Advancing Research and Practice in Digital Curation and Publishing

Summary Report and Recommendations of the Workshop on Next Steps in Research, Education and Practice

University College London

June 26, 2010

Introduction

The "Next Steps in Research, Education and Practice in Digital Curation and Publishing" workshop was held in London on June 26, 2010, following the **Fourth Bloomsbury Conference on E-Publishing and E-Publications: 24 and 25 June 2010: Valued Resources: Roles and Responsibilities of Digital Curators and Publishers** (www.ucl.ac.uk/infostudies/e-publishing/e-publishing2010). Both events were cosponsored and organized by the University College London (UCL) Department of Information Studies and the U.S. Institute of Museum and Library Services (IMLS).

Speakers at the conference on June 24–25 discussed the changing roles of—

- publishers, who are faced with the pressure to take responsibility for "supplementary" material, including an increasing number of datasets of growing complexity;
- librarians, as they increasingly manage digital assets and develop data management services; and,
- researchers, who increasingly create and depend on digital data.

The role of digital repositories, which support all of these activities, provided an organizing theme to explore common ground. Understanding how to support digital curation is a central challenge and opportunity for publishers, libraries, data centers, museums, archives, and other data-centric organizations (see "Note on Terminology" below, for a working definition of digital curation). Although publishers, researchers, librarians and other cultural heritage and information professionals have had long-standing interrelationships, the opportunities for working collaboratively to develop new publications and data curation business models for the digital environment have not yet been sufficiently explored.

The postconference workshop built on ideas introduced in the conference. Nineteen invited participants, representing institutions in Germany, the Netherlands, the United Kingdom, and the United States, discussed challenges and opportunities and made recommendations for future research and action relating to digital curation and e-publishing. (See the Participant List at the end of this report.) Participants particularly considered activities that could be undertaken in the near future in order to promote communities of interest, research agendas, collaborations, and potential models. They also recognized the potential advantages of collaboration across professional, disciplinary, and national boundaries. (See the "Recommendations.")

Note on Terminology

An important driver for building collaborative networks across professional disciplines is the need for a common understanding of terms. Participants noted the need to develop a common vocabulary for terms such as "curation." Museum curators, for example, emphasize interpretation of objects in addition to management and care of collections. The U.K.'s Digital Curation Centre states that digital curation "involves maintaining, preserving and adding value to digital research data throughout its lifecycle."[1] Curation in this context is generally understood to

---

[1] www.dcc.ac.uk/digital-curation/what-digital-curation.

include (1) continuity of access, including archiving and preservation; (2) ease of access, including discoverability; and (3) added value, or context, that makes the information more meaningful. There are also differences of opinion and usage regarding the terms "digital curation" and "data curation."

Discussion Summary

The workshop began with reflections on the preceding two days of the Bloomsbury Conference. Early discussion centered on the concept of value and how it may be defined and measured. Although the conference had introduced the term "value," it did not explicitly identify the context—whether economic, social, or some other kind of value—or address the implied question, "value to whom?"

In an economic context, "value" may be understood to mean the flow of benefit over and above incurred costs. In the current economic climate, it will be important for institutions that depend at least in part on public funds to demonstrate value, whether direct or indirect, to the general public. The recent report of the Blue Ribbon Task Force on Sustainable Preservation and Access makes a case for the public value of digital assets at a high policy level.[2] One of the report's major observations is that the value proposition for public investment in digital preservation can best be made on the benefits of managing digital resources for reuse by the current generation of researchers, not just on some projection of distant future value. It will be important to make this case, with concrete evidence based on specific examples, to a wide range of stakeholders who stand to benefit from well-managed and accessible data, including government agencies, universities, researchers, businesses, and ultimately the general public. Other stakeholders, including publishers, data scientists, librarians, archivists, and funders, could work together to provide enhanced data management services. Given that virtually every aspect of human endeavor is now dependent on digital information, awareness should be raised that maintaining that information at large scale is a critical global challenge.

