Remodeling Archival Metadata Descriptions for Linked Archives

Brian Dobreski, Jaihyun Park, Alicia Leathers, & Jian Qin
School of Information Studies
Syracuse University
Syracuse, New York, USA
What are special collections?

• Cultural heritage objects
  • Archives
    • Text documents
    • Post cards
    • TV program videos
  • Photographs
  • Art and architecture drawings
• Rare books
• Sound recordings
• …
Metadata descriptions for special collections

Object and item level metadata descriptions use an in-house schema based on METS and MODS

Finding aids use *Encoded Archival Description (EAD)*

- Collection level
- Object level
- Item level
Collection level metadata

- EAD record
- Relations to other special collections
  - Herbert Beckhard Papers
  - George Goodwin Collection Relating to Marcel Breuer
  - St. Luke's Church Building Committee Records
- A wide variety of material types
- A large number of people and corporate bodies
Object level metadata

```xml
<mets:metsHdr beginDate="2017-12-07T16:09:13-05:00"/>
<mets:agent id="EDITOR" desc="Syracuse University Library">
  <mets:name> Suyracuse University Library </mets:name>
</mets:agent>

<mets:dmdSec id="dmd1">
  <mets:mdWrap>
    <mets:xlmdWrap>
      <mets:xmlData>
        <dc:identifier>38</dc:identifier>
        <dc:title>Wire</dc:title>
        <dc:creator>Roth, Gordon, B. (Author)</dc:creator>
        <dc:date>1945-05-11</dc:date>
        <dc:subject>Geller House I</dc:subject>
        <dc:type>Correspondence</dc:type>
        <dc:rights>Images supplied by the Marcel Breuer Papers, Syracuse</dc:rights>
        <dc:language>English</dc:language>
        <dc:description>Geller House I. Roth, Gordon, B. (Contractor). Roth,</dc:description>
      </mets:xmlData>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
```
Object and item level metadata

Nature of the object: Correspondence
Type of content: Wire to contact the gas company
Persons and firms: Role (author, contractor, recipient, etc.)

Digital counterpart: images of correspondence
Access and organizations: ID, location, sequence of order
Sample metadata record for object

collection_id: 227
object_id: 78584
object_type: 1
time_stamp_created: 2017-03-01 16:50:23
time_stamp_updated: 2017-03-01 16:50:23
time_stamp_exported: 0000-00-00 00:00:00
object_draft: 0
object_deleted: 0
title: Das M'zuzele
title_alt: Lots of empty data cells

date_original_range: 1917-
date_original_display: 1917
date_issued:
date_issued_display:
description: &lt;p&gt;Foreign Blue Amberol record release: Hebrew Series.&lt;/p&gt;
coverage:
series: Edison Blue Amberol
series_id: 10051
media_type: Sound Recording
open_closed: Open
type: Music

internal_id: eba_10051
bibid: 1572810
alt_repo:
draft:
rights:
geo_code:
donor:
summon_content_type: Music Recording
language: Yiddish
subject: Songs, Yiddish
notes:
index:

Lots of administrative metadata

IDs are local and internal
<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>collection_id</td>
<td>227</td>
</tr>
<tr>
<td>object_id</td>
<td>78584</td>
</tr>
<tr>
<td>item_id</td>
<td>80312</td>
</tr>
<tr>
<td>object_type_id</td>
<td>2</td>
</tr>
<tr>
<td>timestamp_created</td>
<td>2017-03-01 17:00:51</td>
</tr>
<tr>
<td>timestamp_updated</td>
<td>2017-03-01 17:00:51</td>
</tr>
<tr>
<td>timestamp_export</td>
<td>0000-00-00 00:00:00</td>
</tr>
<tr>
<td>item_draft</td>
<td>0</td>
</tr>
<tr>
<td>item_deleted</td>
<td>0</td>
</tr>
<tr>
<td>title</td>
<td>eba100511edited</td>
</tr>
<tr>
<td>date_digital</td>
<td>2016-02-04</td>
</tr>
<tr>
<td>linked_objects</td>
<td>78584</td>
</tr>
<tr>
<td>dimensions</td>
<td>03:01</td>
</tr>
<tr>
<td>duration</td>
<td>uElipsh9p0ihfy0mVAvnrA</td>
</tr>
<tr>
<td>internal_id</td>
<td>scrc/belfer_cylinders/tinfoil2.gif</td>
</tr>
<tr>
<td>file_display</td>
<td>scrc/belfer_cylinders/tinfoil2.gif</td>
</tr>
<tr>
<td>file_archive</td>
<td>eba_100511_1_edited.mp3</td>
</tr>
<tr>
<td>internal_file_path</td>
<td>scrc/belfer_cylinders/tinfoil2.gif</td>
</tr>
<tr>
<td>file_size_display</td>
<td>5.53 MB</td>
</tr>
<tr>
<td>item_html</td>
<td>Checked</td>
</tr>
<tr>
<td>item_download</td>
<td>9772c709025ec3ffe3f13635d4835292</td>
</tr>
<tr>
<td>checksum_archive</td>
<td></td>
</tr>
<tr>
<td>dimensions_digital</td>
<td></td>
</tr>
</tbody>
</table>

