Anatomy of a Digitization Project:
Dissecting the Process

Mary Piorun & Lisa Palmer

Lamar Soutter Library
University of Massachusetts Medical School
Worcester, MA

http://library.umassmed.edu

Scanning Forum 2006
Charlottesville, VA

November 6, 2006

Overview

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

UMass Medical School

- Massachusetts' only public medical school, founded in 1970
- Currently ranked 4th in primary care education among 125 U.S. medical schools by U.S. News & World Report
- 950 students
- School of Medicine, Graduate School of Nursing, Graduate School of Biomedical Sciences
- Clinical partner: UMass Memorial Health Care
- Workplace of 2006 Nobel Prize co-recipient for Medicine or Physiology, Dr. Craig Mello
- Separate graduate campus in UMass system

Lamar Soutter Library

- NLM Regional Medical Library for New England Region
- 235,000 volumes
- Journal subscriptions: 1470 print, 4200 electronic
- Special collections: rare books, history of medicine, consumer health, early childhood, pediatrics
- 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 Systems (Project Leader): Project management, technology, usability
- Associate Director for Research, Education and Information 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

ProQuest Digital Commons (http://umi.com/products_umi/digitalcommons/)

- 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
  - Populate repository quickly
  - Generate visibility
  - Gain support
- Failed opportunities: Massachusetts state Medicaid brochures; UMass Board of Trustees meeting minutes

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 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
  - 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 -- enhances keyword access with MeSH and/or LC subject terms

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 views: administrative, user input, output (end user interface)
- Data entry decisions
  - Use Relation element to provide link to record in OPAC for print version of dissertation
  - Use Rights element 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

