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

User Interface: 25 points
- Customizability
- User friendliness
- Searching/retrieval
- Submission process
- Navigation

Tools: 30 points
- E-mail lists
- Faculty web pages
- E-journal publication
- Alerting service
- Controlled vocabulary lists
- Data feeds
- PDF conversion
- Ability to link related files

Company/Community: 10 points
- Customer service/support
- User documentation
- Company stability
- Customer references
- Number of product installations
- Installed base

Cost: 10 points
- Initial cost
- Annual maintenance fee
- Licensing fee
- Impact on staffing models
- Pricing model

Administration: 25 points
- Setup time
- Statistical reporting
- Interoperability/compatibility
- Maintenance interface
- Long-term maintenance required
- Accepted file formats
- Export of data
- New staff skills required
- Branding/customizing
- Training
- Access control

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

Sample XML Dublin Core View

<?xml version="1.0" encoding="utf-8" ?>
<dc>
<dc-record>
<title>Analysis of RNA Interference in ****C. elegans****: A Dissertation**</title>**
<creator>Grishok, Alla**</creator>**
<description>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....**</description>**
<date>2001-09-27**</date>**
<type>text**</type>**
<format>application/pdf**</format>**
<identifier>http://escholarship.umassmed.edu/gsbs_diss/139**</identifier>**
<publisher>eScholarship@UMMS**</publisher>**
<subject>Caenorhabditis elegans**</subject>**
<subject>RNA Interference**</subject>**
<subject>RNA, Small Interfering**</subject>**
<subject>Academic Dissertations**</subject>**
<subject>Dissertations, UMMS**</subject>**
</dc-record>
</dc>

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>

Preparing the System: Metadata (cont.)

- 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

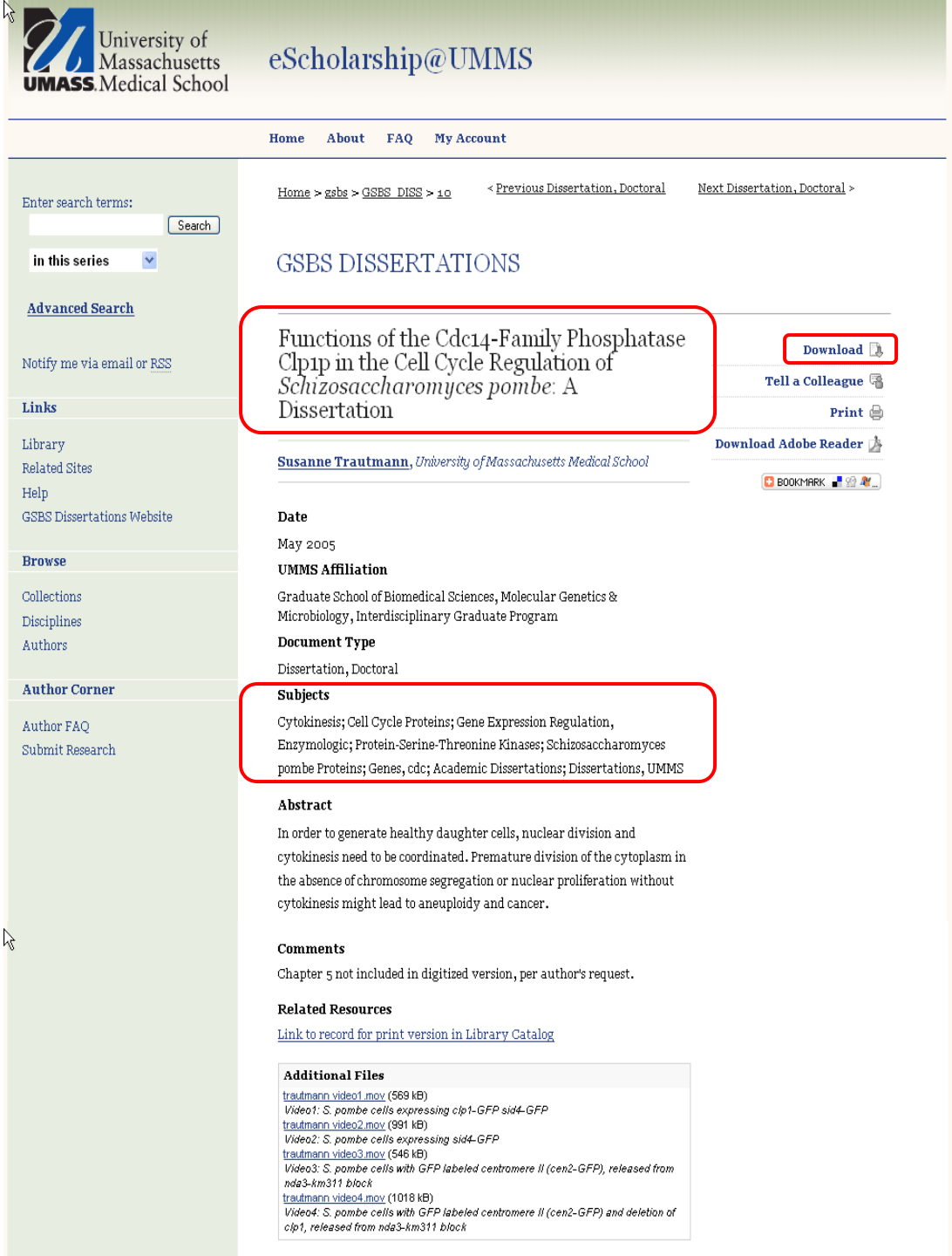
Sample Record

Functions of the Cdc14-Family Phosphatase Clp1p in the Cell Cycle Regulation of *Schizosaccharomyces pombe*
by Susanne Trautmann

Comments →

Relation →

Video Files →



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

Estimates Per Title

	Estimate (Minutes)
Scanning	45
Quality Control	45
OCR Abstract	20
Add to IR	20
Project Management	15
Total	**145**

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

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

	Estimate (Minutes)	Actual (Minutes)
Scanning	45	45
Quality Control	45	25
OCR Abstract	20	30
Add to IR	20	10
Signature Page		25
Replace File		5
Project Management	15	30
Total	**145**	**170**

Under estimated
Over estimated

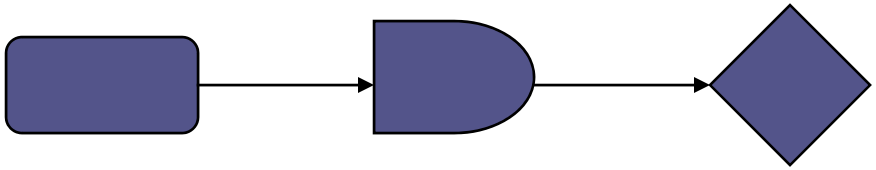
Estimates vs. Actual Project

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

Outsourced Costs

	Actual	
	Hours	**Cost ($)**
Quality Control	133	2,926
OCR Abstract	160	3,520
Signature Page	133	2,926
Total		**9,372**

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/