Aggregation, Integration, and Openness: Current Trends in Digital Libraries.

David Seaman
Executive Director, Digital Library Federation

Abstract

Major US libraries and their users are trending towards a much greater desire for metadata and content that is nimble and agile, that can be combined with other items locally, and that can be integrated with courseware systems, desktop scholars' toolkits, and local archives. Aggregations dictated by publishers (and libraries) – the so-called "data silos" -- are very valuable but increasingly they are no sufficient to our needs if that is the only context in which that content can be accessed. The realities of the local service needs and the growing ambitions of users mean that we need more streamlined, flexible, time-saving, and interactive access than we currently enjoy. Advances in local institutional repositories, our growing ambitions for digital curation, and the developing dialogue between libraries and their users concerning open access scholarship, all argue for richer aggregation, integration, and control than we now have over the bulk of our digital library holdings.

Keywords: *digital library; data silos; open access; institutional repositories; courseware systems; Digital Library Federation; user services; finding systems; data sharing..*

Introduction

The observations in this paper are based in part on my decade in the University of Virginia's digital library, producing and delivering electronic texts and digital images, and working closely with many students, teachers, and scholars who incorporated their use into their daily academic lives; and in part on my experience of the past two years as director of the Digital Library Federation (DLF). To understand the perspective I currently have – one skewed firmly in the direction of large American and European academic libraries, it helps to have some context on the shape and activities of the organization I direct.

Defining the Digital Library Federation

The DLF is a leadership organization, made up currently of thirty-three strategic partner institutions – mostly major US university libraries, including Harvard, Princeton, Yale, Stanford, Michigan, MIT, Chicago, and the University of California, Berkeley; major academic libraries outside of universities such as the Library of Congress, the National Archives, and the New York Public Library; and – most recently – our first non-US member: the British Library. DLF also includes four allied organizations [please see *Appendix I* for a full list of member institutions].

DLF is a young organization still, first created in 1995 by a small number of library directors who felt the need to have an organization that focuses exclusively on their rapidly evolving digital library needs; it prides itself on being nimble and agile in its ability to respond to a rapidly changing set of issues and opportunities. Its work is done in a thoroughly collaborative manner – DLF has a very small central staff and operates through a series of initiatives and working groups whose members are drawn from the experts found in our partner libraries, and very often augmented with librarians, scholars, and software experts from other institutions. Our aim is always to have the right people on working groups we fund, irrespective of what institution they come from.

Working Areas

DLF concentrates on a range of practical and strategic areas of activity that serve various subsets of the membership. Increasingly, we have in addition a central, overarching, strategic goal – the formation of a distributed, open, digital library that encompasses all of our digital holdings and that facilitates much richer discovery and use of our content than the current, disparate, online experience afforded by our various websites – and I will return to that later in the presentation.

User Services

- *Dimensions and Use of the Scholarly Information Environment* [1] – a large survey of the usage habits of digital library users in American universities, whose findings are helping to inform our local strategic planning.

- Learning technologies and courseware – Dale Flecker (Harvard) is currently leading a DLF initiative (funded by the Andrew W. Mellon Foundation) to examine the interface between

courseware systems, publishers' sites, and the library's repositories of content and metadata.

Metadata Standards

- OAI: Open Archives Initiative [2] – DLF has provided some of the core funding for OAI in its early years, and is now moving forwards with Emory, Michigan, Illinois, and OCLC to take the lessons learned from the first round of harvesting services and reflect them back as a set of best practices and enhancements.

- METS: Metadata Encoding & Transmission Standard [3] – in part funded and supported by DLF, this format is rapidly gaining favor as a common way for us to encode descriptive, administrative, and structural metadata within a digital library

- Ongoing work with CrossRef, DOI, and other aspects of the persistent identifier challenge

Resource Management

- Electronic Resource Management Initiative: an XML format for managing content licenses in a much more efficient way [4]

- *The Registry of Digital Masters* (with OCLC) [5]; provides a place for institutions that have created digitized (page-imaged) versions of originally-printed monographs and serials to record what items have been digitized, where they can be accessed, and what specifications were followed in the digitization.