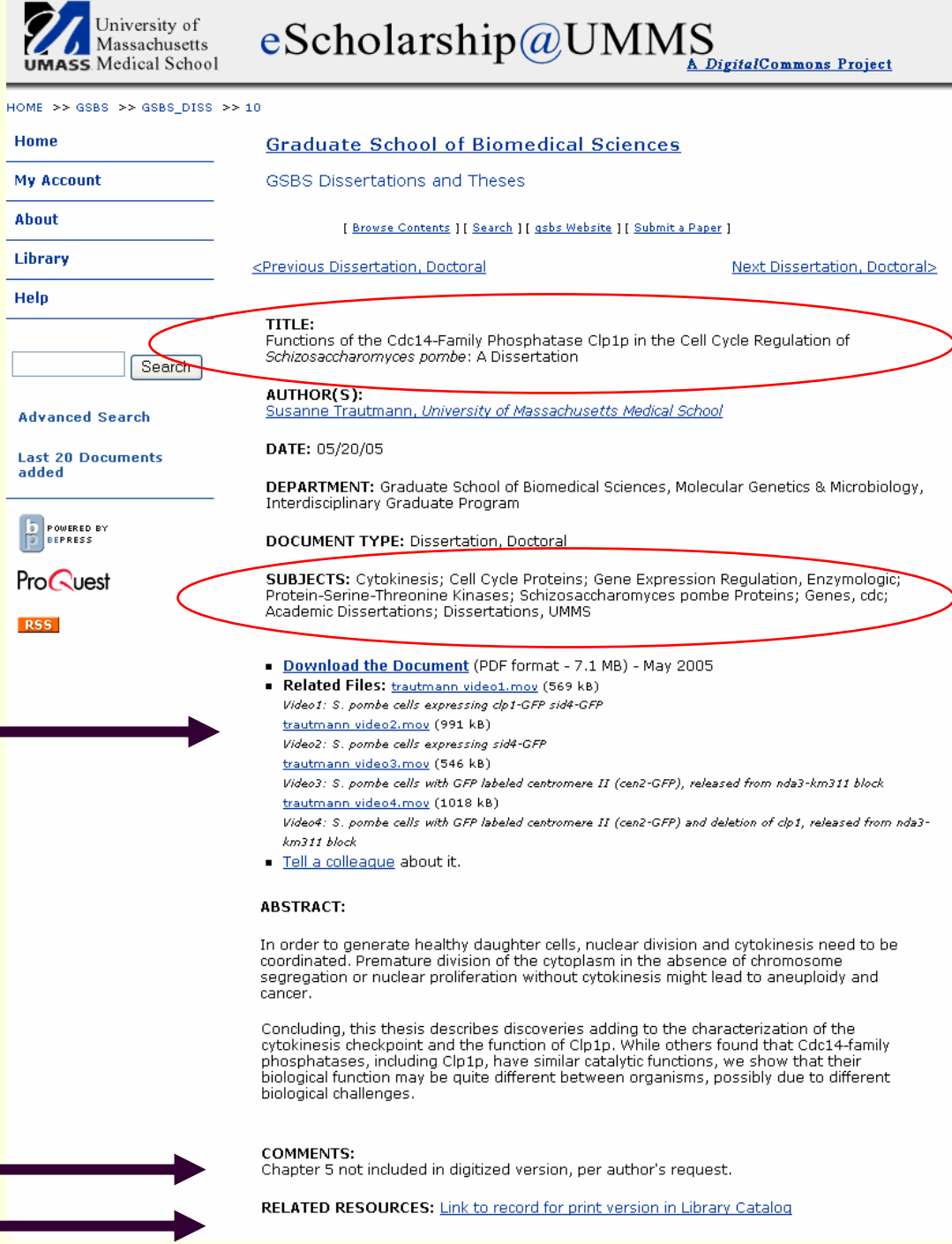
Sample Record

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

Full Text & Video →

Comments →

Relation →



University of Massachusetts Medical School
UMASS
eScholarship@UMMS
A DigitalCommons Project

HOME >> GSBS >> GSBS_DISS >> 10

Home
My Account
About
Library
Help

Search

Advanced Search

Last 20 Documents added

POWERED BY BEPRESS
ProQuest
RSS

Graduate School of Biomedical Sciences

GSBS Dissertations and Theses

[Browse Contents] [Search] [gsbs Website] [Submit a Paper]

<Previous Dissertation, Doctoral Next Dissertation, Doctoral>

TITLE:
Functions of the Cdc14-Family Phosphatase Clp1p in the Cell Cycle Regulation of *Schizosaccharomyces pombe*: A Dissertation

AUTHOR(S):
Susanne Trautmann, *University of Massachusetts Medical School*

DATE: 05/20/05

DEPARTMENT: Graduate School of Biomedical Sciences, Molecular Genetics & Microbiology, Interdisciplinary Graduate Program

DOCUMENT TYPE: Dissertation, Doctoral

SUBJECTS: Cytokinesis; Cell Cycle Proteins; Gene Expression Regulation, Enzymologic; Protein-Serine-Threonine Kinases; Schizosaccharomyces pombe Proteins; Genes, cdc; Academic Dissertations; Dissertations, UMMS

- **Download the Document** (PDF format - 7.1 MB) - May 2005
- **Related Files:** trautmann video1.mov (569 kB)
  Video1: S. pombe cells expressing clp1-GFP sid4-GFP
  trautmann video2.mov (991 kB)
  Video2: S. pombe cells expressing sid4-GFP
  trautmann video3.mov (546 kB)
  Video3: S. pombe cells with GFP labeled centromere II (cen2-GFP), released from nda3-km311 block
  trautmann video4.mov (1018 kB)
  Video4: S. pombe cells with GFP labeled centromere II (cen2-GFP) and deletion of clp1, released from nda3-km311 block
- Tell a colleague about it.

ABSTRACT:

In order to generate healthy daughter cells, nuclear division and cytokinesis need to be coordinated. Premature division of the cytoplasm in the absence of chromosome segregation or nuclear proliferation without cytokinesis might lead to aneuploidy and cancer.

