Digitizing Dissertations for the eScholarship@UMMS Institutional Repository

Lisa Palmer & Mary Piorun

Lamar Soutter Library
University of Massachusetts Medical School
Worcester, MA

http://library.umassmed.edu

MLA TSS Introduction to Library Digitization
Worcester, MA

October 28, 2008
Overview

- Background
  - 1st digitization project
  - Team members and roles
  - Choosing a repository system
  - Identifying manageable first project
- Project: digitizing 300 dissertations in-house 2006-2007
  - Partnership with one of our graduate schools
  - Metadata
  - Permissions process
  - Technical decisions
  - Workflow
  - Skills needed
  - Coordination between and within departments
- Ongoing operations
UMass Medical School

- Massachusetts’ only public medical school, founded in 1970
- Consistently ranked in top 10 in primary care education among 125 U.S. medical schools by U.S. News & World Report
- 990 students and growing
- School of Medicine, Graduate School of Nursing, Graduate School of Biomedical Sciences
- Clinical partner: UMass Memorial Health Care
- Separate graduate campus in UMass system
Lamar Soutter Library

- National Library of Medicine’s Regional Medical Library for the New England Region
- 210,000 volumes
- Journals: 1350 print subscriptions, 5000 subject-focused electronic subscriptions
- Special collections: rare books, history of medicine, humanities in medicine, consumer health, pediatrics, archives
- 42 Library FTE
- Medium-sized academic health sciences library
Team Charge

“Investigate institutional repository products and make a recommendation for the Medical School”
The Team

- Associate Director for Technology Initiatives and Resource Management (Project Leader): Project management, technology, workflow analysis, usability
- Associate Director for Education and Research Services: Outreach to faculty and students, copyright, training
- Catalog Librarians (2): Metadata, indexing, documentation, quality control, usability
System Evaluation

- Research: articles, discussion lists, library websites, users from other libraries, workshops, product demonstrations
- “Score card”
## The Score Card

<table>
<thead>
<tr>
<th>User Interface: 25 points</th>
<th>Cost: 10 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customizability</td>
<td>Initial cost</td>
</tr>
<tr>
<td>User friendliness</td>
<td>Annual maintenance fee</td>
</tr>
<tr>
<td>Searching/retrieval</td>
<td>Licensing fee</td>
</tr>
<tr>
<td>Submission process</td>
<td>Impact on staffing models</td>
</tr>
<tr>
<td>Navigation</td>
<td>Pricing model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tools: 30 points</th>
<th>Administration: 25 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-mail lists</td>
<td>Setup time</td>
</tr>
<tr>
<td>Faculty web pages</td>
<td>Statistical reporting</td>
</tr>
<tr>
<td>E-journal publication</td>
<td>Interoperability/compatibility</td>
</tr>
<tr>
<td>Alerting service</td>
<td>Maintenance interface</td>
</tr>
<tr>
<td>Controlled vocabulary lists</td>
<td>Long-term maintenance required</td>
</tr>
<tr>
<td>Data feeds</td>
<td>Accepted file formats</td>
</tr>
<tr>
<td>PDF conversion</td>
<td>Export of data</td>
</tr>
<tr>
<td>Ability to link related files</td>
<td>New staff skills required</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company/Community: 10 points</th>
<th>Branding/customizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer service/support</td>
<td>Training</td>
</tr>
<tr>
<td>User documentation</td>
<td>Access control</td>
</tr>
<tr>
<td>Company stability</td>
<td></td>
</tr>
<tr>
<td>Customer references</td>
<td></td>
</tr>
<tr>
<td>Number of product installations</td>
<td></td>
</tr>
<tr>
<td>Installed base</td>
<td></td>
</tr>
</tbody>
</table>
Digital Commons
http://www.bepress.com/ir/

- 2-year license purchased in January 2006
- Hosted
- Cool stuff: ability to link video & sound files, data sets
- OAI compliant
- Usage statistics, including monthly readership statistics emailed to authors
- Functionality that would make it easier to promote the repository: email alerts, “paper of the day”
- Faculty researcher pages, online journal publishing
Getting Started

- “eScholarship@UMMS”
  - http://escholarship.umassmed.edu
- Testing with Library staff publications:
  articles, presentations & posters
- Basic customizations to end-user and administrative interfaces
Pilot Project

Needed a manageable first project to

- Populate repository quickly
- Generate visibility
- Gain support
Dissertations

- Graduate School of Biomedical Sciences, founded in 1979
- Good demonstration project
  - GSBS Dean interested in project
  - Reasonable number (~300)
  - Already cataloged and had metadata
- Very few submitted in electronic form
- Not submitted to UMI
Preparing the System: Metadata

- “Supports export to XML Dublin Core (DC) format”
  - not exactly
  - Elements captured: title, creator, description, date, type, format, identifier, publisher, subject
  - Elements not captured: contributor, source, language, relation, coverage, rights
- Added new fields to dissertation template
  - DC elements not captured
  - Department (“UMMS Affiliation”)
  - ID number for bibliographic record in library catalog
RNA interference (RNAi) in the nematode Caenorhabditis elegans is a type of homology-dependent post-transcriptional gene silencing induced by dsRNA. This dissertation describes the genetic analysis of the RNA interference pathway and inheritance properties associated with this phenomenon. We demonstrate that the RNAi effect can be observed in the progeny of the injected animal for at least two generations. Transmission of the interference effect occurs through a dominant extragenic agent.
Preparing the System: Metadata (cont.)

