sciences, such forms of increasingly data-driven research are rapidly becoming an important form of serious scholarship, or *cyberscholarship*.

A distinguishing characteristic of cyberscholarship is that primary resources (e.g., original data and intermediate results) and even derivative works become an integral component of the scholarly record. These materials are fundamental to the replication of experiments and the validation of results, but are not routinely reported in traditional scholarly venues, making it difficult, if not impossible, for other investigators to replicate results and extend them in new directions. Further, the routine collection and curation of primary data enables research by others that would otherwise be either impossible or unaffordable.

One need not look far for innovative exemplars. Pioneering work in a number of disciplines portends a potential transformation driven by principles underlying cyberscholarship. Likely one of the oldest is reflected in the work of the Interuniversity Consortium for Political and Social Research (ICPSR). As reported on its Web site, "established in 1962, ICPSR is the world's largest archive of digital social science data. [It] acquires, preserves, and distributes original research data and provides training in its analysis. [It] also offers access to publications based on [its] data holdings." [6] [#N6]

In another discipline entirely, the introduction to the US National Virtual Observatory (NVO) Web site states that its "objective is to enable new science by greatly enhancing access to data and computing resources. NVO makes it easy to locate, retrieve, and analyze data from archives and catalogs worldwide." [7] [#N7] The Harvard-Smithsonian Center for Astrophysics, [8] [#N8] whose Web site identifies it as "the world's largest and most diverse center for the study of the Universe," provides online access to a wide variety of datasets and services to scientists, students, educators, and the public. These represent major leaps forward from the days when astronomers' and astrophysicists' only recourse was to compete for observing time at observatories, and scientists in other disciplines and the public were all but shut out from direct access to primary data.

Peter Murray Rust suggests that if the current scientific literature were fully available online in a digitized, semantically accessible form, "huge amounts of undiscovered science" would emerge. [9] [#N9] The impediments are less technological than they are social and economic. Conservative traditions of higher education (e.g., promotion and tenure requirements) coupled with consolidation that inhibits market competition and innovation among scholarly publishers (e.g., John Wiley & Sons' purchase of Blackwell Publishing) render such "literature-data-driven science" [10] [#N10] all but inaccessible.

Urgency: Capturing Content at its Source

Sufficient and growing evidence strongly suggests that scholarly materials available in digital form accelerate the pace of research and increase the number of contributors to that research. But for this to occur within a discipline, a critical mass of materials must be readily accessible online and available for computation. Thus, in order to engage a community of scholars, digital content must be routinely and rigorously collected, curated, managed, and preserved.

While papers published in the print media are becoming increasingly available in digital form, with rather few exceptions the primary data on which these papers are based are not readily accessible, nor typically available. There are, of course, many reasons why this is the case, but high on the list is author motivation as fostered by the policies of merit review in higher education, by the standards of publication established by professional associations, and by the reporting requirements of research sponsors. Institutions of higher education value peer-reviewed publications in top-tier venues, giving scant, and rarely any, credit for non-traditional products such as software or datasets, regardless of their scholarly value. Short-term funding (typically 1–3 years) provided for research projects further undermines the motivation for capturing and preserving primary data. Without the requirement or support beyond the contract period for preservation of materials, and without an institutional or disciplinary commitment for long-term preservation, the value of the overall scholarly record is severely diminished. A change of culture is required to ensure the collection and preservation of digital content. Extending scholarly recognition beyond the traditional formal publication, to include peer-reviewed primary source data, would go a long way in this direction. JISC has already taken some initial and informative steps in this direction, as will be discussed subsequently. Collectively, research sponsors, professional associations, and institutions of higher education should not consider research to be complete until the data has become an appropriate part of the permanent scholarly record.

Digital content must not only be captured and preserved, but it must also be organized in such a manner that it is accessible and usable by computer programs (sometimes called "agents"). While it is not the case that a "one size fits all" standard will meet the needs of every discipline, it is also not the case that every discipline (and every sub-discipline within a discipline) has such unique requirements that no level of agreement is achievable.

