LOOKING BACK, MOVING FORWARD
Remediating Duke Digital Collections Metadata

MAGGIE DICKSON
Metadata Architect, Duke University Libraries

“We Can Do It, You Can Too” Virtual Preconference, June 8, 2016
Hosted by ALCTS, the Association for Library Collections and Technical Services
Who’s talking:

MAGGIE DICKSON
Metadata Architect
Duke University Libraries
DUL Digital Collections

- 21 years of digital projects development
- 100+ collections, 200K+ items
- Most collections delivered via custom Django-based discovery and access application

https://repository.lib.duke.edu/dc/wdukesons/dscsi05008
Why Remediate Now?

- Opportunity to migrate to the Duke Digital Repository (Hydra stack)
- Creation of Metadata Architect position

http://library.duke.edu/digitalcollections/adaccess_T0736/
Formation of Metadata Task Group

- Metadata Architect (chair)
- Archivist for Metadata & Encoding for Rubenstein Library
- Catalog Librarian from Technical Services
- Assessment Librarian
- Digital Collections Program Manager
- Digital Projects Developer

http://library.duke.edu/digitalcollections/oaaaarchives_BBB4169
Our Charge

1. Review and assess existing metadata for digital collections, and make recommendations for basic cleanup and consistency
2. Implement recommendations made in task #1
3. Prepare collections for sharing with the DPLA
4. Consult with stakeholder groups on metadata needs
5. Make recommendations for organizational resources, resources, technologies, and standards to support the creation and management of descriptive metadata at DUL
Guiding Principles

- Fitness for Purpose
- Broad Applicability
- Broad Shareability
- Forward-Thinking

http://library.duke.edu/digitalcollections/gedney_CC0037
The Dataset

- 117 MB CSV file
- 163K+ items
- 2M+ statements

http://library.duke.edu/digitalcollections/paverjohn_PAV0065
The Dataset

<table>
<thead>
<tr>
<th>All</th>
<th>collection</th>
<th>category</th>
<th>identifier</th>
<th>field</th>
<th>dc_mapping</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Product</td>
<td>subject</td>
<td>Sanitary Napkins</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Medium</td>
<td>medium</td>
<td>Magazine</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>DCMI_Type</td>
<td>type</td>
<td>Still Image</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Headline</td>
<td>title</td>
<td>Kotex</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Company</td>
<td>creator</td>
<td>Cellucotton Products Company</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Publication</td>
<td>source</td>
<td>Delineator</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Temporal_Coverage</td>
<td>temporal</td>
<td>1921</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Genre</td>
<td>type</td>
<td>advertisements</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Audience</td>
<td>audience</td>
<td>Consumer</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Date</td>
<td>date</td>
<td>1921</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Type</td>
<td>type</td>
<td>drawings (visual works)</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Subject</td>
<td>subject</td>
<td>Beauty and Hygiene</td>
</tr>
<tr>
<td>★</td>
<td>adaccess</td>
<td>advertising</td>
<td>BH0001</td>
<td>Subject</td>
<td>subject</td>
<td>Feminine Hygiene</td>
</tr>
</tbody>
</table>

One item record, formatted as an ‘enhanced’ triple statement, viewed in OpenRefine
Visualization

- Tableau Public
- Free data visualization software

https://public.tableau.com
Tableau Public
Metadata Properties

- 85 unique fields
- 61 ‘duketerms’
- 24 simple Dublin Core
Review & Analysis

Used OpenRefine to determine:
- What collections used the field?
- Was usage consistent within/across collections?
- What kinds of values? Controlled vocabularies? Free text? Both?
- How many unique values?
Documentation of Analysis

