Anatomy of a Digitization Project: Dissecting the Process

Mary E. Piorun
University of Massachusetts Medical School

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/lib_postpres

Repository Citation

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.
This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Library Posters and Presentations by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
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

<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></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Customer service/support</td>
<td></td>
</tr>
<tr>
<td>- User documentation</td>
<td></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>
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
Analysis of RNA Interference in *C. elegans*: A Dissertation

Grishok, Alla

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.}

2001-09-27

text

application/pdf

http://escholarship.umassmed.edu/gsbs_diss/139

eScholarship@UMMS

Caenorhabditis elegans

RNA Interference

RNA, Small Interfering

Academic Dissertations

Dissertations, UMMS
Prepating 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
Functions of the Cdc14-Family Phosphatase Clp1p in the Cell Cycle Regulation of *Schizosaccharomyces pombe*

by Susanne Trautmann

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
Analysis of RNA Interference in C. elegans by Alla Grishok

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 the course of RNA induced by feeding dsRNA. C. elegans accumulate only siRNAs complementary to the target genes. This accumulation depends on the presence of the target sequence and requires activities of several RNA-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.


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

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

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

Skills
- 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

<table>
<thead>
<tr>
<th>Service</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
Permissions Cont...

Contact Method

<table>
<thead>
<tr>
<th>Contact Method</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate School</td>
<td>10</td>
</tr>
<tr>
<td>E-mail</td>
<td>160</td>
</tr>
<tr>
<td>Mail</td>
<td>165</td>
</tr>
<tr>
<td>Both e-mail &amp; mail</td>
<td>75</td>
</tr>
<tr>
<td>No contact</td>
<td>50</td>
</tr>
</tbody>
</table>

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
## Estimate vs. Actual Per Title

<table>
<thead>
<tr>
<th>Task</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>Project Management</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>145</strong></td>
<td><strong>140</strong></td>
</tr>
<tr>
<td></td>
<td>Estimate (Minutes)</td>
<td>Actual (Minutes)</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------</td>
<td>------------------</td>
</tr>
<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>Project Management</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>145</td>
<td>140</td>
</tr>
</tbody>
</table>
## Estimate vs. Actual Per Title

Overestimated

<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>Project Management</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>145</strong></td>
<td><strong>140</strong></td>
</tr>
</tbody>
</table>
## Estimates vs. Actual Project

<table>
<thead>
<tr>
<th></th>
<th>Estimate (May)</th>
<th></th>
<th>Projected (Nov)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours</td>
<td>Cost</td>
<td>Hours</td>
<td>Cost</td>
</tr>
<tr>
<td>Scanning</td>
<td>250</td>
<td>5,000</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>Quality Control</td>
<td>250</td>
<td>6,250</td>
<td>125</td>
<td>2,750</td>
</tr>
<tr>
<td>OCR Abstract</td>
<td>120</td>
<td>3,000</td>
<td>150</td>
<td>3,300</td>
</tr>
<tr>
<td>Add to IR</td>
<td>120</td>
<td>3,000</td>
<td>60</td>
<td>1,500</td>
</tr>
<tr>
<td>Project Management</td>
<td>70</td>
<td>2,450</td>
<td>140</td>
<td>4,900</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>19,700</strong></td>
<td><strong>12,450</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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/