- Added new document type:
  - Dissertation, Doctoral
- Activated live link functionality in Relation, Source, Comments fields
- Changed delimiter for subject field to accommodate MeSH

  **Before:**
  
  `<subject>Libraries</subject>`
  `<subject>Medical; Library Technical Services</subject>`

  **After:**
  
  `<subject>Libraries, Medical</subject>`
  `<subject>Library Technical Services</subject>`
Specified display order of fields for various displays

Data entry decisions

- Use Relation field to provide link to record in OPAC for print version of dissertation
- Use Rights field for information about copyright or permissions
- Comments field
Preparing the System: Metadata (cont.)

- How to re-utilize MARC data from online catalog
  - Small collection
  - Permission granted unevenly
  - Dismissed batch loader functionality
  - Decision: Copy & paste from OPAC; use macros where possible
Functions of the Cdc14-Family Phosphatase Clp1p in the Cell Cycle Regulation of *Schizosaccharomyces pombe*

by Susanne Trautmann
Metadata Skills

Description
Indexing
Authority control
Search and retrieval
Testing
Usability
Quality control
Documentation
Outsource?

- UMI digitization service - $22,500 (in 2006)
  - 2-3 month turn-around
  - Not full-text searchable

- In-house estimate
  - $27,750 (in 2006)
    - Two temporary employees
    - Equipment
    - Project management
    - 14-week turnaround
<table>
<thead>
<tr>
<th></th>
<th>Estimate (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning</td>
<td>45</td>
</tr>
<tr>
<td>Quality Control</td>
<td>45</td>
</tr>
<tr>
<td>OCR Abstract</td>
<td>20</td>
</tr>
<tr>
<td>Add to IR</td>
<td>20</td>
</tr>
<tr>
<td>Project Management</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>145</strong></td>
</tr>
</tbody>
</table>
Recommendation

- Process in-house
  - Gain experience
  - Retain access throughout
  - Tighter control
    - Project
    - Quality
Given

- $10,000 for temporary help
- Circulation staff
- ILL copier/scanners
Process

1. Obtain Permission
2. Scan Dissertation
3. Quality Control
4. Build a Table of Contents
5. Process Abstract
6. Add Dissertation to eScholarship
Permissions

- No process in place
  - Created two forms
    - Alumni
    - Current graduates
  - Forms approved by Legal department
- Contact 300 alumni
  - Access database
  - Local e-mail address
Permissions Cont ...

- 310 authors
  - 250 contacted
- 167 granted permission
- 67% success rate

Skills: Access, Word, Mail Merge, Writing, Searching, Political

Permissions - Scanning – Quality – ToC – Abstract – Adding
Scanning

- Who: Circulation staff
- What: 250 pages – single-sided
- When: Nights & weekends
- Where: ILL office

- Average: 2 per night, 5 on weekends
Scanning

Hardware
- Canon networked printer/copier/scanner
- Image Runner 3300 black & white
- Image Runner C3200 color

Software
- ECopy 3.1
- http://www.ecopy.com
Scanning

- Printout from OPAC
- Scan using eCopy
  - Break file up if called away
  - File is stored on the copier

**Skills:** Teamwork, Work Prioritization, Attention to Detail, Scanner Operation

Permissions - **Scanning** - Quality - ToC - Abstract - Adding
Quality Control

- Assemble the files
- Check for completeness
- Clean up edges
- Verify image quality
- Saving of file in various formats

Skills: Attention to Detail, eCopy, Scanner, Save As, File Management

Permissions - Scanning - Quality - ToC - Abstract - Adding
Table of Contents

- In PDF using bookmarks to build a ToC
  - Title
  - Signature
  - Abstract
  - Chapters
  - References

Skills: Adobe Acrobat Professional, Bookmarks

Permissions – Scanning – Quality – ToC – Abstract – Adding
Process Abstract

- eCopy
- OCR
- Notepad
- Cleanup
- HTML tagging
- Cataloging “In Box”

Skills: Attention to Detail, Proofreading, Basic HTML

Permissions – Scanning – Quality – ToC – Abstract – Adding
Review: 3 Files

- eCopy file for future use
- Searchable PDF
- HTML version of abstract
Add to eScholarship

