Metadata migration to Islandora: Is there an easy way?

6-29-2014

Sai Deng
University of Central Florida, sai.deng@ucf.edu

Lee Dotson
University of Central Florida, lee.dotson@ucf.edu

Find similar works at: https://stars.library.ucf.edu/ucfscholar

University of Central Florida Libraries http://library.ucf.edu

Part of the Cataloging and Metadata Commons

Original Citation


This Conference Presentation is brought to you for free and open access by STARS. It has been accepted for inclusion in Faculty Scholarship and Creative Works by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
Metadata Migration to Islandora: Is There An Easy Way?

Sai Deng, Lee Dotson
University of Central Florida Libraries
What will be covered?

- A Fast Intro to MODS
- How does Islandora ingest data
- UCF Collections and Metadata Migration to Islandora
  - DC-MODS-MARC Transformation Challenges
  - Options to Improve Metadata: XSLT Stylesheet, MODS Editing
  - The Tool: Notepad++
  - Pattern Identification and MODS Editing
  - Issues to Think About
- Final thoughts
A Fast Intro to MODS

- **MODS**: Metadata Object Description Schema ([http://www.loc.gov/standards/mods/](http://www.loc.gov/standards/mods/))

- **MODS Elements and Attributes (Version 3.4)** ([http://www.loc.gov/standards/mods/mods-outline-3-4.html](http://www.loc.gov/standards/mods/mods-outline-3-4.html))
  - **Root Elements**: mods, modsCollection
  - **Top level elements**: titleInfo, name, typeOfResource, genre, originInfo, language, physicalDescription, abstract, tableOfContents, targetAudience, note, subject, classification, relatedItem, identifier, location, accessCondition, part, extension, recordInfo
  - **Attributes and subelements**: for example, name has attributes: ID, type, authority etc. and subelements: namePart, affiliation, role etc.

- **Conversion**
  - **LC MODS mappings**
    - DC to MODS 3.2/3.3/3.4 ([http://www.loc.gov/standards/mods/dcsimple-mods.html](http://www.loc.gov/standards/mods/dcsimple-mods.html))
    - MARC to MODS 3.4 ([http://www.loc.gov/standards/mods/mods-mapping-3-4.html](http://www.loc.gov/standards/mods/mods-mapping-3-4.html)) (3.5 available)
    - MODS 3.4 to MARC ([http://www.loc.gov/standards/mods/v3/mods2marc-mapping.html](http://www.loc.gov/standards/mods/v3/mods2marc-mapping.html))

- **LC MODS XSLT 1.0 Stylesheets (2.0 available)**
  - DC to MODS 3.4 ([http://www.loc.gov/standards/mods/v3/DC_MODS3-4_XSLT1-0.xsl](http://www.loc.gov/standards/mods/v3/DC_MODS3-4_XSLT1-0.xsl))
  - MARCXML to MODS 3.4 ([http://www.loc.gov/standards/mods/v3/MARC21slim2MODS3-4.xsl](http://www.loc.gov/standards/mods/v3/MARC21slim2MODS3-4.xsl))
  - MODS 3.4 to MARCXML ([http://www.loc.gov/standards/mods/v3/MODS3-4_MARC21slim_XSLT1-0.xsl](http://www.loc.gov/standards/mods/v3/MODS3-4_MARC21slim_XSLT1-0.xsl))
Islandora Migration @ UCF Libraries

- **Islandora** is an open source Digital Asset Management System (DAMS) based on FedoraCommons (as base repository), Drupal (as front-end CMS) and solr (as discovery application).

- **Islandora FLVC (Florida Virtual Campus)**
  (https://islandora.pubwiki.fcla.edu/wiki/index.php/Main_Page)

- FL-Islandora is the FLVC instance hosted for the state universities and state colleges in a collaborative environment. All decisions are made collaboratively.

- **Islandora Collection and Metadata Migration @ UCF**
  - **DigiTool collections:** Political & Rights Issues & Social Movements (PRISM, available in DC format), Florida Heritage (MARCXML available)
  - **CONTENTdm collections:** University Archives, Civil War, Harrison Buzz Price, Theses and Dissertations, UCF Community Veterans History... (in CONTENTdm DC format)
  - How to prepare and work with metadata in this process?
FL-Islandora Ingest

- Online ingest: choose content model and metadata form in Islandora GUI.
  - Content models:
    - Unitary (one primary file, e.g. PDF files, single images);
    - Compound (child objects and parent objects, for any type of related objects to display together; parent objects containing metadata only);
    - Paged (hierarchical, for books, newspapers, can contain, e.g. newspaper, issue, page)
  - The metadata form is created by FLVC;
  - Librarians can manually input in the form; the form can pre-populate data through uploading a MARCXML file or from a template.

- Batch import (Zip file importer): ingest a batch of objects through the online interface.
  - Need to prepare zip file (zip the MODS metadata records and content files together, has strict rules)
FL-Islandora Ingest (Cont.)

- Offline ingest: FTP content to server and program handles ingest.
  - MODS record, content files and instructions called "manifest" packaged together; program handles the FTPed content;
  - Content can originally be from other systems, e.g. DigiTool, CONTENTdm. The ingested MODS files can be transformed from DC or MARCXML metadata;
  - DC to MODS, MARC to MODS XSLT stylesheets are involved in the conversion;
  - Offline ingest is FLVC programming, and it is not completed yet.

