The Adolescent Institutional Repository: Metadata Management Perspectives and Challenges

Casey Mullin, Western Washington University

Available at: https://works.bepress.com/casey-mullin/14/
The Adolescent Institutional Repository
Metadata Management Perspectives and Challenges

Casey A. Mullin
Head of Cataloging and Metadata Services
Western Washington University

American Library Association Annual Conference
June 25, 2018
Introduction – A trip down the rabbit hole

• Shortly after I arrived…
• And then…
• And while we’re at it…
• Launched in 2014
• Partnership between the Libraries, the Provost, the Graduate School, and the Office of Research and Sponsored Programs
• Electronic theses (self-deposit)
• Faculty publications
• University-produced periodicals
  • *Window Magazine* (alumni magazine)
  • *Klipsun Magazine* (student-run magazine)
• Journals (managed in CEDAR)
  • *Journal of Educational Controversy* (peer-reviewed)
  • *Occam’s Razor* (student-run)
• Salish Sea Ecosystem Conference
  • Archive hosted in CEDAR
• ...and more (*over 5,600 objects in 60+ collections*)
Leadership Roles
(or, Who’s in charge here?)

- Scholarly Communications Librarian
  - Curatorial oversight and promotion
- IR Manager
  - Reports to Director of Technology & Discovery Services
- Head of Cataloging and Metadata Services
  - Supervises unit that provides remediation/enhancement services
- Other stakeholders, e.g.,
  - University Archivist (curatorial role for select collections)
Collaborative Model of Digital Collections Management

Curatorship

Discovery/Systems Support

Metadata Services
Ultimate goals

• Standardize metadata profiles for all CEDAR collections
  – To borrow a phrase from the PCC, establish a “CEDAR standard record”
• Safeguard the integrity of our metadata for future platform migrations
• Harmonize descriptive metadata practices across all three platforms
Step 1 – Create a “floor”

- **Required elements from the Orbis Cascade Alliance Dublin Core Best Practices**
  - Date
  - Identifier
  - Rights
  - Title
  - Type

- **Recommended elements from OCA DC BP**
  - Creator
  - Description
  - Format
  - Language
  - Relation
  - Spatial coverage
  - Subject

- **Added benefit**: DPLA hub preparedness

**https://www.orbiscascade.org/dublin-core-best-practices/**
Step 1 Challenges

• Dublin Core mappings are stored in the vendor’s “black box”

• Digital Commons native metadata is not in a standardized schema
  – Some OTOB patterns based on content “bucket”
  – Different names for the same thing

• Dublin Core mappings had to be determined collection by collection
  – This lead naturally to the next step…
Step 2 – Fill gaps in existing metadata profiles

- CEDAR collections added over several years, with differing amounts of metadata input during the planning stage
  - Initial collections had no data dictionary
  - Later collections had bespoke data dictionaries, whose format varied over time
- In June 2017, we identified all the gaps and sent a suite of requests to bepress, including:
  - Desired Dublin Core mapping changes
  - Additional elements desired
Step 2 – Fill gaps in existing metadata profiles

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal of Educational Controversy (jec)</td>
<td>Dublin Core (current)</td>
<td>Dublin Core (desired changes)</td>
</tr>
<tr>
<td>Article Title</td>
<td>dc.title</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>dc.creator</td>
<td></td>
</tr>
<tr>
<td>Short Title</td>
<td></td>
<td>dc.title.alternative</td>
</tr>
<tr>
<td>Keywords</td>
<td>dc.subject</td>
<td></td>
</tr>
<tr>
<td>Publication Date</td>
<td>dc.date.created</td>
<td></td>
</tr>
<tr>
<td>Disciplines</td>
<td>dc.subject</td>
<td></td>
</tr>
<tr>
<td>Document Type</td>
<td>dc.type</td>
<td></td>
</tr>
<tr>
<td>Theme</td>
<td></td>
<td>dc.format</td>
</tr>
<tr>
<td>Abstract</td>
<td>dc.description.abstract</td>
<td></td>
</tr>
<tr>
<td>Cover Page Footnote</td>
<td>dc.description</td>
<td></td>
</tr>
<tr>
<td>Peer Reviewed</td>
<td>dc.description</td>
<td></td>
</tr>
<tr>
<td>Recommended Citation</td>
<td>dc.source</td>
<td>Remove mapping</td>
</tr>
<tr>
<td>URL</td>
<td>dc.identifier</td>
<td></td>
</tr>
<tr>
<td>Rights</td>
<td>dc.rights</td>
<td>Requested additional element (Patterned after: Rights in surveyresearch_docs)</td>
</tr>
<tr>
<td>Language</td>
<td>dc.language</td>
<td>Requested additional element (Patterned after: Language in huxley_stupubs)</td>
</tr>
<tr>
<td>Related resource</td>
<td>dc.relation</td>
<td>Requested additional element, free-text, optional element</td>
</tr>
</tbody>
</table>
Step 3 – Create complete data dictionaries for each collection

- Columns used
  - Display label
  - Database label
  - Dublin Core mapping
  - Repeatability (and delimiting method)
  - Required
    - One column became two: Required at time of upload vs. Required during remediation
  - Data type
  - Input conventions
  - Default value
  - Display to public (reality vs. aspirations)
  - Notes column (remediation, questions for bepress, etc.)
### Step 3 – Create complete data dictionaries for each collection

