2008-07-01

Digitizing Dissertations for an Institutional Repository: A Process and Cost Analysis

Mary E. Piorun
University of Massachusetts Medical School

Lisa A. Palmer
University of Massachusetts Medical School

Follow this and additional works at: https://escholarship.umassmed.edu/lib_articles

Part of the Library and Information Science Commons

Repository Citation
https://escholarship.umassmed.edu/lib_articles/94

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Library Publications and Presentations by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Digitizing dissertations for an institutional repository: a process and cost analysis

Mary Piorun, MSLS, MBA, AHIP; Lisa A. Palmer, MSLS

See end of article for author’s affiliations. DOI: 10.3163/1536-5050.96.3.008

Objective: This paper describes the Lamar Soutter Library’s process and costs associated with digitizing 300 doctoral dissertations for a newly implemented institutional repository at the University of Massachusetts Medical School.

Methodology: Project tasks included identifying metadata elements, obtaining and tracking permissions, converting the dissertations to an electronic format, and coordinating workflow between library departments. Each dissertation was scanned, reviewed for quality control, enhanced with a table of contents, processed through an optical character recognition function, and added to the institutional repository.

Results: Three hundred and twenty dissertations were digitized and added to the repository for a cost of $23,562, or $0.28 per page. Seventy-four percent of the authors who were contacted (n=282) granted permission to digitize their dissertations. Processing time per title was 170 minutes, for a total processing time of 906 hours. In the first 17 months, full-text dissertations in the collection were downloaded 17,555 times.

Conclusion: Locally digitizing dissertations or other scholarly works for inclusion in institutional repositories can be cost effective, especially if small, defined projects are chosen. A successful project serves as an excellent recruitment strategy for the institutional repository and helps libraries build new relationships. Challenges include workflow, cost, policy development, and copyright permissions.

INTRODUCTION

Digitization projects in libraries seem ubiquitous as libraries become increasingly involved in the acquisition, development, and management of digital information [1]. Libraries typically target archival and special collections materials such as historical documents and photographs [2]. Projects to digitize vast collections of books began as early as 1971 with Project Gutenberg and are now getting widespread media attention with the launch of Google Book Search, the Internet Archive, and others [3]. In an April 2007 list of ten assumptions about the future that would significantly impact academic libraries and librarians, the Association of College & Research Libraries Research Committee placed digitization at the top of the list, stating, “There will be an increased emphasis on digitizing collections, preserving digital archives, and improving methods of data storage and retrieval” [4].

A related emergent trend in academic libraries is the implementation of institutional repositories (IRs), digital collections that capture and preserve the intellectual output of university communities [5]. A search of OpenDOAR, the Directory of Open Access Repositories, lists 298 academic repositories in North America [6]. Health sciences libraries are among those contributing to this trend; of 125 libraries that responded to a 2006 supplementary survey for the Annual Statistics of Medical School Libraries in the United States and Canada, 28 have established IRs.

Highlights

● The Lamar Soutter Library partnered with the University of Massachusetts Medical School Graduate School of Biomedical Sciences to digitize doctoral dissertations for inclusion in a newly created institutional repository.

● Seventy-four percent of dissertation authors (209/282) gave permission for the digitization. The cost to process the entire dissertation collection in-house was $23,562, only $1,062 more than the estimate to outsource.

● Digitizing the dissertation collection increased access: the print collection was used 723 times in the past 5 years, while the electronic collection was used 17,555 times in 17 months.

Implications

● Digitizing student works is an effective way to begin populating an institutional repository.

● In-house digitization projects can be cost-competitive with outsourced alternatives.

● A repository can be a catalyst for developing relationships in the institution by providing the library with a new avenue for outreach.

● Skills and experience gained from a small project can be applied to larger-scale projects.

* Based on a poster at MLA ’07, the 107th Annual Meeting of the Medical Library Association; Philadelphia, PA; May 26, 2007; and a presentation at the Scanning Forum 2006; Charlottesville, VA; November 6, 2006.

J Med Libr Assoc 96(3) July 2008
and 70 are planning to add or are considering offering a repository [7]. According to Foster and Gibbons, libraries build IRs because they “provide an institution with a mechanism to showcase its scholarly output, centralize and introduce efficiencies to the stewardship of digital documents of value, and respond proactively to the escalating crisis in scholarly communication” [8].

