Electronic Publishing, ETD's and Institutional Repositories

Robert P. Spindler
Arizona State University

Colleges and universities, and especially research universities, have a long tradition of publishing associated with their mission to widely disseminate the research products of faculty. Helen Samuels indicated that "The purpose of the research literature is two-fold: to communicate the results and discoveries in order to expand knowledge and to stake a claim to the results."[1] Historically university presses and journals of scholarly associations hosted by universities conducted this activity, typically resulting in publication of hardcopy monographs and "archival" quarterly journals. In most cases the university library acquires formal scholarly journals and monographs for their subject content, but some university archives acquired materials published by their institution for their evidential as well as informational value. This can result in some duplication of effort given that the same title may be needed for both functions, and archives cannot depend on subsequent transfer of materials from the general collection since the materials may be lost or damaged by circulation and heavy use.

Many institutional archives have defined official university publications as archival records of the institution and have included them in their collection development policies. Samuels acknowledged the importance of collecting university publications, writing "Colleges and universities have a particular obligation to ensure the availability and retention of the works issued by their own institution.[2] Annual reports, student and staff newspapers, newsletters, course catalogs, policy manuals and committee reports offer detailed accounts of the development of universities and their component colleges and departments over time. University publications serve as the richest and most commonly sought sources of institutional memory. But very few universities have established requirements for delivering copies of official publications to the archives, and the number and bulk of these materials could overwhelm archival programs. As a result, archivists often manually populate their publications holdings. Typically, archivists acquire an archival copy of official periodicals by maintaining subscriptions of the key titles with the producing offices, or cold-calling offices to acquire unique titles as they are released.

In the late 1980's and early 1990's, desktop publishing technologies became widely available for use on MacIntosh workstations and PC's. This enabled the inexpensive and local production of a wide variety of less formal hardcopy publications such as departmental newsletters and brochures. Many academic and administrative departments of universities embraced these technologies to inexpensively facilitate services marketing and student recruitment.

Archiving the *products* of desktop publishing did not represent a new challenge for academic institutions, as the paper product could be cataloged and filed as before. However, retention of the electronic *source files* produced by desktop publishing gave academic units a preview of the electronic preservation challenges that would grow with the use of new publishing technologies. Suddenly universities had digital assets in the form of the raw materials used in electronically produced publications (e.g. photographs, text copy) that could be retained and used for other purposes.

Many archives developed a corpus of electronic finding aids in word processing and/or database formats in this period. These assets needed to be maintained indefinitely in order to realize operational efficiencies in updating finding aids when new materials were added to archival collections. Eventually the process of retaining electronic finding aids caused archivists to learn hard lessons about digital preservation and digital asset management, especially when

implementing the data conversions required by software that was not Y2K-compliant. But this experience positioned some archivists to serve as effective and experienced advisors when other areas of the university required their assistance with digital preservation and data migration issues.

Starting in the mid-1990's it was the combination of desktop publishing technologies, development or implementation of encoding standards like Hypertext Markup Language (HTML) and Standard Generalized Markup Language (SGML), and electronic distribution through the Internet that enabled the development of a wide range of new publication forms including general web pages, online staff directories, web-based policy manuals, electronic annual reports and digital technical reports. Online publishing demonstrated the potential to expand distribution, enable real time updating and decrease distribution costs. Academic and administrative units moved quickly to take advantage of this new media, however the costs of digital asset management were generally not perceived or funded in the early years of Internet publication.

Early Electronic Theses and Dissertations and a Case Study

The potential for applying electronic publishing technologies to scholarly research was recognized as early as 1987, when University Microforms International (UMI) called an exploratory meeting in Ann Arbor Michigan. Soon some university staff and administrators began to feel pressure from students who were asked to reverse-engineer their word-processed electronic documents to meet rigid graduate college format standards developed to facilitate hardcopy production. The development of the Internet and desktop publishing for the web enabled particularly motivated graduate students to use these technologies to create different forms of digital content that would serve as attachments to traditional hardcopy theses, or complete electronic texts intended to replace hardcopy.