Production

- Production standards: *Benchmark for Faithful Digital Reproductions of Monographs and Serials*. [6]

- Cataloging Standards: *Describing Cultural Objects and Images* [7]

- Best practices for digital production workflows

Preservation

- Preservation of electronic scholarly journals [8]: a series of recent studies: Yale, Harvard, and the University of Pennsylvania worked with individual publishers on archiving the range of their electronic journals; Cornell and the New York Public Library worked on archiving journals in specific disciplines; MIT's project involved archiving "dynamic" e-journals that change frequently; and Stanford's involved the initial development of LOCKSS ("lots of copies keep stuff safe").

- Global Digital Format Registry [9]

Library Trends

Over the past two years I have had the opportunity to visit many academic libraries and observe the work they are currently doing as they integrate digital holdings, tools, and techniques into their preservation, access, and user service activities. The trends I discuss here are neither necessarily equal in importance, nor uniform in the attention and resources we afford them; however, they do seem to be significant enough to enumerate and put up for general consideration, even though this means a lack of specificity in my treatment of any one of them.

Trend 1: Courseware systems and the Library

Many American universities and colleges have installed courseware systems in recent years – software packages that help faculty manage their teaching through course-specific web pages, and which provide a suite of ancillary functions from automatic grading to online discussion forums. *Blackboard* and *WebCT* are the most common of the commercial systems; larger institutions are just as likely to build their own systems – Stanford's *CourseWork* is a good example[10] – and in a recent development we see this development going on in a coordinated manner across campuses in a project named Sakai: "The University of Michigan, Indiana University, MIT, Stanford, the uPortal Consortium, and the Open Knowledge Initiative (OKI) [have joined] forces to integrate and synchronize their considerable educational software into a pre-integrated collection of open source tools."[13] We also see related standards development such as the Open Knowledge Initiative (OKI) [11], and new public services such as Open Courseware at MIT – where over 500 courses are available free as pre-packaged web publications [12].

The principal concern for libraries with the recent campus deployment of courseware systems is that they are often installed and run by our IT departments without much library involvement, and there is too often a poor interface (both human and technical) between the library content management systems and the courseware systems. There is too little ability to link from the webpage for a course to the digital library holdings that support it. The advent of courseware also brings with it other library opportunities and challenges, including the wholesale archiving of course content in institutional

repository systems, the complicated management of rights to licensed content once downloaded into a teaching software module in a courseware system, and -- in at least one case -- the move of the courseware system itself to be under the control and support of the library. The advent of courseware holds great promise for moving the library into the classroom in ways hitherto unimagined, but so far the reality falls far short of this welcome opportunity to engage even more richly with the work of our teachers and students.

Trend 2: Authentication as an Enabling Technology

This is not the most forceful of trends, but important nonetheless; we are finally seeing a meaningful alternative to the simple and flawed IP address authentication that we all still use to govern access to content licensed for use by a specific institution: Shibboleth, [14] a product of the Internet II middleware initiative, controls access not by the location on the web from which you try to access a resource (IP authentication – "I am legitimate because I come to you from this location") but by trading information about the attributes of a user – "I am enrolled in Fine Arts 101 in the Spring 2004 semester" -- between a user's home institution to a publisher or vendor. This allows the user to be physically located anywhere on the web, and allows for much richer granularity of licensing. Shibboleth has caught the attention of many of us in the larger academic libraries: Pennsylvania State University and North Carolina State University have already conducted a successful test of Shibboleth, using it to control access to resources licensed for use in two classes, one at each institution. Major publishers and aggregators such as EBSCO, OCLC, JSTOR, and Elsevier are all interested or already implementing Shibboleth too. Moreover, the success of Apple's *iTunes Music Store* service gives hope for a more customer-centric direction in other Digital Rights Management (DRM) enforcement schemes – ones that treat the fee-paying customer with some respect, unlike the first wave of systems for e-book and music files. Our ambitions for richer library services are greatly aided by more nuanced protection and authentication schemes, so these are welcome advances.