Medical librarians are just beginning to report their experiences with institutional repositories in the professional literature [9–13]. In one case study, Krevit and Crays [13] describe challenges that the Texas Medical Center experienced in piloting a multi-institutional repository, including copyright concerns and lack of faculty participation. An analysis by Singarella and Schoening [14] of the surveys conducted between 2005 and 2007 by the Association of Academic Health Sciences Libraries and a survey conducted in 2006 by the Association of Research Libraries [15] confirmed that the challenges experienced at the Texas Medical Center were not unique. Libraries are the drivers of IRs at their institutions, as few faculty members identify and self-archive their own materials. Libraries struggle to recruit content and employ a variety of strategies to enlist submissions [16–19]. Content may vary, but a recent study by McDowell reports that student works account for the largest percentage of documents in institutional repositories, approximately 41.5% [19].

The following case study describes a nexus of these two trends: digitization of student scholarly works and institutional repositories. The first digitization project for the Lamar Soutter Library at the University of Massachusetts Medical School (UMMS) was to digitize 300 doctoral dissertations and add the full text to the school’s new IR.

BACKGROUND

Founded in 1970, UMMS encompasses the graduate schools of medicine, nursing, and biomedical sciences. The Lamar Soutter Library holds 175,000 print volumes and provides access to 316 databases, 4,650 electronic journals, and 359 electronic books. The IR is the library’s first comprehensive digital initiative.

In early 2006, the library purchased a license for ProQuest Digital Commons,† a hosted institutional repository, and named the repository “eScholarship@UMMS” <http://escholarship.umassmed.edu>. The team implementing the repository, a previously reported process [12], consisted of representatives from the library’s systems (project management and technical support), cataloging (metadata support), and reference (outreach) departments. In March 2006, the dean of the UMMS Graduate School of Biomedical Sciences (GSBS) expressed interest in digitizing the school’s dissertations. The GSBS had produced 300 dissertations, most of which were available only in print format. The team thought this would be an excellent demonstration project: it was supported by the dean, it was a manageable size, metadata could be reutilized from the library’s online public access catalog (OPAC), and the dissertation authors held the copyright. In May 2006, the library and GSBS partnered to make the dissertations fully searchable on the web.

METHODOLOGY

Outsourcing versus insourcing

The team investigated 2 options for digitizing the dissertations: outsourcing to UMI or performing the work in-house. UMI estimated the cost to be $75 per title ($22,500 total) and 8–12 weeks processing time. The basis for the library estimate was created by library staff scanning and locally preparing 3 sample dissertations. Table 1 shows the library’s cost estimate of $27,750—for staffing, project management, equipment, and software—and 725 hours of processing time (or 18 weeks when represented as a 40-hour work week). In all instances, except for project management, the team assumed the work would be performed by temporary help. The team had 2 issues of concern: at the time, electronic files created by UMI were not full-text searchable, and the graduate school would need to commit to sending all future dissertations to UMI to keep the database current.

The project team recommended that the library process the dissertations in-house, despite longer time to process and higher cost, in order to gain experience, retain access to materials throughout the project, and have tighter control over scanning quality. Library administration accepted the recommendation to do the digitization locally, citing “gaining experience” as the major benefit; however, $27,750 was not available to fund the project. Ten thousand dollars was allotted to hire temporary staff, with the understanding that circulation staff and interlibrary loan equipment would be utilized for scanning and team catalogers would add dissertations to the repository. It was also recognized that the project could not be completed in 18 weeks as staff assigned to the project would need to incorporate the dissertation tasks into their daily workload.

Metadata

To fully utilize metadata from the library’s integrated library system, team catalogers customized default templates in the Digital Commons software designed to control the indexing and display of a collection of records. Customizations were necessary to fully describe the dissertations and incorporated features such as the activation of live link functionality in fields where uniform resource locators (URLs) might be included, the addition of a field to record authors’ UMMS departmental affiliations, and the accommo-

† In July 2007, Berkeley Electronic Press (bepress), the original developers of the software, resumed full support of the Digital Commons product. It is now called bepress Institutional Repository. For more information, see the product description available at <http://www.bepress.com/ir/>. 
Table 1  Estimate versus actual costs and processing times

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Estimate (300 documents)</th>
<th>Actual (320 documents)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
<td>Rate ($)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cost ($)</td>
</tr>
<tr>
<td>Scanning</td>
<td>225</td>
<td>20</td>
</tr>
<tr>
<td>Quality control</td>
<td>225</td>
<td>25</td>
</tr>
<tr>
<td>Optical character recognition (OCR) of the abstract</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>Add to institutional repository (IR)</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>Signature page</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Replace file</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>Project management</td>
<td>75</td>
<td>35</td>
</tr>
<tr>
<td>Equipment/software</td>
<td>10,000</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>725</td>
<td>27,750</td>
</tr>
</tbody>
</table>

* Page length is estimated at 250 per document.