**General Technical Information**

- **Compression:**
- **PPI:**
- **Quality:** Good
- **ScanHW:** Benchmark ADC
- **ScanSW:** ProTools LE 7.4
- **Sample Rate (Audio):** 44.1
- **Sample Rate (Video):**
- **Sampling Ratio:**
- **Codec:** Belfer Audio Archive
- **Bit Depth:** 24 bits
- **Sampling Rate:** 44.1

**Digitized By:**
- Belfer Audio Archive

**File Path:**
- scrc/belfer_cylinders/tinfoil2.gif
- eba_100511_1_edited.mp3
- scrc/belfer_cylinders/tinfoil2.gif

**File Quality:**
- Good
Why ontological modeling?

• Enrich semantics in metadata by establishing a network of related people, events, subjects, places, times, material types, and other features important and meaningful for users

• Transform the structures of metadata to fully realize the technology potentials

• Enable the utilization of semantic data available externally and the publication and sharing of the semantic data about our collections
Metadata samples

• Belfer Cylinders Collection: music and spoken word recordings dating from 1890 to 1929.
  • 1729 physical item records and 1729 digital item records
  • 3000 name records for individuals, groups, and other entities in various roles.

• Ronald G. Becker Collection of Charles Eisenmann Photographs:
  • photographs of 19th century sideshows, circuses, and performers, most taken by Charles Eisenmann or his successor Frank Wendt, dating from 1836 to 1960
  • 1,414 physical item records, 1,416 digital item records, and 1,504 name records describing various roles.

• Ted Koppel Collection: videos of ABC News television programming with Ted Koppel, including approximately
  • 6,600 episodes of Nightline (March 1980-November 2005)
  • 7416 physical item records, 13,610 digital item records, and 72,988 name records of individuals appearing in various roles.
<table>
<thead>
<tr>
<th>Category</th>
<th>Elements common to both</th>
<th>Physical items</th>
<th>Digital items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative</td>
<td>collection_id, internal_id, object_id, object_type_id, time_stamp_created, time_stamp_export, time_stamp_updated</td>
<td>alt_repo, bibid, series_id, date_issued, donor, draft, index, location, notes, object_deleted, object_draft, open_closed, related_items, rights, summon_content_type</td>
<td>item_id, Item_draft, item_deleted, digitized_by, linked_objects, item_html, item_download, checksum_display, checksum_archive, notes, orig_gen, orig_format, orig_notes, dig_notes</td>
</tr>
<tr>
<td>Descriptive</td>
<td>title</td>
<td>coverage, date_issued_display, date_orig_display, description, geo_code, language, media_type, series, subjects, subject_local, title_alt, type</td>
<td>date_digital, duration, color_bw, dimensions, dimensions_digital, physical_description, sound,</td>
</tr>
<tr>
<td>Technical (digital only)</td>
<td>file_display, file_archive, general_technical_information, file_compression_archive, file_ppi_archive, file_quality_archive, file_scan_hw_archive, file_scan_sw_archive, file_type_archive, file_size_display, file_size_archive, internal_file_path, file_bit_depth, sampling_rate_audio, sampling_rate_video, sampling_ratio, codec, file_format, tech_info_file, tech_info_preservation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Decision making issues

• Should every piece of metadata be included in the ontological model?
  • Can this ontological model replace the traditional relational database for all metadata functions (inventorying, managing, organizing, and curating special collections)?

