ETD Lifecycle Management

USetdA 2013
Pre-Conference Workshop
Workshop Overview

9:15am-10am  ETD Lifecycle Management - Interactive Overview

10am-10:45am  Guidance Documents for Lifecycle Management – Overview & Usage

10:45am-11am  Break

11am-11:45am  ETD Lifecycle Management – Overview & Demo

11:45am-12pm  Wrap-Up & Survey
Project Background

- IMLS-funded *Lifecycle Management of ETDs* project (2011-2014)
  - Documenting lifecycle curation practices for ETDs and improving implementations for curation tools
    - Guidance Documents
    - Workshop & Training Materials
    - Lifecycle Management Tools

[http://metaarchive.org/imls](http://metaarchive.org/imls)
Addressing Key Challenges

• The move to electronic theses & dissertations greatly enhances the accessibility & sharing of graduate student research but raises grave concerns about the long-term sustainability of these digital resources:
  – Can we ensure that ETDs acquired from students today will be available to future researchers? In 10 years? In a century?
  – How will libraries identify and institutionalize the best long-term curatorial practices for this important genre of digital content?
  – How will institutions address the entire lifecycle of ETDs?
ETD Lifecycle Management

Interactive Overview
Learning Outcomes

1. Attendees will gain a clear understanding of information lifecycle management models and how they apply to ETD programs

2. Attendees will understand how their current ETD program responsibilities fit within a larger lifecycle for managing ETDs

3. Attendees will understand how they can better coordinate with other ETD program stakeholders at key stages of curation activity
Lifecycle Management for ETDs

- ETD programs provide policies, workflows, and services around crucial lifecycle curation functions:
  - creation guidelines
  - deposit/submission
  - documenting approvals
  - metadata capture
  - rights management
  - ingest into commercial and/or library-based repositories
  - cataloging, access and usage monitoring
Lifecycle Management for ETDs

Federal Law 44 U.S.C. 2901
ISO 15489

DCC Curation Lifecycle Model
Lifecycle Management for ETDs

http://www.dcc.ac.uk/resources/curation-lifecycle-model
ETD Lifecycle Workflows

Create & Receive
- Submission System
- Author
- ETD
- Metadata
- Supplemental Files

Appraise & Select
- Graduate School/IT

Manage
- Libr ary/IT

Store, Access & Re-Use
- Institutional Repository
- ProQuest/Vendors
- ProQuest/Vendors
ETD Stakeholders on the Lifecycle

- **Student authors** create & submit ETDs
- **Graduate Schools** process, approve & embargo ETDs
- **Libraries/IT/Vendors** catalog & archive ETDs
- **Graduate Schools & Libraries/Vendors** also update ETDs
- **ETD Program Planners and Implementers** and **Libraries/IT/Vendors** disseminate ETDs
Interactive Exercise

• What Phase is your ETD Program/Service in right now?
  – Planning
  – Development
  – Testing
  – Implementation
  – Assessment/Improvement

• What questions/issues are you grappling with?

• Where do these questions/issues fall on the lifecycle?
Lifecycle Management for ETDs
ETD Lifecycle Workflows

Create & Receive

Appraise & Select

Manage

Store, Access & Re-Use

Submission System

Graduate School/IT

Library/IT

Institutional Repository

ProQuest/Vendors

Author

ETD Metadata Supplemental Files

ProQuest/Vendors
Questions?
ETD Guidance Documents

Overview & Usage
Learning Outcomes

1. Attendees will gain an understanding of the full range of lifecycle management activities for ETDs

2. Attendees will understand how each Guidance document is most relevant to various ETD program stakeholders and lifecycle management stages

3. Attendees will understand how to make use of the documents in practical ways depending on their stage of ETD program planning or implementation
Guidance Documents for ETD Lifecycle Management

1. Guidelines for Implementing ETD Programs - Roles & Responsibilities
2. Guide to Access Levels and Embargoes of ETDs
3. Briefing on Copyright Issues and Fair Use in ETDs
4. Guidelines for Collecting Usage Metrics & Demonstrations of Value for ETD Programs
5. Overview of Formats, Complex Content Objects, and Format Migration Scenarios for ETDs
6. Metadata for ETD Lifecycle Management
8. Guide to Options for ETD Programs
Guidelines for Implementing ETD Programs – Roles & Responsibilities

- Implementing an ETD program requires identifying the various stakeholders and specifying their roles and responsibilities throughout the entire course of ETD management.