Efforts to build support to meet this challenge must be rooted in clear statements of what criteria of value are applicable to the various players and contexts. This will include broad concepts such as economic and social value, together with more specific contexts such as value linked to broader institutional and government agendas. Research is necessary to show the extent to which new value statements differ from those applied to traditional service propositions, and why.

Institutions need to think about, and sometimes rethink, their value propositions. Should they try to acquire more funding to take on additional roles? Or should they free up money by spending less on traditional roles in order to take on new activities? How can they weigh the costs and benefits of new vs. traditional services? Are there ways to reduce costs to operate more economically? The idea of small-scale pilot projects to test potential models generated a great deal of enthusiasm. Successful exemplars that could be built and shown to work with a relatively

---

[2] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustainable economics for a digital planet: Ensuring long-term access to digital information,* 2010. Available at http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

small investment of funds, and that can be scaled up or adapted by other institutions, would benefit the greater community of knowledge creators, users, and preservers. These examples would also demonstrate value to the public, which would benefit from having online access to information, as well as from advances in science, creativity, and knowledge that result from use of the information.

Some discussion concerned the need in scholarly communication to demonstrate impact beyond journal citations. How can organizations make a case for maintaining information that does not make its way into a publication, such as datasets used to produce conclusions and supplementary data cited in a publication? One project discussed at the conference, Dryad, is preserving such data.[3] Dryad serves as a repository for tables, spreadsheets, flat files, and other kinds of data underlying scientific publications that do not have a home elsewhere. Dryad allows investigators to validate published findings, explore new analysis methodologies, repurpose the data for research questions unanticipated by the original authors, and perform synthetic studies such as formal meta-analyses.

There is a compelling need for information about costs throughout the life cycle of data. The Keeping Research Data Safe project, funded by the U.K. Joint Information Systems Committee, is currently conducting research to collect such data.[4] The increasing costs of research and publication, as well as the exponential growth of information that exists only in digital form, suggest that more work should be done to explore the relative costs and benefits of conducting research and disseminating it in different ways, as well as the need to explore optimal economies of scale for storing digital data and making the data accessible for reuse. Repositories incur costs up front, yet it takes time for them to scale up to a critical mass of content and build an established user base. Data repositories face a particular challenge in demonstrating impact, as the practice of data citation is still in its infancy (unlike journal articles) and may take years to become established. More research on how to demonstrate emerging value or impact from a range of recently established data repositories would be useful. A case must also be made to university administrators, government and corporate officials, research funders, and other stakeholders of the value of preserving both data and publications, and for their support of data repositories at departmental, institutional, or multi-institutional/disciplinary levels. Can it be shown, for example, that data stewardship will help a university lead the field in a particular area of research in which it has already invested by building outstanding research departments, centers, or laboratories? What case might be made for multi-institutional repositories? More attention needs to be paid to the potential value of what publishers call "supplementary material" or "supplementary data," which are often seen as somehow inferior to publications. If impact is measured only in terms of journal citations, it will be difficult to demonstrate or measure the value of the data and other resources that inform and enhance publications.

---

[3] N. Beagrie, "Dryad: Archiving and Preservation of Data" (presentation at Fourth Bloomsbury Conference on E-Publishing and E-Publications, London, U.K., June 24–25, 2010). Presentation slides available at www.ucl.ac.uk/infostudies/e-publishing/e-publishing2010. See also Dryad home page: http://datadryad.org.
[4] www.beagrie.com/jisc.php.

Interpretive publications and the data on which they are based should point to each other, thereby increasing the value of each and enhancing the potential value of data for validation and reuse. An argument was made at the conference for treating data as independent content to be accessed directly, especially when the data are dynamic and not fixed. Although some publishers, such as the Organization for Economic Cooperation and Development, may publish data collections, in most cases publishers would prefer to point to data held elsewhere if they can be confident of their persistence.

Supplementary data are often thought of as being only text, but in fact they include a wide range of other formats that deserve more attention, including multimedia, spreadsheets and graphics, and software and parameter files for models. For example, in disciplines such as chemistry, it is important to provide dynamic access to searchable data in their original form rather than in text form. Scholars often present evidence in seminars and conferences, including simulations or other illustrative material that provides important insights into their work, which does not appear in publications and is not preserved anywhere. Such information may substantially enhance a subsequent publication and therefore have more value than other forms of supplementary material, yet it is challenging to preserve and no one is collecting and preserving it.

