ETD Repository: Drupal, Solr, Islandora, and Fedora Commons

Aaron Collie, Devin Higgins, Lucas Mak, Shawn Nicholson
Library Collections

**Digital**
- Librarian competence varies widely
- Spaghetti Infrastructure (e.g. ad hoc mysql, coldfusion, tomcat, apache, filesystem, flat html, etc)
- Access secondary to storage

**Traditional**
- Librarians professionally trained to collect, store, maintain, navigate, and provide globally envied customer service
- Systematic Infrastructure
- Standard Description
- Access primary to storage
Librarianship

- Librarianship is a profession which has survived the printing press, the publisher, the computer, the internet, and now the google

- That is because we have wonderful job security: entropy
But what is our ROLE?

- The **printing press** did a pretty decent job of replication
- The **publisher** has made a pretty penny on quality assurance
- The **computer** has revolutionized processing
- The **internet** continues to push the boundaries of distribution
- The **google** has (more) lawyers
Digital Information & Systems

- Metadata drives design
- Usable for many applications
- Sustainable over time

(pretty much the mantra for all library services)
Environmental Scan

• Dspace vs. contentDM
• EPrints vs. Digital Commons
• ICA-AtoM vs. Hydra
• Islandora
Clear winner!! (for us)

- Islandora
  ① Series of drupal modules; we like drupal
  ② Backed by fedora commons
  ③ Open Source & big hug community
  ④ Microservice architecture (think linux)
  ⑤ API
Technical Overview
Database light

• FOXML, Gsearch, Apache Solr, Akruba-LLStore
  – It does not require the use of database tables to "look up" the path to each file.
  – It stores files in a deterministic location based on a md5 hash (stored) and a unique id (PID) of each file.
  – The index can be rebuilt from the contents on the filesystem. Preserve the bits.
  – Messaging service can listen for and respond to events
Content friendly

• Content abstraction
  – Agnostic to format, complexity, mereology
  – Identifiers
  – Programmatic control

• Content relationships
  – RDF, Mulgara triplestore

• Content models
  – Predefined routines (pipe to...)
Community driven

• Solution Packs
  – E.g. imagemagick + Djatoka + OpenSeadragon
  – E.g. SHA-512 + cron job + status report

• Drupal
  – drush en antigravity –y
  – Drupal Forms API
Implementation(s)

• 1 mo: VirtualBox
• 6 mo: Hyper-V
• 12 mo: Dell PowerEdge R320 (x2)
• 18 mo: production, mirror, development
• 24 mo: live site
• 30 mo: (out for beers)
• 36 mo: sorry, you said “incremental” what?
Pilot collection

- Effective Spring 2011, MSU no longer accepts bounded dissertations and only accepts electronic submission via ProQuest
- Estimated ~500-600 dissertations per year
- Received every 3 hours via SFTP from vendor
- ZIP with PDF and Metadata
PRAIRIE FEN HYDROLOGY

By

Hassan Abbas

A DISSERTATION

Presented to the Faculty of Michigan State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Civil Engineering

2011

Abstract

Prairie fens provide habitat to more than 50 rare species, including the federally endangered Mitchell’s satyr butterfly. Substantial resources have been invested in their protection; however, these conservation efforts have proceeded without an understanding of the underlying groundwater flow regime that is critical to these fens. Regional scale hydrological investigations could cost hundreds of
"There is something in all this very like democracy": Cultures of Political Discussion in the Victorian Novel