- **Internal Stakeholders:** Institutional Administrators, Graduate Schools, Academic Libraries, Offices of Information Technology

- **External Stakeholders:** Commercial Companies, ETD Organizations, Library Consortia, Access Harvesters/Facilitators, Digital Repository Services, and Digital Preservation Services
Guidelines for Implementing ETD Programs – Roles & Responsibilities

• ETD Program Planning
  – Provide a rationale for establishing an ETD program
  – Advocate the program
  – Propose an implementation plan

• ETD Creation, Submission and ingestion
  – Construct, format and submit ETDs, may make embargo requests
  – Develop submission policies, procedures and standards
  – Administer the process, review and approve ETDs
  – Offer assistance including legal services
  – Prepare submission systems
  – Catalog and ingest ETDs, digitize retrospective theses and dissertations

• ETD access
  – Develop ETD access policies and ETD end user license
  – ETD access management
Guidelines for Implementing ETD Programs – Roles & Responsibilities

• ETD access (cont.)
  – Distribute ETDs in multiple access venues
  – Assist with ETD visibility, accessibility, and searchability

• ETD Archiving and Preservation
  – Develop a formal preservation plan
  – Organize and preserve in reliable media or systems
  – Preserve ETD contents, format, metadata and URLs

• ETD program evaluation and assessment
  – Evaluate ETD submission, support services, and program impact
  – Evaluate ETD cataloging, archiving and preservation practices
  – Evaluate ETD accessibility and usability
  – Evaluate ETD systems: operational, sustainable, and viable
Guide to Access Levels & Embargoes of ETDs

• One of the most contested topics in ETD program planning is the question of ETD embargoes and levels of access restriction
  – As evidenced by both the NDLTD/MetaArchive surveys and the 2010 CNI survey.

• An “embargo” of an ETD means delaying public access to the ETD, either temporarily or permanently.
Guide to Access Levels & Embargoes of ETDs

• Different stakeholders are particularly concerned about embargoes.
  – For instance, does depositing an ETD in a public repository constitute publication and hinder future development/publication of the work?
    • Publishers do not think so. (NDLTD survey)
  – The US Patent Office might consider an ETD to be prior art.

• Policy options for the embargo of ETDs range:
  – No embargo
  – Blanket fixed-length embargo
  – Limited list of fixed-length embargo
  – Embargo renewal
Briefing on Copyright Issues & Fair Use in ETDs

• ETD program may introduce students to issues of copyright
  – Author rights
  – Types of licenses
  – Fair use
  – Commercial publishers

• An ETD program does disservice to both students and institution if it does not provide information to make informed decisions on copyright and fair use.

• Who on the campus can provide guidance about ETD copyright and fair use?
Briefing on Copyright Issues & Fair Use in ETDs

• ETD programs must know
  – What understanding or agreement is in place at your institution regarding intellectual property rights of students?
    • What rights does the university/college exert on student work?
    • When does research belong to the university, and when does it not?
  – Does sponsored research require students to sign University Intellectual Property (IP) agreement or embargo their results?
  – What qualifies as plagiarism?
  – When does publication violate copyright agreements of work reproduced in the ETD?
  – What should students know regarding ETDs and future scholarly publishing efforts?
Guidelines for Collecting Usage Metrics & Demonstrations of Value

- Libraries have a long history of evaluating and studying use of library resources and collections, and ETDs should be no exception.

- Usage data can make a strong case for ETD program support to university administrations.

- Usage reports of all kinds should be prominently featured on the ETD program website, and easily reviewed by all users of the service.
Guidelines for Collecting Usage Metrics & Demonstrations of Value

• Quantitative statistics
  – Easily gathered
  – Frequently used
  – Provide compelling indication of utility, e.g. download statistics
  – Analysis can be performed on the data. e.g. demographics, behavior (new, returning users), technology, and devices.