It is clear that a systemic solution will be required to address this systemic problem. An appropriate solution must ultimately address the full context of scholarly research and communication, addressing not only the technology, but also the deeply rooted policies and processes of our institutions and professional associations. Current traditions, practices, and policies, which have evolved over a long period of time in which print publications were the gold standard, support

complex and stable interdependencies among higher education, professional associations, publishers, and libraries. These are deeply ingrained in our institutions and the economy, and are directly linked to individual professional advancement (of which tenure is the most dominant factor).

For the entire 20th century (and before), journals provided an accepted, well-known, stable, and successful model for scholarly communication. In the 21st century, it seems that traditional print journals provide too little of the data that scholars are coming to expect, and that the cyberinfrastructure can support. Further, the legal and financial framework underlying traditional scholarly communication is perceived by some to have become an impediment to the advancement of contemporary scholarship. While print media provides a proven venue for reporting scholarly work, cyberinfrastructure is needed for replication and extension of work that increasingly depends on computational analysis of large-scale data resources.

Recent attempts at addressing these issues spawned the development of institutional repositories, systems typically run by institutions of higher education to capture and preserve the scholarly output of its faculty. Often run out of university libraries and viewed as an extension of the libraries' critical role in supporting scholarship and preserving the scholarly record, institutional repositories provide the infrastructure enabling scholars to deposit their digital materials in a shared university resource, thereby assuring broader access and preservation without burdening the individual scholar. But for a number of reasons (e.g., faculty allegiance being stronger to a discipline than to an institution, and depositing in the repository being viewed as additional unproductive labor), institutional repositories have had disappointing success. [11] [#N11] They do not adequately address the full spectrum of change that is necessary and "so far tend to look like 'attics' (and often fairly empty ones)." [12] [#N12]

Needed: A New Form of Infrastructure

Digital content does not belong in the attic, but in the foundation, enabling the advancement of scholarship and informing best practices. The Human Genome Project, [13] [#N13] for example, invested substantially in developing the infrastructure to manage its content, with the very clear objective of tailoring its design to the research practices of the field. But these modes of operation have yet to scale downward to impact other disciplines where small teams and individual investigators are the norm. Across the disciplines, digital content has become (or is rapidly becoming) the newest component of *infrastructure*, which has historically included basic structural foundations such as roads and bridges, the electric grid and, more recently, networks. This new addition to infrastructure ensures that digital content, including published papers, primary data sources, software, and related materials is readily available, accessible, and usable by others. Successful systems will provide appropriate access, be sensitive to institutional and disciplinary cultures of both depositors and users, and address ongoing technical issues, such as long-term stewardship of the resources.

In contrast to a traditional library, where an individual selects information by personally browsing collections, searching catalogs, and examining specific items, in a very large digital collection, computer programs acting on behalf of the individual carry out these functions. Rarely are more than small parts of the collection actually viewed, and then only after automated preliminary screening.

Such screening requires content to be organized for computer analysis. Data formats must support machine processing, and application program interfaces (APIs) must adhere to some level of standardization. Formats designed for human reading (e.g., PDF) serve this need poorly; mark-up languages, such as XML, perform rather well. But access requires more than common formats and standard APIs. The software supporting cyberscholarship is, itself, quite complex, typically benefiting from collaborative, scalable open-source development. Further, complex legal and financial barriers continue to impede access to data.

The identification of strategic high-performance computing requirements in the 1980s resulted in the development of large-scale supercomputing centers sustained by ongoing federal investment. It may well be the case that the emerging and potentially even more compelling requirements for cyberscholarship will be better served by *superdata* centers that leverage economies of scale in both technology and operational efficiency than by a multitude of independent institutional repositories. As has been the case with supercomputing, the development of superdata centers would require significant investment, long-term commitment, and extensive expertise.

Additional leverage can be gained through a common base of value-added services that extends the intellectual reach of all, from novices through experts. While many of these services may be specific to a discipline, there are also generic categories of services that have broad application. In particular, services that support self-organizing knowledge over distributed networks driven by human interaction are envisioned. These will serve to build and support relationships among users, repositories, and communities. [14] [#N14] A process is needed to foster the development of these general-purpose services, rather than requiring individual, customized tools for each application.