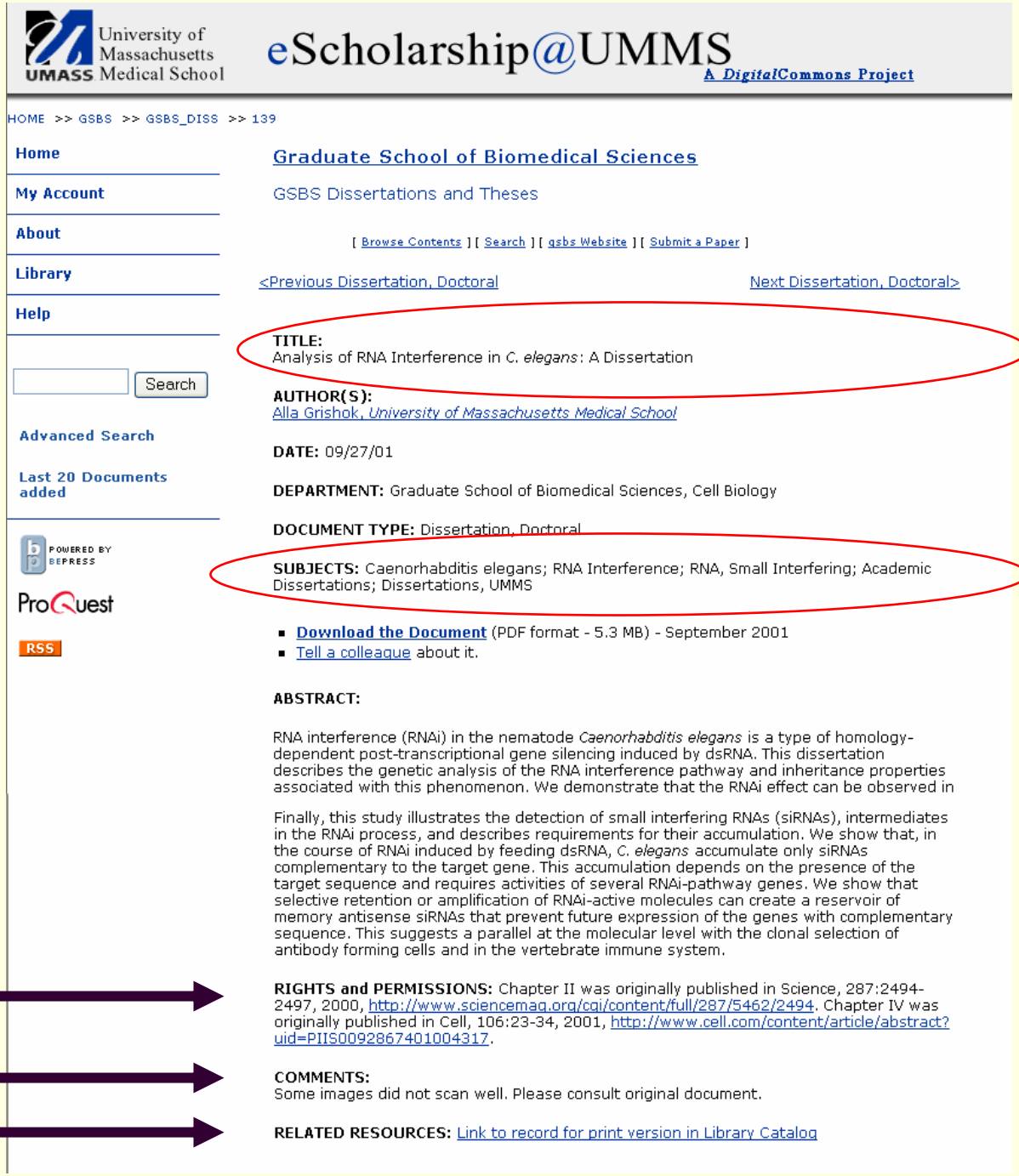
Concluding, this thesis describes discoveries adding to the characterization of the cytokinesis checkpoint and the function of Clp1p. While others found that Cdc14-family phosphatases, including Clp1p, have similar catalytic functions, we show that their biological function may be quite different between organisms, possibly due to different biological challenges.

COMMENTS:
Chapter 5 not included in digitized version, per author's request.

RELATED RESOURCES: Link to record for print version in Library Catalog

Sample Record

Analysis of RNA Interference in *C. elegans* by Alla Grishok



University of Massachusetts Medical School
UMASS
eScholarship@UMMS
A DigitalCommons Project

HOME >> GSBS >> GSBS_DISS >> 139

Home
My Account
About
Library
Help

Search
Advanced Search
Last 20 Documents added

POWERED BY BEPRESS
ProQuest
RSS

Graduate School of Biomedical Sciences

GSBS Dissertations and Theses

[Browse Contents] [Search] [gsbs Website] [Submit a Paper]

<Previous Dissertation, Doctoral Next Dissertation, Doctoral>

TITLE:
Analysis of RNA Interference in *C. elegans*: A Dissertation

AUTHOR(S):
Alla Grishok, *University of Massachusetts Medical School*

DATE: 09/27/01

DEPARTMENT: Graduate School of Biomedical Sciences, Cell Biology

DOCUMENT TYPE: Dissertation, Doctoral

SUBJECTS: Caenorhabditis elegans; RNA Interference; RNA, Small Interfering; Academic Dissertations; Dissertations, UMMS

- **Download the Document** (PDF format - 5.3 MB) - September 2001
- Tell a colleague about it.

