Charleston Conference 2009 Proceedings

Session Title: "Digital Curation and E-Publishing: Libraries Make the Connection"

Authors:

Sayeed Choudhury, Associate Dean for Library Digital Programs and Hodson Director of the Digital Research and Curation Center, Johns Hopkins University Sayeed@jhu.edu

Mike Furlough, Assistant Dean for Scholarly Communications and Co-Director, Office of Digital Scholarly Publishing, Penn State University Libraries mfurlough@psu.edu

Joyce Ray, Associate Deputy Director for Library Services, Institute of Museum and Library Services jray@imls.gov

Abstract:

This paper addresses issues in digital curation, which is the management of digital assets throughout their life cycle for maximum interoperability, discovery, preservation and re-use. The experiences of two institutions—Johns Hopkins University's Sheridan Libraries and Penn State University Libraries—are presented as examples of how libraries can collaborate successfully with publishers to provide preservation and other back-end services to support scholarly publishing. One example (Johns Hopkins) relates to scientific data, while the other (Penn State) relates to the humanities. The paper also discusses the potential of digital curation to expand and enhance library services broadly and describes the investment the Institute of Museum and Library Services has made in digital curation research, education, and practice.

Text:

Introduction

Digital curation is truly a 21st century term. It emerged in the last decade as a concept that provides a comprehensive view of the creation and management of digital data, originally scientific data, that is now being ingested on a massive scale by digital repositories. This concept recognizes that, in order for digital assets to be maintained over a long period of time, they must not only be preserved but must also be created according to high quality standards to ensure interoperability with other data and to enable re-purposing and discovery by future users beyond the original creators and users. The concept became concrete when the Digital Curation Centre (DCC) was established in the UK at the University of Edinburgh, with a number of UK partner institutions. The DCC is funded by the UK's Research Councils E-Science Programme and by the Joint Information Systems Committee. The concepts and practices of digital curation have now expanded beyond the scientific community to encompass all areas of scholarly activity.

What is digital curation? The DCC defines it as "maintaining and adding value to a trusted body of digital information for current and future use; specifically the active management and appraisal of data over

the life-cycle of scholarly and scientific materials." [http://www.dcc.ac.uk/about/--see fig. 1] The term "life-cycle" indicates the importance of archival principles to the management of digital data. [Note that while the terms "digital curation" and "data curation" are frequently used interchangably, a distinction may be made between them. "Data" often refers to scientific data, which may include data not in digital form, while "digital" refers more broadly to all digital content; thus, many people consider "digital curation" to be the broader term, through both terms are often used to mean the same thing.]

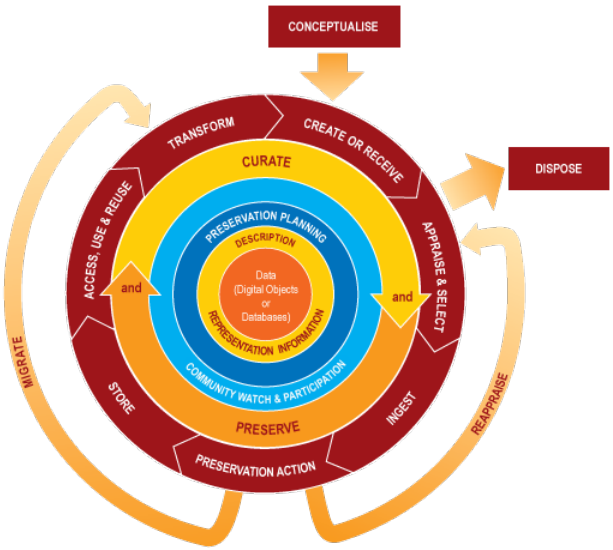

Fig.1 ***credit: Digital Curation Centre***

Running parallel to the emergence of digital curation as a professional field—whether a new field or a subset that cuts across several existing fields remains a matter of debate—is the emergence of e-publishing, or rather, the transformation of traditional paper-based publishing to the electronic environment. The publication process for scholarly materials is now almost exclusively electronic, from acceptance of manuscripts through review and editing, whether or not the final product appears in print.

So we have, on the one hand, a community, or a subset of several communities, that has been working on the "back end" of digital production from the generation of raw data to the construction of an organized product that can be accessed, and, on the other hand, another community—publishers—who work on the "front end" of scholarly communications, from manuscripts to publication. This paper discusses some possibilities for bringing these communities together and demonstrates the role that libraries are playing in making this connection.