• Qualitative evaluation
  – Case studies, surveys, interviews, and visual observations
  – Less commonly performed
  – Provide more nuanced information.

• Recommendation
  – Minimal level
  – Advanced Level: Gathering more data, but requiring more resources.
Guidelines for Collecting Usage Metrics & Demonstrations of Value

- Benefits of Usage Metrics
- Authors, faculty and graduate students
  - Impacts and importance of a title
  - Impacts of their ETDs
- Institution
  - Impacts of the ETD collection
  - Return on Investment (ROI)
  - Satisfaction of graduate students
- Scholarly Society
  - Impacts of ETDs vs. other content
Overall, many ETD programs mandate that the primary item deposited be some form of PDF, sometimes with format checking of the specific characteristics of the PDF.

ETDs can contain non-textual supplementary files:
- Music
- 3D Renderings
- Datasets
- For these files, some flexibility is necessary, but the institution should also provide guidelines.

The data and preferred file formats won’t remain stable over time.
Overview of Formats, Complex Content Objects, and Format Migration Scenarios for ETDs

- Upon deposit, both primary and supplementary files should be checked for:
  - Format validity
  - Viruses
  - Fixity

- Format migrations are anticipated by many ETD repositories. Depending on the migration policy, they may be:
  - Manually batched
  - Automated
Metadata for ETD Lifecycle Management

• Metadata structures much of the ongoing management of ETDs.
• Metadata may be assigned by a mix of: Librarians, Student authors, system, even users.
  • No matter the mix, professionals supervision provides quality control.

• The ETD repository software heavily influences the creation of metadata, but metadata should be thorough and comprehensive.
Metadata for ETD Lifecycle Management

• NDLTD has developed and maintains an ETD specific descriptive metadata schema.

• **PREMIS stands for "PREservation Metadata: Implementation Strategies."
  – The PREMIS data model consists of five interrelated entities: Intellectual, Object, Event, Agent, and Rights.
  – It allows repositories to implement varying workflow and submission models. In light of tracking transitions in the lifecycle of digital objects, this project is experimenting with updating ETD records.)

• An effective ETD metadata management approach can help institutions improve consistency, clarity of data lineage, and relationships so that they can better integrate related resources and ensure long-term access to ETDs.
Guide to ETD Program Planning & Cost Estimation

• Planning an ETD program depends on a number of decisions with varying costs.
  – Staffing
  – Repository platform
  – Equipment main
  – Preservation Strategy
  – Outsourcing

• Prepare a careful plan involving a full range of stakeholders. There will be unexpected changes, but it gives you a place to start.
Guide to ETD Program Planning & Cost Estimation

• Case Study Interviews
  – Large and Small
  – Public and Private
  – Rural and Urban

  – Portland State University
  – Rice University
  – University of Arizona
  – University of North Texas
  – Virginia Tech

  – ETD program personnel?
  – Key technologies?
  – Resource allocation?
  – Etc.
Guide to Options for ETD Programs

- Benefits & Impacts of ETD Programs
  - Moving past debates and to implementations

- Information Resources
  - Lifecycle Management Project Bibliography
  - NDLTD

- Access Policies & Intellectual Property Issues
  - Copyright violations & disputes
  - Access violations
  - Quality control
  - Embargo issues

- Deposit Procedures
  - Mandatory vs. optional
  - File formats
Guide to Options for ETD Programs

• Incorporating ETDs into larger institutional repositories
  – Engages economies of scale
  – Reduces focus on ETD specific issues

• Partnering with other institutions or alliances
  – Improves access to information
  – Shares access to digital preservation resources

• Completely outsourcing to external vendors
  – Reduces work required locally
  – Relinquishes some control to the vendor
ETD Stakeholders on the Lifecycle