Participants discussed the organizational structures needed to support cost-effective and efficient repositories. Repositories have been developed at different scales, ranging from those serving individual institutions to those that operate on a regional or national level, and subject-based repositories that focus on specific disciplines and may be of greater interest to scholars than institutional repositories. Eefke Smit's conference presentation reported on a recent survey of researchers in Europe and the United States that showed particularly positive researcher attitudes toward depositing in either the digital archive of their organization or the digital archive (data center) of their discipline.[5] In addition, the U.S. National Science Foundation (NSF) recently announced that, beginning in 2011, all grant applicants will be required to submit data management plans with their applications. For those disciplines currently not served by a disciplinary repository (estimated as approximately 80 percent of NSF grants), researchers will need to think seriously about whether and how they will provide continued access to their data after the conclusion of their research. Participants noted that a study of the data management plans submitted with grant applications would provide important insights into what arrangements researchers are making for preservation of their data.

Institutional repositories may be too small to achieve the necessary economies of scale if they continue to operate independently, or without aggregation to achieve virtual critical mass. Workshop participants felt that the potential of cross-institutional approaches should be given more serious consideration. One promising model in the United States is the Committee on

---

[5] E. Smit, "Entering the Data Era: Digital Curation of Data-Intensive Science…and the Role Publishers Can Play: The STM View on Publishing Datasets" (presentation at Fourth Bloomsbury Conference on E-Publishing and E-Publications, London, U.K., June 25, 2010). Presentation slides available at www.ucl.ac.uk/infostudies/e-publishing/e-publishing2010.

Institutional Cooperation (CIC), a consortium of Big Ten universities plus the University of Chicago.[6]

In considering what to keep, it would be useful to know the cost of replacing data. It could be instructive to look at how insurers determine valuation as a potential model. But what value should we put on something that is irreplaceable? More work is needed on what economic value is, and what other kinds of value should be factored into the assessment. Some of the work that has attempted to quantify value (return-on-investment studies, for example) is dubious when applied to the public sector. Such studies try to determine how much money is generated indirectly by an investment of funds, and may try to quantify benefits that are difficult or impossible to measure in monetary terms; also, such studies do not consider how much return might have been generated by investing the funds some other way. Not all value can be quantified in monetary terms, so other measures of value should be explored.

Some emerging data sources might provide more insight into costs of the scholarly communication process. Now that most publishers have converted to automated manuscript management systems, it should be possible to look at data such as time lapse from submission to acceptance or rejection of a journal article, and the relative costs of each stage in the process. It might be possible to partner with scholarly societies and commercial publishers to get anonymized data for analysis, which may provide better insight into the operational patterns of the peer review process, including historically concealed costs. Greater transparency could help to encourage more experimentation with new models.

The field of library and information science could undertake research on many of the questions raised during discussion, but some noted that fundamental understanding of data as information is lacking. It is important to work with domain researchers to understand how data are generated and used as evidence, as well as what is done with data afterwards. Useful tools such as the Data Curation Profile, developed by Purdue University Libraries' Distributed Data Curation Center and the University of Illinois Urbana Champaign Graduate School of Library and Information Science, are becoming available to help structure such conversations.[7] At the same time, publishers need to understand the processes of review and dissemination of data sources. An information research network could help promote awareness and synthesis of relevant research and tools.

Partnerships among libraries, institutional repositories, and data stores within departments (a distributed model currently being explored by DSpace@Cambridge[8]) could help to achieve efficiencies and economies of scale in repository management and enhance collaboration. Data centers and research projects have also recruited digital curation graduates of schools of library and information science.