<table>
<thead>
<tr>
<th>ID</th>
<th>Label</th>
<th>Type</th>
<th>Mime type</th>
<th>Size</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ</td>
<td>PDF of this object.</td>
<td>Managed</td>
<td>application/pdf</td>
<td>1.07 MB</td>
<td>download</td>
</tr>
<tr>
<td>DC</td>
<td>Dublin Core record of this object.</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>5.06 MB</td>
<td>download</td>
</tr>
<tr>
<td>MODS</td>
<td>MODS record of this object.</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>8.13 MB</td>
<td>download</td>
</tr>
<tr>
<td>MARCXML</td>
<td>MARCXML record of this object.</td>
<td>text/xml</td>
<td>8.72 KB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDLTD</td>
<td>NDLTD record of this object.</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>5.7 KB</td>
<td>download</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>Federated Object to Object Relationship Metadata.</td>
<td>Inline XML</td>
<td>application/rdf+json</td>
<td>589 B</td>
<td>download</td>
</tr>
<tr>
<td>TN</td>
<td>TN</td>
<td>Managed</td>
<td>application/octet-stream</td>
<td>3.63 KB</td>
<td>download</td>
</tr>
<tr>
<td>PREVIEW</td>
<td>PREVIEW</td>
<td>Managed</td>
<td>application/octet-stream</td>
<td>11.8 KB</td>
<td>download</td>
</tr>
<tr>
<td>FULL_TEXT</td>
<td>FULL_TEXT</td>
<td>Managed</td>
<td>text/plain</td>
<td>557.14 KB</td>
<td>download</td>
</tr>
<tr>
<td>ID</td>
<td>Label</td>
<td>Type</td>
<td>Mime type</td>
<td>Size</td>
<td>Operations</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------</td>
<td>--------------</td>
<td>--------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>PDF</td>
<td>DC</td>
<td>Managed</td>
<td>application/pdf</td>
<td>1.07 MB</td>
<td>download, delete</td>
</tr>
<tr>
<td>MODS</td>
<td>MARCXML record of this object.</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>5.06 KB</td>
<td>download</td>
</tr>
<tr>
<td>NDLTD</td>
<td>NDLTD record of this object.</td>
<td>Inline XML</td>
<td>text/xml</td>
<td>8.13 KB</td>
<td>download, edit, delete</td>
</tr>
<tr>
<td>TN</td>
<td>TN</td>
<td>PREVIEW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FULL TEXT</td>
<td>FULL TEXT</td>
<td></td>
<td></td>
<td>5.7 KB</td>
<td></td>
</tr>
</tbody>
</table>
Metadata

• Sources
  – MSU Library Catalog
    • Original cataloging done for MSU ETD
  – Student-supplied metadata (ProQuest XML)
    • Broad subject categories, keywords, names of advisors & committee members, possible typos

• Targets
  – MarcXML (already exists in OPAC for some ETDs)
  – MODS (MSU-L preferred schema)
  – Dublin Core (required by OAI-PMH and Fedora)
  – NDLTD ETD-MS (international standard for ETD)
Existing Catalog Records of MSU theses and dissertations

• Have
  – Library of Congress Subject Headings (LCSHs)
  – Local accession number
  – Name authority control per local policy
  – MARC 502 dissertation note (degree name, program/academic unit, degree year)

• Don’t have
  – Access points or notes for advisors or committee members
  – Summary/abstract
Metadata Reconciliation, Transformation & Enrichment

• If cataloged, get XML from the catalog
  – Reuse OPAC data if available
    • Subject headings and controlled names in catalog records
  – Enrich XML records derived from existing catalog records with unique data captured from ProQuest
    • Advisor, committee members, subject categories, copyrights/embargo info, abstract
• If not cataloged, create target XML records directly from ProQuest XML
Metadata for Access
Problem: Metadata as Data
Academic “Topics” at MSU
Thinking about Data Structures

• Moving from discrete XML files (each with data about one item) to JSON objects (containing all data).

XML ->
Python lxml ->
Python NetworkX ->
Gephi to visualize networks
Back to the Beginning

• Problem: How to make metadata analysis work for the library.

• Analyzed metadata (in the end) is also metadata about the collection.

• New browsing and exploring options available to the user.

• Convert static graphs into interactive tools for users.
Psychology

The graph below represents the connections between the selected program (in this case Psychology) and all other graduate programs at MSU for roughly 2011 to the present. Connections are based on shared features in the categories below: Subject headings, publication date, keywords, geographic headings, topics, and advisors or committee members.

Subjects
- Psychology (65)
- Clinical psychology (16)
- Organizational behavior (10)
- Neurosciences (7)

Advisors/Committee Members
- Ryan, Ann Marie (17)
- Schmitt, Neal (11)
- Davidson, William (9)
- Scharbrock, John (8)

Kinesiology → Psychology

Topics
- African American studies
- Behavioral sciences
- Educational tests & measurements
- Management
- Psychology
- Public health
- Social psychology

Committee/Advisors
- Campbell, Rebecca
- Cesario, Joseph
- Kerr, Norb
- Reckase, Mark
- Scharbrock, John
- Stanbury, Kathy
- Wittenbaum, Gwen

Geographic
- United States (6)
- Middle West (2)
- Michigan (2)
- Malawi (1)
Digital Information & Systems

• Metadata drives design
• Usable for many applications
• Sustainable over time

(pretty much the mantra for all library services)
More things to come...
Questions?

MSU Libraries
Aaron Collie
Lucas Mak
Devin Higgins
Shawn Nicholson

Contact for more information: collie@msu.edu

Credits for Icons
Tag designed by Garrett Knoll from the Noun Project
3 Book Icons designed by Julien Deveaux from the Noun Project
File Cabinet designed by Alex Hartmann from the Noun Project