Student authors create & submit ETDs

Graduate Schools & Libraries/Vendors also update ETDs

Libraries/IT/Vendors disseminate ETDs

Libraries/IT/Vendors catalog & archive ETDs

ETD Program Planners and Implementers

Submission System

Library/IT

Institutional Repository

ProQuest/Vendors
1. Guidelines for Implementing ETD Programs
2. Guide to Access Levels and Embargoes of ETDs
3. Briefing on ETD Copyright Issues and Fair Use in ETDs
4. Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs
5. Overview of Formats, Complex Content Objects, and Format Migration Scenarios for ETDs
6. Metadata for ETD Lifecycle Management
7. Guide to ETD Program Planning and Cost Estimation
8. Guide to Options for ETD Programs
Guidelines for Implementing ETD Programs

1. Roles & Responsibilities
2. Guide to Access Levels and Embargoes of ETDs
3. Briefing on ETD Copyright Issues and Fair Use in ETDs
4. Guidelines for Collecting Usage Metrics and Demonstrations of Value for ETD Programs
5. Overview of Formats, Complex Content Objects, and Format Migration Scenarios for ETDs
6. Metadata for ETD Lifecycle Management
7. Guide to ETD Program Planning and Cost Estimation
8. Guide to Options for ETD Programs

Lifecycle Guidance

- Conceptualise
- Create or Receive
- Transform
- Appraise & Select
- Preserve
- Ingest
- Store
- Pandemic Action Preservation
- Reappraise
- Disposal
How to Use These Documents

1. Share them with your ETD Program Manager(s)
2. Coordinate a targeted document review with individual ETD Program stakeholders
3. Organize a brown-bag lunch discussion w/ ETD Program stakeholders
4. Document recommendations & next steps for program implementation or improvement
Providing Feedback

• Please take the next month to read one or more documents
• You can agree to formal review on Question 9 of the Post Workshop Survey sheet
• Make use of the review sheet for the *Guidance Documents for Lifecycle Management*
• We will follow-up in 1 month by email to gather feedback
Questions?
Break
Lifecycle Management Tools

Overview & Demonstration
Learning Outcomes

1. Attendees will gain an understanding of the importance of modular curation for ETDs

2. Attendees will gain an understanding of how to begin making use of the project’s documented modular curation tools

3. Attendees will leave with a set of demonstrated examples of the tools in order to encourage their uptake and implementation
Exploring Modular Curation for ETDs

The project has aimed to develop and disseminate a set of software tools to address specific needs in managing ETDs throughout their lifecycle. These tools have been intended to be created as completely modular micro-services, i.e. single function standalone services that can be used alone or incorporated into larger repository systems.
Exploring Modular Curation for ETDs

• Methodology
  – Hold interviews with project partners and external ETD programs
  – Research open source modular curation tools and micro-services
  – Research existing open source submission and repository software systems for interoperability with curation tools
ETD Lifecycle Workflows

Create & Receive  
Appraise & Select  
Manage  
Store, Access & Re-Use
ETD Lifecycle Workflows

- Create & Receive
- Appraise & Select
- Manage
- Store, Access & Re-Use

ETD Metadata
Supplemental Files

Submission System

Graduate School/IT

Library/IT

Institutional Repository

ProQuest/Vendors

Author
ETD Lifecycle Workflows

Create & Receive

Appraise & Select

Manage

Store, Access & Re-Use

Submission System

Graduate School/IT

Library/IT

Institutional Repository

ProQuest/Vendors

ETD Metadata
Supplemental Files

Author
ETD Lifecycle Workflows

Create & Receive
Appraise & Select
Manage
Store, Access & Re-Use

Submission System
Graduate School/IT
Library/IT
Institutional Repository
ProQuest/Vendors
ProQuest/Vendors

ETD Metadata Supplemental Files
...here

Author
ETD Lifecycle Workflows

Create & Receive → Appraise & Select → Manage → Store, Access & Re-Use

- Submission System
- Graduate School/IT
- Library/IT
- ProQuest
- Institutional Repository
- ProQuest/Vendors
- Author

ETD Metadata Supplemental Files
Exploring Modular Curation for ETDs

- Proposed Modular Curation Services
  - ETD Format Recognition
  - Virus Checking
  - PREMIS Metadata Event Record-keeping
  - Digital Drop Box with Metadata Submission Functionality
Modular Curation Tools