Libraries, Publishers and the Sciences

The Sheridan Libraries at Johns Hopkins University (JHU) have embarked upon a major data curation program that has most recently resulted in one of the two awards through the National Science

Foundation's (NSF) DataNet solicitation. This program reflects years of collaboration between the Sheridan Libraries and the Department of Physics and Astronomy at JHU. One of the most important realizations from initial engagements with astronomers is the concept of levels of data that may be applicable to other sciences and perhaps even to the humanities. Astronomers, much like other scientists, recognize different levels of data beginning with the "raw" data generated directly by telescopes ("level zero" data). The data are then processed and calibrated into more refined versions that are increasingly accessible by domain specialists (levels one and two). Eventually the data are processed sufficiently that they can be "published" as a data release or level-three data that can be shared via standard mechanisms such as websites (e.g., SkyServer at http://www.sdss.org). These community-based data releases serve as the foundation for research by professional astronomers and exploration by amateur astronomers or citizen scientists. The analyses of the level-three data or data releases eventually results in yet one more level of refined data which are typically cited directly within publications. It is these level-four data that represented the initial target of JHU's data curation efforts.

The astronomy community has previously attempted to capture and preserve these level-four data by directly appealing to astronomers to deposit them into the Astronomy Digital Image Library (ADIL). While ADIL represented an earnest and useful attempt at data preservation, it did not generate a great deal of engagement. The main lesson from ADIL's experience is that it is critical to embed data curation activities as part of existing workflows or processes. Through conversations with the American Astronomical Society (AAS), the Sheridan Libraries designed a system that incorporate data capture and preservation as part of the publishing process. With funding from the Institute of Museum and Library Services (IMLS) and Microsoft Research, JHU is developing a process and prototype system to demonstrate this data curation approach.

The development of this data curation prototype system has demonstrated the importance of bringing together library, publisher and professional society perspectives. While the principle requirement relates to the library's goal of preserving and providing access to data, the overall design of the system accounts for connections between data and publications and workflow associated with publishing systems. By bringing the library, publisher and professional society perspectives together, each community is able to accomplish its primary objective without compromising the other community's existing process, practices or systems. Undoubtedly, some give and take is required from each group, but the overall data curation goal provided an overarching framework for common goals. In this sense, the library acted as the "hub" bringing together the various players and provided the technical "glue" to bring together the various components. The technical architecture for this system uses community protocols (e.g., SWORD) to facilitate adoption of the specific tools and services and reduce the burden on publishers and scientists who would use the system.

One of the most important protocols for this data curation prototype is the Open Archives Initiative Object Reuse and Exchange (OAI-ORE). OAI-ORE is a protocol for describing compound objects that include different types of distributed objects. The Resource Map (ReM) is a description of aggregations of objects that are connected structurally and semantically. JHU has used OAI-ORE ReMs to express the data models underlying the data curation prototype. In addition to the expression of connections between articles and data, OAI-ORE also provides a mechanism to understand the provenance of data—

that is, who or what manipulated the data and for what purpose. The importance of provenance is important for the validation of scientific data, but it is also a concept well understood in the humanities community.

While the technical components were important for the development of this data curation prototype system, it is especially important to note the role of a particular individual at AAS who acted as the human "interface" between the various players. This individual could easily be classified as a "data scientist" – an individual with knowledge of a specific domain or discipline yet also a deep knowledge of data management. In this particular case, the Sheridan Libraries was fortunate that AAS had such an individual who provided invaluable feedback. Libraries would be wise to consider developing such expertise and capacity in-house. In this regard, the recent developments at Penn State are well worth examining.

Libraries, Publishers and the Humanities

In July 2009, Penn State University Libraries successfully concluded a search for the newly defined role of Digital Collections Curator, a position created after long-term planning and service development undertaken with both the Penn State Press and the campus computing division, Information Technology Services (ITS). In 2005, the Libraries and Press jointly established the Office of Digital Scholarly Publishing to develop experimental programs for collaborative online publishing of standard scholarly genres such as journals, conference proceedings, and peer-reviewed monographs [see http://www.libraries.psu.edu/odsp]. In 2008, as part of a university-wide effort, the Libraries and ITS undertook joint strategic planning that defined a "Cyberinfrastructure, e-Content, and Data Stewardship program" to provide a cohesive suite of access, security, discovery, preservation, curation, repository, archival, and storage services for born-digital data. These planning activities identified overlapping concerns about content presentation and preservation that are shared among libraries, publishing and humanities scholars.