Workshop participants returned several times to the issues raised by "small science" and "small humanities" (virtually all humanities research). It was noted that small science is big: In 2007,

---

[6] CIC home page: www.cic.net.
[7] www.datacurationprofiles.org.
[8] www.dspace.cam.ac.uk.

the National Science Foundation awarded 12,000 grants totaling $2.8 billion; 254 grantees received 20 percent of the total, while more than 9,000 received grants of under $300,000.[9] Many of these smaller projects tend to be in disciplines not well served by national or international disciplinary repositories. What role can publishers and libraries play in disseminating and preserving the data generated by those disciplines? Furthermore, even discipline-based data centers could benefit from participation in a larger data curation environment in order to avoid building silos that inhibit cross-disciplinary reuse of data.

In the cultural heritage sector, aggregations of content from many institutions of various sizes can be very significant. Europeana, the digital library funded by the European Union, is a case in point.[10] To date, it contains representations of 12 million cultural artifacts, with a semantic search engine prototype available through the ThoughtLab section of the portal, and other enhancements to be provided in the near future.[11] In the United States, the University of Illinois, with funding from IMLS, began in 2002 to build a metadata repository of content from U.S. libraries, archives, and museums. This repository has become the largest curated online collection of resources on U.S. history, with nearly 900 collections and more than 1 million item-level records.[12]

These types of aggregations provide great potential for experimental collaboration with publishers. Publishers could add value by developing selected publicly available content for publication, helping to raise awareness and increase use of the online content. For example, reproduction of publication-quality images on paper is quite expensive, limiting the number of images that can be reproduced in print, while web sites can host many high-quality images and large quantities of related resources. These considerations could lead to the development of new business models for hybrid paper and web publications.

A good example of publisher-library collaboration in the humanities was presented at the conference by Patrick Alexander and Mike Furlough of Penn State University.[13] Penn State has created an Office of Digital Scholarly Publication that is jointly supported by the university press and the libraries. One current project involves the publication of a bibliography of German-language broadsides published in North America between 1730 and 1830, supported by a subvention from the Deutsche Forschungsgesellschaft (DFG, or German Research Foundation). Penn State Press plans to publish the work with about 20 illustrations, and the Penn State Libraries will provide open access to the electronic edition, which will include images of many more, possibly all, of the facsimiles. The presenters argued that bibliography is a

[9] P. B. Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends* 57 (2008), 280–99.

[10] Europeana Digital Library home page: www.europeana.eu/portal.

[11] www.europeana.eu/portal/thoughtlab.html. In addition, Europeana has devised an entirely RDF-graph–based data model, to be operational by May 2011, that will enable technical integration of Europeana into the emerging paradigm of Linked Open Data.

[12] Opening History home page: http://imlsdcc.grainger.uiuc.edu/history.

[13] P. Alexander, and M. Furlough, "Humanities Publishing and Data Curation: Eternal Life and Eternal Damnation" (presentation at Fourth Bloomsbury Conference on E-Publishing and E-Publications, London, U.K., June 24, 2010). Presentation slides available at www.ucl.ac.uk/infostudies/e-publishing/e-publishing2010.

primary activity in humanities scholarship and that the underlying database is both an object for curation and a source for further scholarly research.

Participants recognized that considerable attention needs to be paid to users, but some noted that too often researchers who conduct user studies are forgetting their own role as analysts, not just reporters. Information science researchers need not just to consider what people say they want, but also to investigate what users would want if they knew what they *could* have and identify longer-term and larger-scale benefits to a field of study that users may not be in a position to recognize. The relationship between the digital and physical worlds is an exciting area for research. It is of particular interest to scholars in the digital humanities. How do people make connections between the digital and the physical? Will people want to use semantic functionality as a bridge between them? The deployment of "intelligent agents" (as they are known in the artificial intelligence community) to create services around digital content could provide new opportunities to demonstrate impact. The knowledge of archivists, librarians, museum professionals, scholars, and publishers can be brought to bear to address these issues.