ABSTRACT:

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

Finally, this study illustrates the detection of small interfering RNAs (siRNAs), intermediates in the RNAi process, and describes requirements for their accumulation. We show that, in the course of RNAi induced by feeding dsRNA, *C. elegans* accumulate only siRNAs complementary to the target gene. This accumulation depends on the presence of the target sequence and requires activities of several RNAi-pathway genes. We show that selective retention or amplification of RNAi-active molecules can create a reservoir of memory antisense siRNAs that prevent future expression of the genes with complementary sequence. This suggests a parallel at the molecular level with the clonal selection of antibody forming cells and in the vertebrate immune system.

RIGHTS and PERMISSIONS: Chapter II was originally published in Science, 287:2494-2497, 2000, http://www.sciencemag.org/cgi/content/full/287/5462/2494. Chapter IV was originally published in Cell, 106:23-34, 2001, http://www.cell.com/content/article/abstract?uid=PIIS0092867401004317.

COMMENTS:
Some images did not scan well. Please consult original document.

RELATED RESOURCES: Link to record for print version in Library Catalog

Rights and Permissions

Comments

Relation

More on Metadata

- Currently catalogers handle submissions
- Odd problems
  - Some fields do not display if the record has no abstract

<u>Skills</u>

Description

Indexing

Authority control

Search and retrieval

Testing

Usability

Quality control

Documentation

Outsource?

- UMI digitization service - $22,500
  - 2-3 month turn-around
  - Not full-text searchable
- In-house estimate
  - $29,820
    - 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

Permissions Cont...

Contact Method

Graduate School	**10**
E-mail	**160**
Mail	**165**
Both e-mail & mail	**75**
No contact	**50**

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

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
Project Management	15	30
Total	**145**	**140**

Estimate vs. Actual Per Title Underestimated

	Estimate (Minutes)	Actual (Minutes)
Scanning	45	45
Quality Control	45	25
OCR Abstract	**20**	**30**
Add to IR	20	10
Project Management	**15**	**30**
Total	**145**	**140**

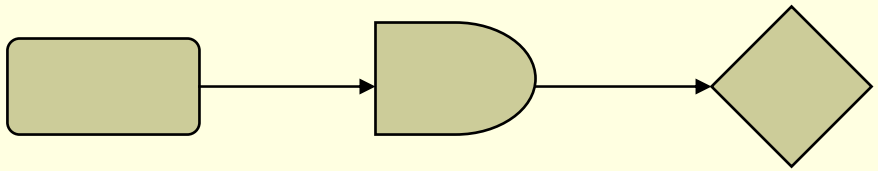
Estimate vs. Actual Per Title Overestimated

	Estimate (Minutes)	Actual (Minutes)
Scanning	45	45
Quality Control	**45**	**25**
OCR Abstract	20	30
Add to IR	**20**	**10**
Project Management	15	30
Total	**145**	**140**

Estimates vs. Actual Project

	Estimate (May)		Projected (Nov)	
	Hours	**Cost**	**Hours**	**Cost**
Scanning	250	5,000	250	
Quality Control	250	6,250	125	2,750
OCR Abstract	120	3,000	150	3,300
Add to IR	120	3,000	60	1,500
Project Management	70	2,450	140	4,900
Total		19,700		**12,450**

Visio in Your Handout



Long Term Coordination

- Become part of the dissertation approval process
- Cataloging and Systems

Evaluation

- Budget
- Usage statistics
  - 169 dissertations submitted (most in full-text)
  - 2122 full text downloads since 5-30-2006
  - One 2005 dissertation on dengue fever has been downloaded 144 times
- Visibility

Future Directions

- Administrative
  - Document policies and procedures
  - Manage copyright issues
  - Create a marketing and promotion plan
- Content recruitment
  - Graduate School of Nursing dissertations
  - Specialized student scholar groups
  - Open access journal articles

Conclusion

- Success factors to date
  - Library funding, support, management, skills
  - Buying a hosted product
  - Support of Graduate School Dean
- Future success
  - Continued funding
  - Dedicated repository staff
  - Increased faculty and department participation
  - Greater campus awareness

Thank You!!

Mary Piorun, MSLS, AHIP
Associate Director, Systems
508-856-2206 – Mary.Piorun@umassmed.edu

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

Presentation URL:
http://escholarship.umassmed.edu/lib_postpres/23/