Most often early efforts were delivered in the form of floppy discs or CD's that were bound in with a hardcopy thesis text. Librarians had some experience with electronic attachments in commercial publications, and descriptive cataloging standards for this form of early multimedia were quickly developed. However, the library and archives world was not well prepared for the advent of wholly new forms of scholarship that were being encouraged and approved by some faculty committees.

In 1995 Keith Voegele, a doctoral candidate in the Computer Science Department at Arizona State University, received approval of his web-based interactive dissertation entitled *Tessellation of Bibliographic Data: An Example Using Categorical Data.* The dissertation consisted of three components, a website containing citations to literature concerning web-based visualization technologies and some text describing creation of the site; a recordable CD where the student "archived" the C and Hypertext Markup Language (HTML) files that comprised the site (at the request of his review committee); and a hardcopy volume that contained a bibliography of the citations to visualization literature and text about creation of the site and how it worked. The website also included an interactive "tessellation" diagram, in which site visitors were invited to add their own new citations which would then be automatically plotted in the electronic diagram. Technical services librarians and the university archivist concluded that the most complete version of the dissertation was the website itself.

The student's committee demonstrated remarkable foresight in requiring the storage of site files on CD-R, but the hardcopy documentation did not convey when the CD-R had been written and whether the review committee saw the same version of the site that was copied to CD-R or

something else. Since the site was interactive the university had lost the opportunity to preserve an exact copy of the document that was approved by the committee. When the archivist attempted to open the CD-R he discovered it was formatted for Macintosh computers and the files were not write-protected in any way.

In an email exchange and conversation with the author, the archivist discussed the potential for rebuilding the site from the hardcopy documentation and the files on CD-R. Voegele suggested it would be impossible to recreate the site from these sources since there were certain compilers and other pieces of software resident on the server that could not be copied due to licensing and CD-R space limitations.[3] As a result the university could rebuild and display some pages that looked like the pages in the site, but could not create an exact or even a near reproduction of its functionality. In March 2000 the archivist returned to the site for the first time in a few years only to discover it had been deleted from the College of Engineering server.

This case study illustrates how certain academic disciplines and progressive faculty encouraged students to use new technologies and develop “new media scholarship” faster than university administrators were able to respond to the opportunity in the mid-1990’s. As a result some instances of early ETD’s have been lost because most universities had not established the policy base and infrastructure to physically acquire and reliably maintain the digital products. However, development of formal electronic publishing programs was already in progress at other institutions.

ETD’s and the Networked Digital Library of Theses and Dissertations

One of the earliest and boldest electronic university publishing ventures was formally initiated in 1992, when the Coalition for Networked Information, Virginia Tech (VT), the Council of Graduate Schools, and University Microforms International (UMI) issued a call for participation in the project entitled "The Capture and Storage of Electronic Theses and Dissertations". The 1992 project intended to establish an SGML Document Type Definition for the encoding and Internet distribution of theses and dissertations, but in 1994 VT identified both SGML and Portable Document Format (PDF) as their primary production format standards. VT received funding from the Southeastern Universities Research Association (SURA) to support their pilot project for 1996/1997 as part of the Monticello Electronic Library. In January 1997 VT shocked the academic world when they established an ETD submission *requirement* for all graduate students. The stated goals were to improve the quality of graduate education, improve the information and technical literacy of graduate students, and enable increased and simultaneous use of university theses and dissertations. VT posted over 1,000 ETD’s to the Internet by April of 1998.[4]

The advantages of this form of electronic publishing were apparent to several other major universities including Cornell, Berkeley, and Michigan, although Virginia Tech was the recognized pioneer in developing ETD’s. Virginia Tech computer science professor Ed Fox lead the effort to establish the Networked Digital Library of Theses and Dissertations (NDLTD), and received a three-year federal grant from the Department of Education in January 1996. The grant was followed by gifts from several corporate sponsors. UMI began accepting electronic submissions to its thesis and dissertations publishing program in 1997. The NDLTD attracted 29 member institutions by July of 1999, but by June 2004 the NDLTD boasted 184 members including several universities from Europe and Asia. Sixty-two of those members established some form of electronic submission requirement, while the balance of the NDLTD offered the option of electronic submissions. [5]