Curation and preservation of digital materials pose more challenges. Programs such as the Library of Congress' National Digital Information Infrastructure and Preservation Program [NDIIPP] [15] [#N15] are tackling this problem in the long term, but they may have little to preserve unless other agencies like the NSF and NIH take more aggressive steps to assure that research data generated by short-term grants is not discarded at the end of the grant. This will require more than an unfunded mandate, however. Neither researchers nor their institutions typically have the resources, the incentives, or the responsibility to preserve the primary data and intermediary products of their work.

Some disciplines are thinking beyond simply *saving* data. They are working toward peer-review systems analogous to those used for reviewing scholarly papers, in which data can be judged on qualities of coherence, design, consistency, reliability of access, and related criteria. Under JISC sponsorship in the UK, for example, a new kind of scholarly resource called a *data journal* is being established, where practitioners submit data sets for peer review and dissemination. [16] [#N16]

Future: Framing a 7-Year Road Map

The workshop concluded that the current laissez-faire approach would not result (at least in the near term) in a broad-based content infrastructure supporting cyberscholarship. Alternatively, a seven-year program of research, development, and implementation is proposed that would enable progress toward the following goal:

> Ensure that all publicly-funded research products and primary resources will be readily available, accessible, and usable via common infrastructure and tools through space, time, and across disciplines, stages of research, and modes of human expression.

This goal is not tied to a specific view of scholarship, but is intended to nurture both cyberscholarship and conventional forms of research. The aim is to go far beyond a system that merely replicates the traditional methods used for physical media in the digital domain.

A target date of 2015 is proposed to reach this goal, staged through a phased process in which a series of prototypes informs development of a stable infrastructure. The goal is aggressive, but achievable when coordinated with other initiatives in the US, the UK, and elsewhere. A three-phase program is proposed, including a three-year research prototype phase that explores several competing alternatives and a one-year architecture specification phase that integrates the best ideas from the prototypes, followed by a three-year research and implementation phase in which content infrastructure is deployed while research on value-added services continues.

Each phase includes four classes, or "stages," of activity: *administration*, *behaviors*, *research,* and *infrastructure*. The *administration* stage is conducted by the responsible funding agencies (e.g., NSF, IMLS, NEH, NIH, Library of Congress, and DoD). Administrative agents working in collaboration both nationally and internationally establish program objectives and budgets, and solicit and support cyberscholarship research proposals.

The second stage addresses *behaviors*. Development of cyberscholarship infrastructure requires organizational adaptation and attention to personal motivations and incentives. Too often ignored to the peril of real progress, this area is of such fundamental significance to cyberscholarship that it is called out here as a separate stage to emphasize its importance.

The third stage, *research*, focuses specifically on developing the tools and services that comprise the technological underpinning of cyberscholarship. This includes tools leading to the automatic ingest, identification, indexing, management, and analysis of scholarly communications across language barriers and among disciplines. To be widely adopted, these tools need to be seamlessly integrated into an infrastructure that provides transparent services to the scholarly community.

A series of instrumented, exploratory pilot projects kicks off the fourth stage, *infrastructure*. The objective of this stage is to deploy the first generation of cyberscholarship infrastructure derived from advanced research and experimentation in tools, services, institutional behaviors, and personal incentives.

Conclusions

A highly multidisciplinary, international group of scholars, administrators, and government officials were invited to a workshop in Phoenix, Arizona co-sponsored by NSF and JISC to examine the seeming contradiction between the rapid growth of digital materials and the slower adoption of digital repositories throughout the scholarly community. The participants considered research practice, which is increasingly enabled by network infrastructure, and the shortcomings

of traditional scholarly publication to communicate the results of that research effectively. They observed that scholarly communication, as well as research practice, is undergoing significant transformation as a result of the growth of digital technologies, but that the collective infrastructure supporting that transformation has not kept pace.

The workshop concluded that:

- The widespread availability of digital content creates opportunities for new forms of research and scholarship that are qualitatively different from traditional ways of using academic publications and research data (this is called cyberscholarship).
- Digital content must be captured, managed, and preserved in ways that are significantly different from conventional methods to support cyberscholarship
- Development of the infrastructure requires coordination at a national and international level.
- Development of the content infrastructure requires a blend of interdisciplinary research and development that engages scientists, technologists, and humanities scholars.
- A seven-year timetable is recommended for research and development to support implementation of the content infrastructure.