Format Recognition

- Unix command-line programs (e.g., file) w/shell scripts
- JHOVE/2
- DROID
- FITS

Virus Checking

- Clam AV

Preservation Metadata

- DAITSS Description Service
- PREMIS Event Service (UNT)
Investigating Open IR Systems

• Dspace
  – Built-in support for format validation, virus checking & some preservation metadata (plugin architecture)

• EPrints
  – Built-in support for format validation, virus checking & some preservation metadata (plugin architecture)

• Open-ETD
  – No built-in support for any curation tools (no plugin architecture)

• ETD-db
  – No built-in support for any curation tools (no plugin architecture)
ETD Technologies on the Lifecycle

FTP, ProQuest UMI, EPrints, Open-ETD, etc.

ProQuest PQDT, DSpace, EPrints, Open-ETD, etc.

ProQuest UMI, EPrints, Open-ETD, etc.

ProQuest, DSpace, EPrints, ETD-db, Open-ETD, etc.
Lifecycle Curation Tools

- Conceptualise
- Virus Checking
- Dispose

Libraries UNT
PREMIS Event Record-Keeping

UNIX
JHOVE2
DROID
FITB

Preservation Metadata
ETD Lifecycle Workflows

Create & Receive  Appraise & Select  Manage  Store, Access & Re-Use

Submission System  Graduate School/IT  Library/IT

ProQuest/Vendors  Institutional Repository  ProQuest

Author

ETD Metadata Supplemental Files

UNIX

JHOVE2
Overall Findings

• Only a few submission & IR software systems have plugin architectures
• By the time ETDs arrive at the IR it is often too late to apply curation services
• Existing curation services have clear standalone usage instructions and available APIs
• ETD Programs have technical staff at their disposal to help implement modular curation
Next Steps

1. Simplify existing curation tool instructions for ETD stakeholders in:
   - Graduate Schools
   - Libraries
   - Vendors

2. Develop recommendations for using tool outputs, reports, metadata, etc.

3. Testing usage and implementations with project partners and other interested programs (Q7 on Post-Workshop Survey)
Project Wiki Workspace

Functional Requirements for ETD Micro-Services

Overview

Each of the Lifecycle Management Tools aims to serve as a standalone micro-service that can be called via command line or web interfaces in order to ensure that the systems can be easily integrated in existing environments in a modular way. Each micro-service will have clear documentation that will enable implementers to deploy the tool in their own setting. The intent of researching, developing and documenting these four micro-services is that they will catalytically enhance existing repository systems being used for ETDs, which often lack simple mechanisms for these functions.

The micro-service packages produced in the course of this project will include the following tools:

ETD Format Recognition Service

Accurate identification of ETD component format types is an important step in the ingestion process, especially as ETDs become more complex. This micro-service should:

1. enable batch identification of ETD files through integration of function calls from tools like JHOVE2, DROID, and other format identification tools; and
2. structure micro-service output in ad hoc tabular formats for importation into repository systems used for ETDs such as DSpace, and the ETD-db software, as well preservation repository software such as iRODS and DAITSS and preservation network software such as LOCKSS.

The project has successfully researched all of the primary format recognition utilities in current use, including JHOVE2, DROID, FITS and even the UNIX file command. Based on our research of these tools, a thorough analysis of integrations with existing repository systems, and interviews with numerous and diverse ETD programs, project work has shifted to documenting the proper usage of these tools in an ETD Program context.

See ETD Format Recognition Tools Documentation for ETDs.

PREMIS Metadata Event Record-Keeping Service

One gap highlighted in the needs analysis was the lack of simple PREMIS metadata and event record keeping tools for ETDs. This micro-service needs to:

1. generate PREMIS Event somnanta units to track a set of transitions in the lifecycle of particular ETDs using parameter calls to the micro-service; and
2. provide profile conformance options and documentation on how to use the metadata in different ETD repository systems.

See PREMIS Event Service Documentation.

Virus Checking Service

Virus checking is an obvious service needed in ETD programs, as students’ work is often infected unintentionally with computer viruses. This micro-service should:

1. provide the capability to check ETD component files using the ClamAV open source email gateway virus checking software;
2. record results of scans using the PREMIS metadata event tracking service; and
3. be designed such that other anti-virus tools can be called with it.