The NDLTD established a series of international symposia and working committees to develop shared templates, technical standards and a Unicode-based multilingual library catalog system to be developed by VTLS. ETD symposia and international conferences offered presentations by early adopters that addressed encoding standards, submission standards, and system architectures. Some universities simply allowed or required students to electronically submit their final theses product to the graduate college upon approval by their committee, while other institutions sought to reinvent the entire evaluation and submission process for works in progress, seeking efficiencies in the student editing, committee review and approval and graduation administration processes. Submissions of electronic copies and release forms to UMI were often included in the ETD process improvement work. [6]

The adoption of ETD's and the development of the NDLTD serve as mileposts in the development of electronic publishing by universities because they enabled substantial production efficiencies and greater accessibility for products of scholarly inquiry. ETD implementations could also be fast-tracked because universities could unilaterally impose publishing requirements on the graduate student body, a more difficult proposition in relation to faculty or administrative publication. The potential for ETD's to transform scholarly publication has yet to be realized in most disciplines, although certain visually dependent fields such as architecture and dance are quickly adopting more sophisticated combinations of digital text, sound and video. These efforts challenge archivists' perceptions of the fixity of documents, and create substantial new challenges for long-term preservation.[7]

Electronic Administrative Publications

As early as 1997 universities recognized the efficiencies of Internet distribution for certain publications that archivists and records managers would consider vital institutional records. Course catalogs and university policy manuals became popular targets for conversion to the web because of their volatile content and the expense of broad hardcopy distribution. In addition, web-based course catalogs would support the development of e-commerce applications for course registration and the subsequent development of online curricula.

Since these efforts were perceived and administered as process efficiency efforts, concerns for long-term retention and accessibility were not generally addressed in the early implementations. Here production efficiencies clashed with the litigation defense or public records requirements of universities, as the advantages of fast dissemination and real-time updating eclipsed long-term needs for institutional memory and accountability. Many institutions implemented electronic course catalogs and web-based policy manuals without attention to their record keeping needs.[8]

Now communications from high level administrators, academic senate and university committee reports are routinely distributed in electronic formats, most often encoded in HTML or converted to PDF files. But most universities are not collecting and maintaining these materials in a systematic or reliable way, and in most cases archivists are not at the table when electronic publication projects are planned and executed.

Electronic Publishing by Faculty

Meanwhile, some university faculty were embracing the possibilities of what has recently been termed "new media scholarship." In the early 1990's a steady stream of new scholarly reference materials were published as multimedia products, often combining hardcopy books with CD-ROM versions or supplements for wide distribution and fast revision. Most early CD-ROM

publications were built for use at a single workstation, but eventually these products were redesigned, loaded on “juke boxes” containing many similar products and made accessible through servers in order to accommodate multiple simultaneous users. As the Internet developed and the multi-user capacity of these products was exceeded, many of them were converted to web-based products in the late-1990’s. Many faculty and students were first exposed to scholarly electronic publication through their use of CD-ROM products available through academic library networks.

One group that has always valued fast and broad dissemination of research findings has been the physics community. In August, 1991 a group of physics faculty and professionals led by Paul Ginsparg of the Los Alamos National Laboratory founded *arXiv*, a pre-print publication server that would allow scientists to self-post their work in progress without external review. The physics community embraced this new form of scholarly communication as *ArXiv* boasted 170,000 article submissions in the first ten years. According to Ginsparg, “The original objective of the e-print arXiv was to provide functionality that was not otherwise available, and to provide a level playing field for researchers at different academic levels and different geographic locations -- the dramatic reduction in cost of dissemination came as an unexpected bonus.”[9]

But it was this “unexpected bonus,” within the context of rapidly increasing academic journal subscription costs, that sparked a wide and deep re-examination of scholarly publishing and the peer review system by librarians, faculty and university administrators. These changes in scholarly publishing and the advent of new forms of scholarly communication could forever change the roles and responsibilities of archivists in acquiring and preserving electronic university publications and other valuable academic content.

SPARC, D-Space, and Institutional Repositories

In the late 1980’s, research sponsored by Association of Research Libraries (ARL) identified a “crisis” in escalating costs of scholarly journals, especially those in science, technology and mathematics. One study concluded that the price-per-page of 160 core journals exceeded the growth in costs by 2.6% to 6.7% a year. This could mean that publishers were enjoying operating profits of 33% to 120% a year.[10] University Libraries across the country struggled to justify sufficient increases of acquisition budgets so universities could in some cases buy back the research of their own faculty.