Handoff to Cataloging
Add to eScholarship

- Step 1: Add record to eScholarship
  - Copy/paste from OPAC
  - Author, title, department, date, subjects
  - Document type, comments, abstract, link to OPAC, upload PDF
- Step 2: Add full-text link to record in online catalog
- Step 3: Move files to “Added to eScholarship” folder
- Step 4: Update alumni database

Skills: Cataloging, Organization, Multi-tasking, HTML, Access, Teamwork

Permissions – Scanning – Quality – ToC – Abstract – Adding
Decision

- No permission form on file
  - Scan dissertation and add record to eScholarship without full-text

- Pro: Process was working well, under budget
- Con: Adding records without the full text
Workflow without Permission

- Step 1: Add record to eScholarship
  - Copy/paste from OPAC
  - Author, title, department, date, subjects
  - Document type, comments, abstract, link to OPAC
- Step 2: Add comment: seeking permissions
- Step 3: Move abstract to “Added to eScholarship” folder
- Step 4: Update alumni database
When Permission is Obtained

- Step 1: In eScholarship, edit comments and upload PDF
- Step 2: In online catalog, add notes and full text link
- Step 3: Move PDF to “added” folder

**Results:** More steps, more coordination, higher risk of errors, user frustration
The Unexpected

- Privacy
  - Dean’s signature

- Signature Page
  - Dean, advisor, committee members

- Resolution
  - Create blank signature page
    - Adobe Illustrator
  - Replace files in eScholarship
Estimate vs. Actual Per Title

<table>
<thead>
<tr>
<th></th>
<th>Estimate (Minutes)</th>
<th>Actual (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Quality Control</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>OCR Abstract</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Add to IR</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Signature Page</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Replace File</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Project Management</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>145</td>
<td>170</td>
</tr>
</tbody>
</table>

Under estimated
Over estimated
## Estimates vs. Actual Project

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th></th>
<th></th>
<th>Actual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
<td>Cost ($)</td>
<td>Hours</td>
<td>Cost ($)</td>
<td></td>
</tr>
<tr>
<td>Scanning</td>
<td>225</td>
<td>4,500</td>
<td>240</td>
<td>4,800</td>
<td></td>
</tr>
<tr>
<td>Quality Control</td>
<td>225</td>
<td>5,625</td>
<td>133</td>
<td>2,926</td>
<td></td>
</tr>
<tr>
<td>OCR Abstract</td>
<td>100</td>
<td>2,500</td>
<td>160</td>
<td>3,520</td>
<td></td>
</tr>
<tr>
<td>Add to IR</td>
<td>100</td>
<td>2,500</td>
<td>54</td>
<td>1,890</td>
<td></td>
</tr>
<tr>
<td>Signature Page</td>
<td></td>
<td></td>
<td>133</td>
<td>2,926</td>
<td></td>
</tr>
<tr>
<td>Replace File</td>
<td></td>
<td></td>
<td>26</td>
<td>910</td>
<td></td>
</tr>
<tr>
<td>Project Management</td>
<td>75</td>
<td>2,625</td>
<td>160</td>
<td>5,600</td>
<td></td>
</tr>
<tr>
<td>Hardware / Software</td>
<td>10,000</td>
<td>990</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>725</td>
<td>$27,750</td>
<td>903</td>
<td>$23,562</td>
<td></td>
</tr>
</tbody>
</table>
## Outsourced Costs

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
</tr>
<tr>
<td>Quality Control</td>
<td>133</td>
</tr>
<tr>
<td>OCR Abstract</td>
<td>160</td>
</tr>
<tr>
<td>Signature Page</td>
<td>133</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
</tr>
</tbody>
</table>


Visio in Your Handout
Long Term Coordination

- Between the Library and the Graduate School on the dissertation approval process
- Between Cataloging and Systems
Project Evaluation

- **Budget**
- **Usage statistics as of 10-26-2008**
  - 382 dissertations submitted
  - 33,385 full text downloads
  - One 2005 dissertation on dengue fever has been downloaded over 800 times
- **Visibility**
Future Directions

- Administrative
  - Manage copyright issues
  - Create a marketing and promotion plan
- Self-submission of dissertations
- Content recruitment
  - Specialized student scholar programs
  - Open access journal articles
  - GSBS student publications
Conclusion

- Success factors
  - Library funding, support, management, skills
  - Buying a hosted product
  - Support of Graduate School Dean
  - Small, defined project

- Future success
  - Continued funding
  - Dedicated repository staff
  - Increased faculty and department participation
  - Greater campus awareness
Thank You!!

Mary Piorun, MSLS, MBA, AHIP
Associate Director
Technology Initiatives and Resource Management
508-856-2206 – Mary.Piorun@umassmed.edu

Lisa Palmer, MSLS
Catalog Librarian
508-856-4368 - Lisa.Palmer@umassmed.edu

Presentation slides and handouts:
http://escholarship.umassmed.edu/lib_articles/96/

Detailed journal article about this project:
http://escholarship.umassmed.edu/lib_articles/94/