The time is right for a focused, international effort to experiment, explore, and finally build the infrastructure for cyberscholarship.

Acknowledgements

This work was funded in part by the National Science Foundation through grant number IIS-0645988 and the UK Joint Information Systems Committee.

Notes

1. R. Larsen and William Y. Arms, “The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship,” report of a workshop held in Phoenix, Arizona, April 17–19, 2007, available at www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf [http://www.sis.pitt.edu/~repwkshop/NSF-JISC-report.pdf] [#N1-ptr1]
2. “NSF’S Cyberinfrastructure Vision for 21st Century Discovery,” available at www.nsf.gov/attachments/102806/public/NSFCyberinfrastructureVisionDraft-4.0.pdf [http://www.nsf.gov/attachments/102806/public/NSFCyberinfrastructureVisionDraft-4.0.pdf] [#N2-ptr1]
3. William Y. Arms, “Cyberscholarship: High Performance Computing meets Digital Libraries,” in this issue of the *Journal of Electronic Publishing*. [#N3-ptr1]
4. Electronic Cultural Atlas Initiative, http://www.ecai.org/ [http://www.ecai.org/] [#N4-ptr1]
5. See http://www.ch.cam.ac.uk/staff/pm.html [http://www.ch.cam.ac.uk/staff/pm.html] [#N5-ptr1]
6. Interuniversity Consortium for Political and Social Research, http://www.icpsr.umich.edu/ [http://www.icpsr.umich.edu/] [#N6-ptr1]
7. National Virtual Observatory, http://www.us-vo.org/ [http://www.us-vo.org/] [#N7-ptr1]
8. The Harvard-Smithsonian Center for Astrophysics, http://cfa-www.harvard.edu/ [http://cfa-www.harvard.edu/] [#N8-ptr1]
9. Peter Murray Rust, “Data-Driven Science — A Scientist’s View,” available at http://www.sis.pitt.edu/~repwkshop/papers/murray.html [http://www.sis.pitt.edu/~repwkshop/papers/murray.html] [#N9-ptr1]
10. Ibid. [#N10-ptr1]
11. Philip M. Davis and Matthew J. L.Connolly, “Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University’s Installation of DSpace,” *D-Lib Magazine* 13, no. 3/4 (March/April 2007), available at http://www.dlib.org/dlib/march07/davis/03davis.html [http://www.dlib.org/dlib/march07/davis/03davis.html]

[#N11-ptr1]

12. Laura Brown, Rebecca Griffiths, and Matthew Rascoff, “University Publishing in a Digital Age,” July 26, 2007, available at http://www.ithaka.org/strategic-services/Ithaka%20University%20Publishing%20Report.pdf [http://www.ithaka.org/strategic-services/Ithaka%20University%20Publishing%20Report.pdf] [#N12-ptr1]

13. Human Genome Project, http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml [http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml] [#N13-ptr1]

14. Luce, Rick, “EDatabase Lessons for an EData World,” position paper submitted to the NSF/JISC Repositories Workshop, April 17–19, 2007, available at http://www.sis.pitt.edu/~repwkshop/papers/luce.html [http://www.sis.pitt.edu/~repwkshop/papers/luce.html] [#N14-ptr1]

15. National Digital Information Infrastructure and Preservation Program, http://www.digitalpreservation.gov/ [http://www.digitalpreservation.gov/] [#N15-ptr1]

16. Waters, Donald J., “Doing Much More Than We Have So Far Attempted,” *EDUCAUSE Review* 42, no. 5 (September/October 2007), available at http://www.educause.edu/apps/er/erm07/erm0756.asp [http://www.educause.edu/apps/er/erm07/erm0756.asp] [#N16-ptr1]

Product of Michigan Publishing (http://www.publishing.umich.edu), University of Michigan Library (http://www.lib.umich.edu/) • jep-info@umich.edu (mailto:jep-info@umich.edu?subject=Journal%20of%20Electronic%20Publishing) • ISSN 1080-2711