Trend 3: Digital Archiving, Curation, and Preservation

Not surprisingly, preservation and archiving are among the most active areas for digital library endeavors. They address our core competencies, and provide a fertile ground for our natural abilities as custodians of scholarly works to think and plan over a long period of time. Major academic libraries are leading the push towards institutional repositories that store the intellectual assets of our faculty in all its forms – databases, images, teaching modules, computer simulations, finished scholarship – on the assumption that this is both a service to the individual scholar, a rich source of re-useable material for others, and a necessary part of our university infrastructure as we transition to a generation of scholars for whom all new scholarship is digital. The DSPACE Federation [15] is the best-known of these efforts, and DSPACE repositories are now implemented in a number of American and European libraries.

The international activity in the area of digital preservation is noteworthy: there is rich ongoing work in Australia and New Zealand [16]; the UK's JISC has just funded a Digital Curation Centre to provide national research and advice on the storage and lifecycle management of digital objects [17]; and in the US we have The National Digital Information Infrastructure and Preservation Program, led by the Library of Congress [18]. The US Congress has already made $25 million available for the planning and prototyping stages, with an additional $75 million (to be matched by recipients 1:1, to yield $150 million) to be made available in the second phase of this preservation infrastructure.

Trend 4: Digital Production (and Tools for Use)

Most large US academic libraries are producing digital objects locally, drawing on their physical collections for items that are good early candidates for digitizing, and are often being driven by demand for certain works in electronic form from their teaching and research faculty. Indeed, in many places we are seeing a shift from a series of discrete projects (sometime undertaken with outside grant money) to an ongoing production process, in which it is assumed that some level of digitizing is a permanent part of the service that the library offers. As we move forwards in our digital production efforts we become much more attuned to the need for common benchmarks of quality – largely present when dealing with text and image materials – and of registries such as the DLF/OCLC *Registry of Digital Masters*: a MARC-based catalog of page-image reproductions of printed monographs and serials.

Centralized production within an institution makes it easier to discuss larger-scale centralization, and we are seeing discussions taking place again around the subject of regional, multi-institutional digital production enterprises, to extend the economies of scale even further.

Much of this activity recognizes that a large library is particularly well-suited to digital production – it has the material, it has existing metadata, it has an ability to raise outside funding from donors and grants, it has a high degree of technical proficiency and often existing digital library delivery software, and – in America at least – it has an inexpensive workforce in its students, who typically work a part-time job while studying. Ongoing digitizing activities also reflect a growing sense that there simply is not enough digital content available to some of our users. The sciences are well-served, as are some areas of law and business, but typically the humanities and social sciences are still content-poor. One aim of the DLF's Distributed Library initiative is to encourage and coordinate a greater local investment in digitization. Nationally, we see ambitious proposals such as the Digital Opportunity Investment Trust (DOIT), now seeking adoption by the US Congress. [19] This "digital gift to the nation" proposes to take a portion of the proceeds from the future sale of telecommunications bandwidth to create a $20 billion trust fund, held by the US Treasury, the interest from which ($1 billion per year, approximately) would fund massive new investment in digital content from our libraries and museums, new research into digital pedagogy and life-long learning, and rich new tools for students, teachers, and the general public.

The recognition that we need tools as well as content ands context is also evident to us as librarians as we work especially with humanities scholars and students. I have already alluded to courseware portal-based tools such as Sakai, but we are beginning to see greater library interest in the mechanisms to use and transform our standards-based content.

Trend 5: Service Layers/Deep Sharing

Arguably the most active and overarching trend in the libraries I see is the growing dissatisfaction with the fragmented data landscape we have to offer our users, and the need for richer abilities (on both large and small scales) to integrate the content we buy and license with that which we build, and to re-shape the various commercial offerings into services and collections that make sense in a local context..