Including a university press in Penn State's broad plans for content stewardship had highlighted the limits of the Libraries' approach to content development and production, which, except for electronically submitted dissertations, was exclusively focused on digitization and re-formatting of physical collections. Working with the Press's production department to move manuscripts to online and physical formats demonstrated that the Libraries' operational practices and policies did not yet easily accommodate collaboratively produced work, or born-digital work produced entirely out of the Libraries' immediate control. Penn State has made a virtue of decentralization in its digital library functions. With no central "digital library" group, the constituent activities are part of the normal operations of the Preservation, Cataloging, and Information Technology units, as well as significant infrastructure support from Digital Library Technologies, a division of ITS. This is a strength in that it links the "digital library" to well-established librarianship in core service units. But such an approach limits development of a unified program to support emerging needs or experimental approaches, which often require more dedicated resources.

Digital curation is a useful label for that collection of challenges newly located at the intersection of publishing, collections development, preservation, and the humanities. Humanistic scholarship depends upon establishing and interpreting relationships among primary sources, artifacts, and ideas. Beginning in the mid-90s, digital humanities projects like the Valley of the Shadow or the Walt Whitman Archive worked with research libraries and smaller archives to provide electronic access to previously inaccessible archival materials, simultaneously exploiting the Web's ability to establish and represent novel structural and interpretive relationships among those materials [see http://valley.lib.virginia.edu/ and http://www.whitmanarchive.org/]. Such projects were sometimes referred to "digital thematic collections," hinting at the close relationship between library and humanistic practices underlying these projects [see Kenneth M. Price, "Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?" http://www.digitalhumanities.org/dhq/vol/3/3/000053.html]. Kathleen Fitzpatrick, in Planned Obsolescence: Publishing, Technology, and the Future of the Academy (forthcoming, NYU Press), builds on this notion, arguing that digital humanities authorship is essentially a form of curation: a process of selection, managing, and adding value to existing materials.

Relationships among digital objects are as critical to preserve as the objects themselves, because many such objects may have no discernable purpose or limited meaning without the context. The book or article, the citations, and the data that was created during the research are in themselves less meaningful than they are as a collective. It is this network of relationships, and the data models that represent them, that libraries are now struggling to enable and to make durable parts of the scholarly record. As libraries begin to collaborate with University Presses and act as publishers or service providers, this collaboration will re-enforce and make more critical their role as long-term preservation agents. The Office of Digital Scholarly Publishing at Penn State distributes online, open-access versions of peer-reviewed monographs that are published by the Press in its Romance Studies series [http://romancestudies.psu.edu]. A persistent URL for the e-book version is included on the copyright page of each book, which makes a clear commitment to that e-book's durability going forward. Putting a standard monograph series online didn't make the Library a publisher, but it linked the Library's role as a preservation agent more directly to its emerging role as a distributor.

The Penn State Press and its authors want to be able to do more with the online version than reproduce a physical text. While the term "Data curation" often evokes huge scientific databases produced by esoteric research projects, humanists produce data too, and they include the bibliographic information and much of the other "stuff" that we have in our libraries. Could the Library publish an extended bibliography? Build links to the journal articles in the citations? Digitize some of the key texts referenced in the monograph? How about all of them? Create a searchable database and map of the 19th century (non-English) government records used as evidence in the book? Idiosyncrasy is the common element in humanities research, even though the long-form argument is the gold standard for professional advancement. Each case, when brought forward to the Library, was presented as completely unique, dependent not only upon the subject matter, but also the methods and working style of the author, who had collected his or her data in one particular way or the other.

Libraries are based on standardization, imposing organizational patterns on existing knowledge. Now we need to balance the uncertainty of the future technology and the novelty inherent in innovative

scholarship and with the need to standardize on data formats and protocols to support replicable services. To serve as a responsible steward of Romance Studies and any ancillary materials, the Libraries and the Press must ensure a degree of uniformity of these projects, and yet do so without negatively affecting the conception of the scholarly work. Ultimately these materials would be as valuable as they are connected to a wider set of digital content created by other scholars. If authorship is a form of curation, then by extension, publication needs to entail that activity too.

Within this context, Penn State created the Digital Collections Curator role to bring a combination skills and knowledge that would help Libraries better respond to these emerging issues. The Libraries' job posting announced that "Digital Collections Curator will lead the Libraries' efforts to develop and plan user focused services that enable the effective creation, sharing, discovery, and use of digital content in support of research, teaching and learning," with responsibilities to:

- Lead development of an inclusive, user-focused agenda for digital scholarly content stewardship.
- Investigate, recommend, and develop plans for user-focused and repository- based services to effectively manage the sustainable creation, collection and distribution of high-value digital scholarly content.
- Manage a broad set of existing digital collections and repository content, including: reformatted materials (images, books, newspapers, manuscripts, etc.), publication related content (journals, conference proceedings, monographs, hybrid formats, post & pre-prints, working papers, etc.), as well as the potential and emerging needs for data collections in a wide array of disciplines.