• If not, what metadata elements should be included in the ontological model?
  • What is exactly the purpose of this ontological model?
    • Increase interactivity and discoverability of metadata and objects in collections
    • Administrative metadata are local and not necessarily included in the model, but there should be a way to link to them automatically when necessary
Approaches to ontological modeling

• Principles:
  • No creation of new metadata, but utilizing what have been created
  • Select only the elements that would facilitate interactivity and discoverability
  • Sustainable, meaning the process can be automated to reduce reinventing the wheel while maintaining consistency
  • Use an infrastructural approach whenever possible

• Steps:
  • Understand humanities scholars’ needs and the ways of using special collections
  • Collect information about previous models
  • Analyze metadata elements and structures to identify classes and properties
    • Reusable classes and properties from existing models?
    • What new classes are absolutely needed?
  • Tune the model based on test with examples
Models that already exist

CIDOC-CRM credit: https://slideplayer.com/slide/7984641/

- Overly sophisticated models can make automatic metadata conversion very difficult
- Stay focused: interactivity and discoverability

Record in Context – Conceptual Model (RiC-CM)
Reusable classes and properties from existing models

• Criteria for external classes and properties to be “reusable”
  • Able to enrich semantics for content
  • Able to establish meaningful links between entities
  • Able to map to existing metadata descriptions

• How to keep the criteria consistent – documentation, documentation, documentation
The initial Linked Archive Model
Revised Linked Archives Model
Agent’s role types determine relation types

Agent \(\rightarrow\) creates \(\rightarrow\) Collection/Object
Agent \(\rightarrow\) participates in \(\rightarrow\) Collection/Object
Agent \(\rightarrow\) is_subject_of \(\rightarrow\) Collection/Object

Collection level:
- Author
- Subject

Object level:
- Author
- Interviewee
- Appearance
- Show host
- Photographer
- Composer
- Performer
- Contractor

Collection level:
- Repository
- Creator
- Processor

Object level:
- Subject
- Publisher

Agent à creates à Collection/Object
Agent à participates in à Collection/Object
Agent à is_subject_of à Collection/Object
Any class in the model may serve as subject.
Collections

- George Goodwin Collection Relating to Marcel Breuer
- St. Luke's Church Building Committee Records
- Herbert Beckhard Papers
- Marcel Breuer Papers

Mapping between model and metadata:

- Subject
  - is_about: Collection
  - is_related_to: Object

Collections

- George Goodwin
  - interviewed: Mary Hagerty
  - about: Marcel Breuer
  - Title of an audio recording object

- Marcel Breuer Papers
- St. Luke's Church Building Committee Records
- Herbert Beckhard Papers
Event as a type of subject

Event name: Iran hostage crisis
Date range: 1980-1981
Person: (lots)
Organization: (countries, national agencies, and other institutions)

Nightline: Iran: Day 444

100+ Nightline TV programs

It is now day 444 for the hostages in Iran. Last minute financial problems still delay the release of the hostages. On this special edition of Nightline, we'll have the details of what's causing the hold-up; we'll have reports from Washington and West Germany; and we'll talk live to a former U.S. ambassador to Iran and with other experts about this final act in the hostage drama. Includes commercials.
Transforming metadata descriptions into Linked Data

- Person class
- Organization class
- Subject class: OCLC FAST categories provide a great source for mapping subject terms in the metadata already created
- Geolocation class: can be enriched by GIS data and LC Geographic Areas terms
- Collection, object, and item class: the identifiers will be indicative of each of these levels

LC or local authority names enriched by role
Conclusion and next step

• There are various ontological models for linked data transformation, but not a single one can meet all of our requirements.

• Reuse of classes and properties in relevant models needs further exploration.

• Transformation from relational database to linked data format requires careful mapping, likely refining the model.

• Linked archives is an effort of reusing components from existing models for enriching the semantics in metadata descriptions to achieve the goal of discoverability and interactivity.
Thank you! Questions?