- Used Google Sheets to record analysis, as well as recommendations for remediation
  - Great for concurrent use
  - Allowed for comments/discussions
  - Easily shared with interested stakeholders
<table>
<thead>
<tr>
<th>Current Dublin Core Mapping</th>
<th>Field Label</th>
<th># of instances</th>
<th># of Coll.</th>
<th># of Val.</th>
<th>Collection Usage</th>
<th>Free text/CV/Date</th>
<th>Usage Notes</th>
<th>Issues</th>
<th>Property Recommendations</th>
<th>Value Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>title</td>
<td>81568</td>
<td>48</td>
<td>61239</td>
<td>Used in every collection except for road, mma, adaccess</td>
<td>Free text</td>
<td>Messy, slight variations across items that have the same titles (generally advertising collections)</td>
<td>Same as Title</td>
<td>Require field. Set policies for type of collection (archival, doc_photo, etc.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>title</td>
<td>78736</td>
<td>9</td>
<td>37420</td>
<td>Used in advertising collections</td>
<td>Free text</td>
<td>Same Title</td>
<td>Require field. Set policies for type of collection (archival, doc_photo, etc.)</td>
<td>Delete duplicate title entries</td>
<td></td>
</tr>
<tr>
<td></td>
<td>alternative</td>
<td>746</td>
<td>8</td>
<td>655</td>
<td>duc; frankspeda; gamble; hasm; italianposters; profiam; quintets; russianposters</td>
<td>Free text</td>
<td>25 instances of value 'None' in Frank Espada collection. Some instances seem more like description than title (eg. 'Photographs of the principal individuals involved: James. B. Duke'). Some values that are showing up in alternative are in title field in CONTENTdm, was mapping changed when METS was generated?</td>
<td>Same as Title</td>
<td>Merge with Title field. Use Headline as collection-specific label</td>
<td></td>
</tr>
<tr>
<td></td>
<td>alternative</td>
<td>2097</td>
<td>1</td>
<td>1920</td>
<td>hasm</td>
<td>Free text</td>
<td>Sometimes authoritative value of proper name, sometimes general heading (eg. 'Museums'), sometimes appears to have been transcribed. Lots of brackets, question marks, etc.</td>
<td>4626 instances of some form of 'N/A'; 2281 instances of some form of 'unknown'. Lots of 'Various'. Lots of brackets, question marks, etc. Slight variations in terms that result in many different values for the same entity.</td>
<td>Merge with Creator. Use Company as collection-specific label</td>
<td>Keep as legacy field.</td>
</tr>
<tr>
<td></td>
<td>alternative</td>
<td>1373</td>
<td>1</td>
<td>1306</td>
<td>hasm</td>
<td>Free text</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>4626 instances of some form of 'N/A'; 2281 instances of some form of 'unknown'. Lots of 'Various'. Lots of brackets, question marks, etc. Slight variations in terms that result in many different values for the same entity.</td>
<td>Merge with Creator. Use Company as collection-specific label</td>
<td>Keep as legacy field.</td>
</tr>
<tr>
<td></td>
<td>creator</td>
<td>89009</td>
<td>10</td>
<td>14155</td>
<td>Used in advertising collections</td>
<td>Mix of free text/authoritative</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>4626 instances of some form of 'N/A'; 2281 instances of some form of 'unknown'. Lots of 'Various'. Lots of brackets, question marks, etc. Slight variations in terms that result in many different values for the same entity.</td>
<td>Merge with Creator. Use Company as collection-specific label</td>
<td>Cluster and edit authoritative (authoritative-like) headings and question marks. Replace 'N/A' and 'unknown'. Move description.</td>
</tr>
<tr>
<td></td>
<td>composer</td>
<td>17636</td>
<td>3</td>
<td>6888</td>
<td>Mix of free text/authoritative</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>Transcribed values with authoritative values. Question marks.</td>
<td>Transcribed values mixed with authoritative values (heading for GF Handel alongside 'A Lady', lots of initials). Some could be authoritative but are not (Pagannini). Lots of quotation marks, allcaps (mostly Mazzoni), brackets, N/A.</td>
<td>Merge with Creator. Use Composer as collection-specific label</td>
<td>Cluster and edit authoritative (authoritative-like) headings. Remove transcribed description.</td>
</tr>
<tr>
<td></td>
<td>creator</td>
<td>19119</td>
<td>26</td>
<td>899</td>
<td>Used in all collections</td>
<td>Mix of free text/authoritative</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>Merge with Creator. Use Composer as collection-specific label</td>
<td>Move 26 instances of creator heading. Cluster at (or authoritative-like) heading marks. Remove t Description. Make this a lot easier.</td>
</tr>
<tr>
<td></td>
<td>creator</td>
<td>446</td>
<td>1</td>
<td>29</td>
<td>behindthewell</td>
<td>Controlled vocabulary</td>
<td>Inverted personal names. One corporate body (Fox Television). Very clean</td>
<td>Local controlled vocabulary - most likely not found outside of Duke? Except for Fox.</td>
<td>Remap to contributor.</td>
<td>Remap to contributor.</td>
</tr>
<tr>
<td></td>
<td>creator</td>
<td>6826</td>
<td>2</td>
<td>3009</td>
<td>hasm; sheetmusicindex</td>
<td>Mix of free text/authoritative</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>Sometimes authoritative value of proper name, sometimes transcribed.</td>
<td>Merge with Creator. Use Composer as collection-specific label</td>
<td>Merge with Creator. Use Composer as collection-specific label</td>
</tr>
</tbody>
</table>
Remediation: Workflow