Ann Okerson, in a consulting report commissioned by ARL, recommended that, “ARL should strongly advocate the transfer of publication of research results from serials produced by commercial publishers to existing non-commercial channels. ARL should specifically encourage the creation of innovative non-profit alternatives to traditional commercial publishers.”[11] Over the next several years ARL and other national organizations attempted to initiate dialogues with various stakeholders including university administrators and faculty to find alternatives to the rapidly escalating journal subscription costs.

In May of 1997, Ken Frazier, Director of Libraries at the University of Wisconsin, Madison, suggested the formation of a membership organization that would fund the establishment of ten alternative non-profit electronic journals. This suggestion led later that year to the formation of SPARC, the Scholarly Publishing and Academic Resources Coalition. “SPARC is a membership organization whose mission is to restore a competitive balance to the STM journals publishing market by encouraging publishing partners (for example, societies, academic institutions, small private companies) to launch new titles that directly compete with the highest-priced STM journals or that offer new models that better serve authors, users and buyers.”[12] SPARC has

been very successful launching several alternative and low cost publishing initiatives that have attracted the attention of faculty in many science, technology and medical fields.

Meanwhile, technical developments in systems design and the Internet resulted in several new tools for scholarly publication and scholarly communication. Perhaps the most important of these developments was the creation of server software entitled D-Space. D-Space was initially conceived in 2001-2002 as a joint research and development project of MIT and visiting scientists from Hewlett-Packard. “D-Space is an open source software platform that enables institutions to capture and describe digital works using a submission workflow module, distribute an institution's digital works over the web through a search and retrieval system, and preserve digital works over the long term.” [13] D-Space resources were opened for public access September 30, 2002.

D-Space is the flagship technology of several applications that were built in 2000-2002 to enable self-publication. The ETD community had been experimenting with student self-publication vehicles for several years, and alternative self-publication software such as Berkeley Electronic Press (known as Bepress) and Fedora[14] were directed at faculty and academic institutions. But the technology had the potential to do much more than facilitate self-publication. Creators and promoters of D-Space and Fedora recognized that the applications could be used to centrally capture, preserve and make available all kinds of digital objects including unpublished digital assets, raw data sets, electronic theses and dissertations, research works in draft, university websites and other electronic university records.

However, ARL and SPARC seized upon these technologies as a potential solution to the serials crisis since they could be used as a vehicle for inexpensive alternative scholarly publication. In October of 2002 over three hundred library administrators and a small group of archivists from the US and Canada attended *Institutional Repositories: A Workshop on Creating an Infrastructure for Faculty-Library Partnerships,* convened by SPARC and ARL at the historic Mayflower Hotel in Washington DC. There, several university library administrators including James Neal of Columbia University and Ann Wolpert of MIT promoted use of these technologies in institutional repositories (IR’s). Speakers principally addressed the potential for IR’s to host low-or-no-cost scholarly journals, but they also recognized the potential role of IR’s is supporting digital asset management for colleges and universities. [15]

Suddenly, the academic library community had taken interest in certain functions that had traditionally been assigned to archival personnel, specifically acquisition, preservation and access for electronic faculty papers and publications. Universities were making a substantial investment in the infrastructure to support those functions by creating institutional repositories. But until recently the archival profession did not recognize the opportunity to collect and preserve electronic faculty and student materials represented by IR’s. The Society of American Archivists has its first conference session on institutional repositories in August, 2004.