Significant work is going on in this arena, and we have some helpful tools and protocols at our disposal: the Open Archives Initiative, OpenURL [20], and CrossRef [21] all address different parts of the "data silo" problem. Even so, there exists a fundamental need to have content that encourages local re-organization and creation of services, and that permits individual users to progress beyond browsing and searching on sites created by others.

Scholarly publishers and digital libraries alike produce isolated silos of data that integrate poorly with others. A good silo is a lovely thing – but not always sufficient to our needs if the data contained in it – journal articles for example – can only be accessed through that one interface and alongside other content published by that producer or aggregator. A website containing all the items published by – say – Oxford University Press may well be a wonderful thing when that is what you need, but often the arrangement of content you want is other than this one. Libraries don't shelve physical books by publisher and users don't often work this way – it runs counter to our normal patterns of behavior. And yet too often that is the ordering principle of our digital library holdings. Too often we build product that can only appear on our terms, in our interfaces, in our tools, on our site.

Libraries face a chronic inability to repackage content for local use – in this, we are failing in our service mission to our customers. It is not dissimilar to the challenge we faced before the web, with content on CD-ROM that was isolated one from another. We've moved the problem online, but have not solved it fundamentally. Now you can suffer data isolation from the comfort of your home.

An illustrative example: imagine a 19th-century history undergraduate course, PDA-equipped and using (say) *WebCT*. The professor finds 100 relevant objects at 25 online archives (images, letters, and newspapers) and 20 journal articles in 10 different online journals. What can he or she do now other than create an online bibliography to this disparate material? In some cases the journals do not support persistent linking at the article level; none of them are available from their source for the PDA; they can't be searched all at once, or annotated; they are a mishmash of clashing aesthetics and functions. Such a data landscape is not encouraging of deep engagement with the content, of personalization, or of re-use.

Hopping in and out of many different web sites is also time-consuming. The DLF's *Dimensions and Use of the Scholarly Information Environment* survey made clear that lack of time is a critical issue: 38.8% of the total sample of respondents and 60.2% of the faculty reported "not having enough time" as their major problem in using online resources, [22] and the current isolation of data sites makes it very difficult for a library to address this problem with customized local aggregations and services.

What do we need to move forward?

Malleability: We need the data that resides on publisher and library sites to be much easier for us to re-shape for local customized delivery and analysis. We need to match the delivery format with the immediate needs and location of our users.

Management: We need the ability for a library to build local services that allow users to interact richly across vendors. Publishers could do much to help libraries be data aggregation services for the libraries' customers. Even the consistent use of OAI records, with all their limitations acknowledged, would be a real step forward.

Multiplicity: PDA, wireless, ebook, text-to-speech, and print-on-demand are all here or coming, and content that cannot go there will increasingly under-achieve.

Mixability: too often we invite our users to visit sites and watch content channels (a passive use, rather like TV); sometimes their needs are better served by the ability to sample, re-use and re-package – perhaps to form a personal library, or a classroom presentation (rather like the music mix that takes pieces from lots of CDs and creates a new compilation).

Mass: we need more content, and more innovative use will drive more creation.

The Distributed, Open, Digital Library

In its founding charter in 1995, the DLF called for the creation among its members of a distributed, open, digital library – a coherent view of its dispersed collections. The organization has recently re-affirmed this as a major strategic goal, which is beginning to take shape in two main areas of activity:

a) A Scholarly Finding System, initially based on OAI metadata harvesting, but in a second phase also encompassing research on the "next-generation" tools that we need to navigate dispersed and growing collections (richer metadata; federated searching of content; semantic searching and clustering; visualization of result sets). Currently, our dispersed content is too hard to find, and we cannot simply turn to general tools such as Google to fill the need – our digital objects are often part of that hidden web that Google does not see, and we need to utilize the metadata we have in the discovery process.

b) Digital Object Sharing: we recognize that the needs of some users, and some of our library service ambitions, are most easily served if the digital object can be easily moved from repository to user or from repository to repository. This "deep sharing" requires a host of new infrastructure, rights management, policy, data description, and data structuring agreements.