Workshop participants also discussed the relationship among research, education, and practice. The digital environment has helped to break down traditional boundaries, such as those separating libraries, museums, and publishers. Today's students find it easier to move across information environments, and consideration should be given to what this means for the future of organizations as well as for education. In library and information science education, the curriculum tends to be stratified toward either the technology-oriented or people-oriented ends of the spectrum. Perhaps the conclusion to be drawn is that both perspectives are important, and a better understanding is needed of how they interact (again, connections between the digital and the physical). Research and education should be viewed as intertwined; students who have done successful research are now excelling in the workplace and will become leaders for change. UCL, which is unusual among library and information science schools in having a Centre for Publishing within its Department of Information Studies, is playing an important educational role in making connections among publishers, librarians, and digital curators. In the United States, IMLS has promoted and funded the development of digital curation programs in graduate schools of library and information science since 2006. This funding has supported the development of robust programs (including core curricula, specialized elective courses, and required internships in established digital repositories) in a number of institutions, including the University of Illinois Urbana Champaign, the University of North Carolina at Chapel Hill, and the University of Tennessee.

Conclusion

Participants recognized that this was only a beginning conversation about a potentially fruitful new relationship among archivists, librarians, museum professionals, scholars, and publishers, with the advantage of building upon earlier discussions between data librarians and data archivists.[14] The recommendations that follow provide a general roadmap for how these

---

[14] www.iassistdata.org.

relationships might be developed, with specific actions suggested to promote ongoing discussion and collaboration.

Recommendations

Recommendation 1: **Foster communication and collaboration among publishers, data scientists, librarians, archivists, and other practitioners who manage data, as well as funders, researchers, and educators to advance research and practice in digital curation and publishing.**

- *Leverage existing networks to promote ongoing discussion by key organizations such as the Association of American University Presses, the Society for Scholarly Publishing, and graduate schools of Information (iSchools).*

- *Explore venues and opportunities in which to engage stakeholders and facilitate collaboration and awareness of issues related to digital curation and publishing, including the potential of the International Digital Curation Conference to play a role in this effort.*[15]

- *Identify and involve stakeholders, including publishers, funders, scholars, and scientists, to develop pilot programs and undertake research in digital curation and publishing.*

- *Investigate the interest of funding bodies, including IMLS in the United States, Joint Information Systems Committee in the United Kingdom, SURF in the Netherlands, the DFG, and the European Commission, in grant support of collaborative activities, including additional workshops. Future workshops could target a range of representatives from commercial and nonprofit publishers, including university presses; scholarly societies; libraries; museums; data centers; and library and information science researchers and educators.*

- *Explore opportunities for promoting collaboration and exchange of information between educational programs in digital curation and publishing. Consortia such as the iSchools group and the International Digital Curation Education and Action (IDEA) Working Group could be useful for this purpose.*[16]

Recommendation 2: **Develop research agendas around digital curation and publishing.**

There need not be a single agenda, given the size and diversity of the professions involved, but various stakeholder groups should be engaged to consider what research would be useful. Suggested topics include the following:

---

[15] www.dcc.ac.uk/events/conferences/6th-international-digital-curation-conference. A postconference workshop following the 2010 conference will include consideration of this report.
[16] For more on IDEA, see http://ideaworkgroup.org.

- *Investigate and make recommendations to establish common understanding of terminology within the new research area of web science, such as "curation," including formal conceptual abstractions and fundamental concepts.*

- *Undertake a gap analysis of tools for performing digital curation; several participants specified the need for lighter tools.*

- *Further investigate costs and impacts (not necessarily in the same study), with attention to the development of better, nonmonetary measures of value and impact.*

- *Conduct research on the concept of public value (e.g., economic, social, research) and how it may be demonstrated both within the broader context of institutional agendas and for different kinds of content and uses.*

Recommendation 3: **Develop a virtual research information network to monitor the changing environment in which researchers access and share data and documents, and in which libraries and publishers must now engage.**

The resource should promote current awareness of research and practice in digital curation and publishing, as well as synthesis of past and current research and practice.