Penn State sought candidates with a "masters degree in library science or a relevant field, three years experience in the creation and management of digital resources, the ability to lead and work collaboratively in an evolving and decentralized environment, a commitment to user focused design, development, and service provision; and communication skills that will support work with both technology experts and novices." Implicit in this role is the need to make things up as you go along, basing the approach in a strong understand of the scholarly community to be served.

The emerging data curation field will require practitioners to understand deeply the process of scholarship and how that process can be modeled in systems and processes of complex organizations such as libraries and publishers. In fact, the person eventually hired into this role at Penn State had earned an advanced degree in the humanities prior to earning a MLIS from a program with a specialty in data curation. (The curriculum was developed in part with funding from the IMLS, see below). To be successful in the future as long-term stewards of information, libraries, archives, and museums must build on existing practices that have historically served us well, but we must also open up these practices to accommodate new forms of messiness and chaos that will challenge traditional roles as the expert curator.

IMLS Investments in Digital Curation

As we can see from these two cases, libraries are re-inventing themselves to adapt to changes in scholarly practices brought about by technology. Other changes are being forced not only by new technologies but also by economic pressures. Library acquisition budgets have been drastically reduced while the cost of books and journals is rising. Many newspapers have ceased publication, while more and more content appears only in digital form, often created and posted online without any commitment to or plan for preservation [e.g., YouTube, Flickr, and blogs]. This data is likely to be used for research purposes by humanists and social scientists, which suggests that capture and preservation of this raw content along with the published analysis would form a valuable part of the scholarly record for validation, replication, and future re-use. How can this new scholarly workflow be supported and documented? Who will be responsible for preservation of this large body of primary and secondary sources? Who will be responsible for preservation of scientific datasets, especially in domains that lack large repositories? Should researchers be responsible for managing their own data, or can the research to publication to preservation process be supported by services to make the workflow smoother and the preservation task easier?

To address these questions, IMLS has funded several research and demonstration projects to investigate the potential role of libraries in digital curation. In addition to the Johns Hopkins project discussed in this paper and funded in 2006, IMLS made an award in 2007 to the Purdue University Library's Distributed Data Curation Center, in a partnership with the University of Illinois Urbana Champaign, to address the question, "Which researchers are willing to share data, when, with whom, and under what conditions?" The case studies of researcher data/metadata workflow, and curation profiles describing policies for archiving and making available research data across different disciplines, are being used to develop system requirements for managing data in a repository and recommendations for implementing results under diverse systems. In addition, the project will describe the roles of librarians and identify the skill sets they need to facilitate scholarly communication and data sharing. This project has important implications for academic libraries and points to a new role for libraries in the provision of data management and preservation services to researchers throughout the data's life cycle [see http://d2c2.lib.purdue.edu/]

Developments in digital curation with regard to the role of libraries also have major implications for education in library and information science, especially in the academic library and archives specializations. In 2006, IMLS called for proposals in the Laura Bush 21st Century Librarians program to develop programs of study in digital curation in graduate schools of library and information science. Awards were made to three institutions: the University of Arizona, the University of Illinois at Urbana Champaign, and the University of North Carolina at Chapel Hill. Each of these programs is now well established, and all are attracting students at the master's and post-master's level, including opportunities for online study and summer institutes. Other programs have since been funded, including, notably, the University of Michigan and the University of Tennessee, and numerous schools have developed programs and courses of study in digital preservation and stewardship. In all, graduate schools of library and information science are preparing students who can help libraries make the

transition from guardianship of collections to the provision of services in the networked world. Evidence is growing that these skills are needed, and the sooner the better.

Conclusions

We suggest that academic libraries and publishers can succeed in today's challenging environment of economic scarcity through collaboration, with each contributing the things they do best. For publishers, this means adding value to the end-products of research, including processes such as peer review and skills such as high-quality editing and presentation. For libraries, it means building on their traditional preservation mission and their awareness of standards that cut across disciplines. As the preservation of digital data becomes increasingly important for different phases of the research process—from raw data to intermediate data sets that can be analyzed and interpreted to create a published conclusion—the ability to preserve large amounts of interoperable data over a long period of time becomes critical. It also becomes exciting, because we can begin to imagine future interdisciplinary scholarship that expands knowledge in ways that are currently unknown.

With the recent move of several universities to place the university press under the administrative auspices of the university libraries [including, notably, the University of Michigan and Purdue], the potential expansion of library services to include the "front end" of publication support in addition to the "back end" of digital curation seems even more likely.