In both of the areas of activity above there is much to do, and some basic new behaviors to learn, including the routine provision of harvestable metadata when we create a digital object, and the conscious creation of content that can work in someone else's system.

Conclusion

The transformation from isolation to integration is our central challenge and opportunity, with some enormous payoffs when we get it right. Innovative users and library services providers need malleable content with which to engage and innovate; it is not sufficient simply to offer the current fragmented set of websites defined by publishers, aggregators, or libraries as the only way to access our rich, standardized, and re-purposeable content.

****

Notes

1) *Dimensions and use of the scholarly information environment* http://www.diglib.org/pubs/scholinfo/

2) *Open Archives Initiative* http://www.openarchives.org

3) *METS: Metadata Encoding & Transmission Standard* http://www.loc.gov/standards/mets/

4) *Electronic Resource Management Initiative*. See: http://www.library.cornell.edu/cts/elicensestudy http://www.diglib.org/standards/dlf-erm02.htm

5) *The Registry of Digital Masters* (with OCLC) http://www.diglib.org/collections/reg/reg.htm

6) *Benchmark for Faithful Digital Reproductions of Monographs and Serials.* DLF: Washington DC, 2002. http://www.diglib.org/standards/bmarkfin.htm

7) *Cataloging Cultural Objects: A Guide to Describing Cultural Works and their Images*. Visual Resources Association, 2004. http://www.vraweb.org/CCOweb/index.html

8) *Archiving Electronic Journals: Research Funded by the Andrew W. Mellon Foundation.* Edited, with an Introduction, by Linda Cantara. DLF: Washington, DC. 2003. www.diglib.org/preserve/ejp.htm

9) The Global Digital Format Registry http://hul.harvard.edu/gdfr/

10) *CourseWork*, Stanford University http://getcoursework.stanford.edu/

11) The Open Knowledge Initiative: http://web.mit.edu/oki/

12) MIT OpenCourseWare: http://ocw.mit.edu/index.html

13) The Sakai Project: http://www.umich.edu/~sakai/index.html

14) Shibboleth Project: http://shibboleth.internet2.edu/

15) DSPACE: http://www.dspace.org/

16) Ongoing work on digital preservation in Australia and New Zealand can be found at http://www.nla.gov.au/preserve/ and http://www.natlib.govt.nz/en/whatsnew/4initiatives.html

17) JISC Digital Curation Centre: http://www.ucs.ed.ac.uk/bits/2004/february_2004/

18) The National Digital Information Infrastructure and Preservation Program http://www.digitalpreservation.gov/ http://www.diglib.org/forums/fall2003/fallforum03.htm#p18

19) Digital Opportunity Investment Trust (DO-IT): http://www.digitalpromise.org/

20) The OpenURL Framework for Context-Sensitive Services (NISO) http://www.niso.org/committees/committee_ax.html

21) CrossRef: http://www.crossref.org/

22) *Major Problems* summary table, from *Dimensions and use of the scholarly information environment* www.diglib.org/pubs/scholinfo/question25017.htm

Appendix I

DLF Partners and Allies
The British Library
California Digital Library
Carnegie Mellon University
Columbia University
Cornell University
Council on Libraries and Information Resources
Dartmouth College
Emory University
Harvard University
Indiana University
Johns Hopkins University
Library of Congress
National Archives & Records Administration
Massachusetts Institute of Technology
New York Public Library
New York University
North Carolina State University
Pennsylvania State University
Princeton University
Rice University
Stanford University
University of California, Berkeley
University of Chicago
University of Illinois at Urbana-Champaign
University of Michigan
University of Minnesota
University of Pennsylvania
University of Southern California
University of Tennessee
University of Texas at Austin
University of Virginia
University of Washington
Yale University

Allies
Los Alamos National Laboratory Research Library
Online Computer Library Center
Research Libraries Group
Coalition for Networked Information