- *Explore the potential to expand an existing web resource (such as the Digital Curation Centre or the Digital Curation Exchange) to address this recommendation.*[17]

Recommendation 4: **Seed some demonstration projects to aggregate and publish small science and small humanities content and tools in hybrid web/paper formats.**

- *The following characteristics of successful projects were proposed:*
  - Projects' intellectual focus should be on a topic that has demonstrable interest to multiple user communities, and proposers should "think like publishers" in terms of selecting themes that will be of interest to "the market."
  - Projects should involve multiple partners rather than being focused at one institution, and should bring together participants with complementary content and skill sets. Proposers should be encouraged to think outside the boundaries of their own institutions to identify and serve wider communities of interest. International projects should be encouraged.
  - Projects should result in "exemplary containers" that can be repurposed for other projects. Proposers should focus on lightweight delivery platforms and workflows that are not too tailored for one particular subject area but can be used and adapted by other projects. The demand for the HUBzero "platforms for scientific collaboration" created at Purdue University shows that there is strong interest in off-the-shelf solutions, but these need not be as complex or expensive as

---

[17] For more on the Digital Curation Exchange, see www.digitalcurationexchange.org.

HUBzero, which has powerful simulation and computing capabilities that many projects may not need.[18]

- o Proposers should be careful to balance the desires of the intellectual creators of the subject collection with the needs of the larger population of intended users.

- *In addition, these recommendations to funders were proposed:*
    - o The opportunity and need to support citizen science and serve the needs of unaffiliated scholars when aggregating, publishing, and curating subject-based collections should be recognized; museums, libraries, and archives could play a role in leading such projects and engaging these contributors.
    - o Attention should be paid to a longer-term goal of interoperability between many small projects, and thus there should be a focus on using common standards. Promoting collaboration between aggregations is important.

Recommendation 5: **Investigate and propose strategies for curating, identifying, and linking data to publications.**

- *Investigate issues involved in curating supplemental data. These include basic problems of identifiers, integrity, format, metadata, and access, as well as problems for challenging formats such as multimedia and multipart works.*

- *Promote cooperation between publishers and libraries to cross-link supplemental data to repositories using standard formats, identifiers and protocols, and supporting metadata.*

- *Promote cooperation between publishers and libraries to help expose content to larger audiences and drive traffic to data through such means as (a) providing richer and more granular linking; (b) providing XML versions as well as PDF versions of data to enable use as data; (c) pointing to datasets; (d) using the Linked Data field;[19] (e) finding new sources of data; and (f) establishing the role of editor in creation of datasets.*

Workshop Participants

Patrick Alexander, Director, Penn State University Press, and Co-director, Penn State Office of Digital Scholarly Publishing

Chris Batt, Chris Batt Consulting

Neil Beagrie, Director of Consultancy, Charles Beagrie Ltd.

Scott Brandt, Associate Dean for Research and Professor of Library Science, Purdue University Libraries

Peter Burnhill, Director, EDINA National Data Center, University of Edinburgh

---

[18] For more on HUBzero, see https://hubzero.org.
[19] http://esw.w3.org/LinkedData.

Mike Furlough, Assistant Dean of Scholarly Communications, Penn State University Libraries, and Co-Director, Penn State Office of Digital Scholarly Publishing

Tula Giannini, Dean, Pratt Institute School of Information and Library Science

Stefan Gradmann, Professor, Berlin School of Library and Information Science (B-SLIS), Humboldt University (Berlin)

Carolyn Hank, Doctoral candidate, School of Information and Library Science, University of North Carolina at Chapel Hill (*recorder*)

Cliff Lynch, Director, Coalition for Networked Information, United States

Wilma Mossink, Legal Advisor, SURFfoundation, the Netherlands

David Nicholas, Director, Department of Information Studies, University College London (UCL), and Director, UCL Centre for Publishing and CIBER research group

Carole Palmer, Professor and Director of the Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois Urbana Champaign

Joyce Ray, Associate Deputy Director for Library Services, Institute of Museum and Library Services, United States

Allen Renear, Associate Dean for Research and Associate Professor, Graduate School of Library and Information Science, University of Illinois Urbana Champaign

Carol Tenopir, Professor of Information Sciences and Director of Research, College of Communication and Information, University of Tennessee, and Director of Center for Information and Communication Studies

Claire Warwick, Reader, Digital Humanities, Department of Information Studies, UCL, and Director, UCL Centre for Digital Humanities

Anthony Watkinson, Senior Lecturer, Centre for Publishing, Department of Information Studies, UCL

Charles Watkinson, Director, Purdue University Press