- Began in December, ongoing
- Collection-by-collection or collection grouping
- All changes documented by field in collection-level Google docs

http://library.duke.edu/digitalcollections/gedney_CM0126
Remediation Tools: OpenRefine

- openrefine.org
- wiki: https://github.com/OpenRefine/OpenRefine/wiki
- Features:
  - Faceting
  - Filtering
  - ‘Cluster and edit’ feature
  - Regular expression transformations
  - Reconciliation against external data sources (eventually)
Remediation Tools: Text editor/Regex

- TextWrangler (It’s free! It’s great!)
  - Faster processing on large dataset
  - Better for complex regular expressions - supports grep
Remediation Tools: Ruby scripting

- Self-education:
  - Lynda.com coursework
  - RailsBridge.org workshops

About RailsBridge:

We teach people to code because we believe that the people making technology should accurately reflect the diversity of those using it.

We want to push for all kinds of diversity in tech: gender, race, sexual orientation, ability, and class. Women were the first population we focused on, but aren't the last.

We value acceptance, enthusiasm, and hard work. The RailsBridge community thrives on optimism and a love for making great things.
Remediation: What It Consists Of

- Field normalization
  - Fields merged/renamed/remapped to Dublin Core
  - Usage aligned with DPLA
  - Consistent application of core fields across collections
- Value normalization
  - Values mapped to appropriate controlled vocabularies/thesauri (LCSH, LCNAF, AAT) when possible
  - Normalization of geographic, date, format terms
Implementation of EDTF

Extended Date Time Format - LoC date encoding standard
- https://www.loc.gov/standards/datetime
- Extension of ISO 8601
- Allows for the encoding of ‘squishy’ dates
  - Uncertain/approximate dates
  - Partial dates
  - Seasons

http://library.duke.edu/digitalcollections/adaccess_BH1212
Implementation of EDTF

~12K unique date values in our metadata

- Process
  - Identified patterns
  - Decided how to implement standard
  - Decided how to display EDTF encoded dates
# Implementation of EDTF

<table>
<thead>
<tr>
<th>Existing date value</th>
<th>EDTF-encoded date</th>
<th>Display format</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 1940?</td>
<td>1940-05?</td>
<td>May 1940?</td>
</tr>
<tr>
<td>ca. 1931 Feb. 22</td>
<td>1931-02-22~</td>
<td>Circa February 22, 1931</td>
</tr>
<tr>
<td>1930s-1940s</td>
<td>193u/194u</td>
<td>1930s to 1940s</td>
</tr>
<tr>
<td>1920s Jan. 7</td>
<td>192x-01-07</td>
<td>January 7, 1920s</td>
</tr>
<tr>
<td>Fall 1961</td>
<td>1961-23</td>
<td>Fall 1961</td>
</tr>
<tr>
<td>Undated</td>
<td>uuuu</td>
<td>Undated</td>
</tr>
</tbody>
</table>
Implementation of EDTF

- Transformed dates using regex
- Developed Ruby gem for interpretation/display of EDTF-encoded dates:
  - ‘edtf-humanize’
- Available on github & rubygems.org
Synchronization with EAD Finding Aids

- EADID used to dynamically display source collection information
- ArchivesSpace identifier used to create deep link back to item in context in finding aid
- ARKs for items stored in DAOs in finding aid to provide permanent link to item in digital collection
What All This Gets Us

- A better, cleaner user experience
- Meaningful cross-collection searching and faceting
- Data that allows for date or geographic visualization
- Well-positioned for transition to linked data
- Better ‘shareability’ - DPLA, etc
- Easier migration when we have to do it all over again!
Next Steps/Challenges

- Further remediation necessary
- Reconciliation with Linked Data sources
- Programmatic approach for auditing repository metadata quality
- Tighter integration with other DUL descriptive practices

http://library.duke.edu/digitalcollections/protfam_prfad02346
THANKS!
maggie.dickson@duke.edu