**CEasar Data Dictionaries - Excel**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Database Labels</th>
<th>Narrative</th>
<th>Type</th>
<th>Required at Time of Entry</th>
<th>Required as Part of Immediate Update</th>
<th>Data Type</th>
<th>Input Connections</th>
<th>Display in CEasar?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISBN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Right Statements**

<table>
<thead>
<tr>
<th>Right Statement</th>
<th>Right(s)</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Related Resources**

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Resource Name</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Additional Information**

- See [specific URL for additional information](https://example.com)
- Contact [specific contact information](mailto:contact@example.com)

---

*...times 30*

Mullin, 6/25/18
Rinse and repeat

• Steps 2 and 3 unfolded concurrently over several months
• As we made certain discoveries, it was necessary to go back to the first data dictionary and start again
• Continued to identify element gaps, display lacunae and remediation opportunities
• “What have I gotten myself into?!?”
• This led to Step 4…

Step 4 – Document, document, document

• **Problem #1**: Giant spreadsheet shows the “what”, but I also wanted to know the “why,” “when,” and “who”

• **Solution**: create a registry of CEDAR collections
### Step 4 – Document, document, document

#### Excerpt from CEDAR registry

<table>
<thead>
<tr>
<th>Collection Name</th>
<th>Huxley College Graduate and Undergraduate Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Name (abbreviated)</td>
<td>huxley_stupubs</td>
</tr>
<tr>
<td>Digital Commons “Bucket” (i.e., ir_etd, ir_series, ir_community, ir_book, ir_journal, ir_event_community)</td>
<td>ir_etd</td>
</tr>
<tr>
<td>Month/Year Added to CEDAR</td>
<td>2015-09-08</td>
</tr>
<tr>
<td>History of Staff Roles and Contributions (e.g., content upload, metadata creation and remediation, etc.)</td>
<td>Collection originated from Huxley faculty, Stefan Freelan and Leo Bodensteiner to Peter Smith for ContentDM in 2007-12. Kris Rex did the uploading, metadata creation and remediation to transfer to CEDAR. Kris Rex continues to do the uploading and metadata creation.</td>
</tr>
</tbody>
</table>
| Remediation Checklist | • Check LCNAF for subjects-names and advisor entries; establish names as needed  
| | • Add season to dates that lack it  
| | • Add standardized rights statement/CC license and possibly free-text rights statement |
| Other Comments | Collection contains native objects and objects that are linked from other collections/sources for discovery purposes. Remediation checklist pertains only to native objects. |
Step 4 – Document, document, document

- **Problem #2**: Giant spreadsheet is unwieldy for broader audiences and ineffective at prompting focused conversations with stakeholders

- **Solution**: distill open metadata issues and questions into human-friendly Word document
  - 30+ matrices → 11 pages
## Excerpt from Metadata issues document

### Unresolved issues by collection

<table>
<thead>
<tr>
<th>Collection name (abbreviated)</th>
<th>Issue</th>
<th>Response/Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>wwu et</td>
<td>“Creative Commons License” field: is this redundant with the “Rights” field? If so, can it be removed?</td>
<td></td>
</tr>
<tr>
<td>huxley_stupubs</td>
<td>The “Department field” is applicable only if document is also a thesis; field is redundant in this series, but is used to mimic what is in the Masters thesis collection in Alma/Primo. Should we still keep it since we are linking Huxley theses to the series?</td>
<td></td>
</tr>
<tr>
<td>ir_series facpubs collections</td>
<td>In a few places, Creative Commons License statements appear in the Required Publisher’s Statement field. Should they be moved to the Rights field?</td>
<td></td>
</tr>
</tbody>
</table>
Step 5 – Discuss

• Periodic meetings with stakeholders (G7)
• Metadata issues document serves as a “punch list” of decision points
• Decisions reached to date
  – Genre/Form source vocabulary: AAT
  – Rights statements: 3-fold solution
    • Free-text boilerplate (determined at collection level)
    • Creative Commons License (drop-down)
    • Rightstatements.org statement (drop-down)
• Another suite of bepress requests sent May 2018
• Next action item: staff workflow document
Step 6 – Create “master” data dictionary

- Core fields: common set, based on OCA DC BPs and Digital Commons functional requirements
- Extended fields: “long tail” of fields used only for certain collections
  - Are all still needed?
  - Is there semantic overlap?
- **Objective**: streamlined metadata intake process for new collections
  - Close the boutique, open the factory
Step 7 – Remediate

- Staff have proceeded with “low-hanging fruit” using batch revise process
  - Language
  - Format (MIME type)
- Other items await bepress action and/or stakeholder decisions
- Job security?
  - LCSH and AAT for all 5,600+ objects
  - Authority control
Rinse and repeat

• Steps 1-7 continue, more or less concurrently, but progress has been achieved on all fronts

• Lessons learned:
  – Persuade IR stakeholders to engage metadata expertise early and often
  – Document, document, document
  – The metadata will outlive the system; be prepared for future migrations
  – Be patient; this is an iterative process
  – Take the time; unique and local content deserves your “love”
  – Many hands make light work
Thank you!!

Casey.Mullin@wwu.edu
(or come visit!)