ClamAV was closely researched for any needed scripting and improvements, along with investigations into its potential for systematic integration with various repository systems. No special scripting is needed and this utility is best documented as a standalone service in conjunction with various ETD workflows prior to depositing in an IR. The parallel of this service can be recorded and used by the PREMIS Event Service. There are no other open-source anti-virus services as robust or prepackage as ClamAV.

See Virus Checking Documentation for ETDs.

Digital Drop Box (with metadata submission functionality)

This micro-service addresses a frequently sought function to provide a simple capability for users to deposit ETDs into a remote location via a web form that gathers requisite submission information requested by the ETD program.

The submission Information should:

1. generate PREMIS metadata for the ETD files deposited;
2. have the capacity to replicate the deposited content securely upon ingest into additional locations by calling other Unix tools such as rsync; and

See Digital Drop Box Documentation.
Coming Fall 2014

- Experimenting with Archivematica to model curation services for ETD preservation
  - Michigan State University
  - Simon Fraser University

- Researching a “dropbox” solution for ETD submissions that can facilitate preservation metadata
  - Working closely with DataStage team
Questions?
Lifecycle Management Tools

Digital Record Object Identification (DROID)
Demonstration
DROID Demo

Command-line to your installation directory (mine is “temp”)
  – Matthews-MacBook-Air:~ metamat$ cd Desktop/temp

Drop into your DROID installation directory
  – Matthews-MacBook-Air:temp metamat$ cd droid-binary-6/

Run the DROID start command
  – Matthews-MacBook-Air:droid-binary-6 metamat$ ./droid.sh

2013-07-23 12:36:32,377  INFO Starting DROID.
2013-07-23 12:36:35,656  INFO Creating profile: 1374597395656
2013-07-23 12:36:35,706  INFO Opening profile: 1374597395656
DROID Demo
DROID Demo
DROID Demo
DROID Demo
DROID Demo
DROID Demo

<table>
<thead>
<tr>
<th>Resource</th>
<th>Extension</th>
<th>Size</th>
<th>Last mod.</th>
<th>Ids</th>
<th>Format</th>
<th>Version</th>
<th>Mime type</th>
<th>PUID</th>
<th>Method</th>
<th>Hash</th>
</tr>
</thead>
<tbody>
<tr>
<td>/Users/metakat/Desktop/ETDs_for_Tools_Demo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DS_Store</td>
<td>m4a</td>
<td>6 KB</td>
<td>7/22/13</td>
<td></td>
<td>DS_store file (MAC)</td>
<td></td>
<td></td>
<td>fmt/394</td>
<td>Signature</td>
<td></td>
</tr>
<tr>
<td>ETD_Submission_1_Supplemental_file_1.m4a</td>
<td>m4a</td>
<td>29.1 MB</td>
<td>1/30/13</td>
<td></td>
<td>MPEG-4 Media File</td>
<td></td>
<td></td>
<td>fmt/199</td>
<td>Signature</td>
<td></td>
</tr>
<tr>
<td>ETD_Submission_File.pdf</td>
<td>pdf</td>
<td>203.8 KB</td>
<td>1/30/13</td>
<td></td>
<td>Acrobat PDF 1.5 – Portable Document Format</td>
<td>1.5</td>
<td>application/pdf</td>
<td>fmt/19</td>
<td>Signature</td>
<td></td>
</tr>
<tr>
<td>ETD_Submission_Metadata.xml</td>
<td>xml</td>
<td>1.6 KB</td>
<td>5/20/13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ETD_Submission_Supplemental_file_2.pptx</td>
<td>pptx</td>
<td>49.4 KB</td>
<td>12/5/12</td>
<td></td>
<td>Microsoft Powerpoint for Windows</td>
<td>2007 onw...</td>
<td></td>
<td>fmt/215</td>
<td>Container</td>
<td></td>
</tr>
</tbody>
</table>
DROID Demo
DROID Demo
DROID Demo
DROID Demo
DROID Demo

Questions?