Convergence and Opportunity

The emergence of institutional repositories and the parallel development of online learning management systems have resulted in a unique intersection of formal and informal electronic publishing, the creation of online research and instructional communities, and opportunities for electronic records management and archiving, but the relative roles of the various stakeholders are still being sorted out. Clifford Lynch, Executive Director of the Coalition for Networked Information quipped, “I think that what we're seeing here in some sense is a convergence of sort

of traditional records concerns, the movement of a lot of the teaching and learning processes to digital form, the real transformation of how we're doing scholarship -It's getting real hard to tell what's a record, what's research, what's teaching and learning." Lynch closed his presentation by citing the variety of professional stakeholders who should be consulted in the process of building institutional repositories and preserving electronic publications and records associated with university research.[16]

Archivists now have a seminal opportunity to attract investment in some of our core archival functions and work with librarians, technology professionals, records managers, research administrators, university presses and faculty in the development of institutional repositories. Archivists have valuable perspectives on many related issues including donor relations, description, citation and branding, and especially digital preservation. The work of acquiring, preserving and making accessible university publications is no longer the province of archivists, now we have many more allies and some very sophisticated tools to achieve the goals we share with our universities, and the public at large.

rev. 2004/09/02

---

[1] Helen Willa Samuels, *Varsity Letters: Documenting Modern Colleges and Universities,* Metuchen, N.J., The Society of American Archivists and Scarecrow Press, 1992. p.132.

[2] Samuels, p. 133

[3] Electronic mail from Keith Voegele to Robert Spindler, October 23, 1995. See also Robert P. Spindler, *Preserving Interactive or Multi-Version Web-Based Documents*, 2000. http://www.public.asu.edu/~spindler/ETD2000.Preservation.proceedings.webversion.htm (Accessed June 23, 2004).

[4] Networked Digital Library of Theses and Dissertations: History Description and Scope, n.d. http://www.ndltd.org/info/description.en.html (Accessed September 2, 2004**;** Crowe, Martha J., *Publication of Electronic Dissertations*, Ithaca, New York, Cornell University Library, 1998 http://www.library.cornell.edu/staffweb/ETDSTUDY.HTML (Accessed May 18, 2004); Fox, Edward A., *etal,* Networked Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources, *D-Lib Magazine*, September, 1996 http://www.dlib.org/dlib/september96/theses/09fox.html (Accessed June 15, 2004); *ETD Project Team, The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations*, Blacksburg, Virginia, Networked Digital Library of Theses and Dissertations, 1998. http://www.ndltd.org/talks/ndltd980623.pdf (Accessed June 15, 2004)

[5] Fox, Edward A. Networked Digital Library of Theses and Dissertations, Blacksburg, Virginia, NDLTD, 1998 http://www.ndltd.org/talks/ASU990706.ppt (Accessed June 15, 2004);

[6] See http://www.ndltd.org/meetings/index.en.html for links to the ETD Symposia.

[7] Depocas, Alain, Jon Ippolito and Caitlin Jones, *The Variable Media Approach: Permanence Through Change*, New York: Guggenheim Museum, 2003. (138 pp.) http://www.variablemedia.net/ Accessed February 9, 2004.

[8] Arizona State University was one example of an institution that did address record keeping for electronic policy manuals and course catalogs in 1997. See Spindler, Robert P., "Preserving Web-Based

Records", *Preservation and Access for Electronic College and University Records Conference Presentations*, December, 1999. http://www.asu.edu/ecure/1999/spindler-presentation.html

[9] Ginsparg, Paul, *Creating a Global Knowledge Network*, Ithaca, New York, Cornell University, 2001. http://arxiv.org/blurb/pg01unesco.html (Accessed June 16, 2004)

[10] Case, Mary M., *Capitalizing on Competition: The Economic Underpinnings of SPARC*, Washington DC, Association of Research Libraries, 2002. http://www.arl.org/sparc/announce/case040802.html#notes (Accessed September 1, 2004).

[11] Case, 2002

[12] Case, 2002.

[13] D-Space definition, 2002. http://libraries.mit.edu/dspace-mit/what/definition.html (Accessed September 2, 2004)

[14] http://www.bepress.com/; http://www.fedora.info/ (Accessed September 2, 2004)

[15] A description of the workshop is available at http://www.arl.org/sparc/meetings/ir02.html (Accessed September 2, 2004) Raym Crow, a consultant for SPARC, issued an important position paper in advance of the workshop entitled "The Case for Institutional Repositories", available at http://www.arl.org/sparc/IR/ir.html (Accessed September 2, 2004).

[16] Lynch, Clifford, ECURE 2004 Keynote Address, 2004. http://www.asu.edu/ecure/2004/keynote/ (Accessed September 2, 2004)