...
areas were underestimated, such as the time to OCR the abstract and overall project management. Issues that contributed to this miscalculation include the extra time to correct the scientific notation in the OCR process and the total project management time required to obtain permissions from authors to digitize their work.

**Equipment and software**

The work was accomplished using existing library scanning equipment. The library already owned copies of the software used throughout the process: Microsoft Access, eCopy, Adobe Acrobat, and Adobe Illustrator. Because eCopy came with a scaled down version of the Readiris OCR software, the library purchased 3 copies of the full Readiris program for a total of $990; however, these were not used in the project because the 2 versions conflicted. Thus, the original estimate of $10,000 for equipment and software was too high.

**Labor**

Actual labor costs, as shown in Table 1, were $22,572 versus the estimated costs of $17,750. In the initial estimate, a temporary worker was assigned the task of adding the dissertation to the repository; however, 2 staff catalogers performed this work at a resulting higher rate. The $10,000 allotted for a temporary worker paid for quality control, OCR work, and editing of the signature pages for a total cost of $9,372. This labor cost would have been $6,446 if the extra step of editing the signature pages had not been necessary.

**Budget**

Total project costs were $23,562 ($990 software, $22,572 labor) or $0.28 per page (Table 1). This is $4,188 less than the original estimate of $27,750 to process the dissertations in-house and $1,062 more than the estimated cost to outsource the dissertations to UMI.

**Usage**

Historical circulation data from May 1999 through November 2007 show the library’s print dissertation collection was used 723 times. This is in stark contrast to the first 17 months the electronic collection was available (June 2006 through November 2007). Downloads of full-text PDF dissertations from eScholarship@UMMS totaled 17,555, with 10,497 originating from Google searches.
The usage statistics for this collection indicate that by disseminating the dissertations through eScholarship@UMMS, which is indexed by Google, access and use increased substantially. Studies indicate that individuals who publish their research online in addition to publishing in traditional scholarly venues are cited more often than those who rely solely on paper publications [24–27]. In digitizing the GSBS dissertations, the library has assisted in making the school’s research more widely available.

The team faced challenges such as workflow, cost concerns, policy development, and permissions. Communication and coordination between internal and external departments was vital and minimized errors. As the team learned, regardless of the amount of planning and thought that goes into a project, there is always the possibility that each record or file will need to be reworked. Decisions made in processing the dissertations set a precedent for future collections, such as adding documents without the full text if permission has not been obtained. The team acknowledged this could result in user frustration because they cannot get access to the full text. The team has worked hard to contact as many dissertation authors as possible to keep incomplete records to a minimum.

Nolen and Costanza described their experience in populating the repository at Trinity University, which
also focused on student works, by noting, “it’s important to start small, choosing projects that have usefulness to our constituents” [28]. The Lamar Soutter Library also found having a small, defined project had many benefits. It allowed the team to experience an early success and manage staff and resources by gradually incorporating the work. The team also gained experience with Digital Commons, metadata standards, and copyright. Additionally, this project served as a recruitment strategy to other campus departments through coordinated promotion by GSBS and the library for further population of the institutional repository. New materials recruited include student works, nursing dissertations, and faculty publications, a small portion of which required digitization.

For UMMS, digitizing dissertations proved to be a successful and cost-effective recruitment strategy and helped the library build stronger relationships at the medical school to secure future content. The team’s quick response to the dean’s privacy concerns built a foundation of trust for future work. Currently, all dissertations are submitted to the library in both print and electronic format along with a signed permission form to digitize the work. The library anticipates that building this relationship with students will make it easier to recruit future scholarly works over the life of a researcher’s career at the medical school.

REFERENCES


16. Lynch CA, Lippincott JK. Institutional deployment in the United States as of early 2005 [Internet]. D-Lib Mag [Internet]. 2005 Sep;11(9) [cited 1 Aug 2007]. [http://www.dlib.org/dlib/september05/lynch/09lynch.html].


26. Eysenbach G. Citation advantage of open access articles. PLOS Biol. 2006 May;4(5) [cited 26 Nov 2007]. DOI: 10.1371/journal.pbio.0040157.
27. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLOS One. 2007 Mar;2(3) [cited 26 Nov 2007]. DOI: 10.1371/journal.pone.0000308.

AUTHORS’ AFFILIATIONS

Mary Piorun, MLS, MBA, AHIP, mary.piorun@umassmed.edu; Associate Director for Technology Initiatives and Resource Management; Lisa A. Palmer, MLS, lisa.palmer@umassmed.edu, Catalog Librarian; Lamar Soutter Library, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655

Received October 2007; accepted December 2007