- FL-Islandora has been customized by FLVC and it differs from a standard Islandora installation.

- Read more at: https://islandora.pubwiki.fcla.edu/wiki/index.php/Main_Page
  (Find and click "FL-Islandora Documentation" Google doc)
FL-Islandora Ingest (Cont.)

Final display and results:
- Brief and full record display (with file/page/image);
- MARC display (only to logged in users);
- Provide downloadable MARCXML.

FLVC Stylesheets:
- FLVC DC to MODS
  (https://github.com/FLVC/FLVC_Islandora_XSLTs/blob/master/DC2MODS_DigiTool.xsl)
- FLVC MARC to MODS
  (https://github.com/FLVC/FLVC_Islandora_XSLTs/blob/master/MARC21slim2MODS3-4_FLVC.xsl)
- FLVC MODS to MARC
  (https://github.com/FLVC/FLVC_Islandora_XSLTs/blob/master/MODS3-4_MARC21slim_XSLT1-0_FLVC.xsl)
- MARC to HTML XSLT (for web display)
DC-MODS-MARC Transformation Challenges

- DC-MODS-MARC transformation presents great challenges:
  - Mapping is never perfect, or adequate (not in both ways).
    - The source schema (DC) has a very generic data representation;
      - DC: Only 15 elements. Has Qualifiers (but no "official" QDC-MODS stylesheet? Difficult to have one too)
  - The intermediate schema (MODS) cannot get enough values from DC for elements, sub-elements and attributes;
    - MODS: 20 top level elements, many sub-elements and attributes, more granular than DC
  - The target schema (MARC) does not really fit into the non-ILS type digital library environment? To keep it to enable better data sharing with an ILS?!
    - MARC 21: More granular than MODS, can have multiple MARC elements for a single MODS element
**DC-MODS-MARC Transformation Challenges (Cont.)**

○ Data ambiguity in DC-MODS mapping, for example,

<table>
<thead>
<tr>
<th>DC</th>
<th>MODS</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td><code>&lt;titleInfo&gt;&lt;title&gt;</code></td>
<td>Subtitle undistinguished. Non-filling characters, part name and number unmarked. Title type unclear: alternative, uniform...</td>
</tr>
<tr>
<td>Creator</td>
<td><code>&lt;name&gt;&lt;namePart&gt;&lt;role&gt;&lt;roleTerm type=text&quot;/&gt;creator...</code></td>
<td>Creator type unclear: personal, corporate, conference. Sub-elements value unclear: <code>&lt;affiliation&gt;, &lt;role&gt;, &lt;description&gt;</code>...</td>
</tr>
<tr>
<td>Contributor</td>
<td><code>&lt;name&gt;&lt;namePart&gt;</code></td>
<td>Name type unclear. Sub-elements value unclear.</td>
</tr>
</tbody>
</table>

○ Data ambiguity in MODS-MARC mapping, for example,

<table>
<thead>
<tr>
<th>MODS</th>
<th>MARC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;title&gt;</code> with no <code>&lt;titleInfo</code> type attribute <code>&lt;subTitle&gt;</code>...</td>
<td>245 $a with ind1=1 245 $b</td>
<td>Title types (alternative, translated, uniform...) undistinguished; Unable to distinguish subtitle (if source data is from DC).</td>
</tr>
<tr>
<td><code>&lt;name&gt;</code> with no type attribute</td>
<td>720 ind1=blank ind2=blank (uncontrolled)</td>
<td>Authority control is not specified (when names are controlled). Cannot distinguish 1XX/7XX.</td>
</tr>
<tr>
<td><code>&lt;name&gt;</code> with type=&quot;personal&quot; roleterm &quot;creator&quot;</td>
<td>100 ind1=1 ind2=blank</td>
<td>Not able to get name types from DC. Type value can be added. However, DC can have 1 creator, 2 or more creators in one record; when mapped to MARC, 100/700 can be problematic (unless we only take the first creator as the main author and define it in the stylesheet).</td>
</tr>
<tr>
<td><code>&lt;name&gt;</code> with type=&quot;personal,&quot; no roleterm &quot;creator&quot;</td>
<td>700 ind1=1 ind2=blank</td>
<td>The main author or all author names might not have the roleterm &quot;creator.&quot; When mapping to MARC, the 100/700 relationship is not clear.</td>
</tr>
</tbody>
</table>
DC-MODS-MARC Transformation Challenges (Cont.)

- Element relationship (DC-MODS-MARC conversion)
  - Relationship between different authors: cannot distinguish between 1XX (main entry, such as 100 Personal Name, 110 Corporate Name, 111 Meeting Name), and 7XX (added entry, such as 700, 710, 711);
  - Author and title relationship: cannot generate correct indicators for 1XX and 245 (title) based on whether the author is the main entry;
  - Subject relationships: unable to divide geographic, temporal, topical and genre subdivisions.

- Local elements
  - Many local elements for the collections (to be discussed. e.g. for University Archives, Civil War, Harrison Buzz Price, UCF Community Veterans History...)
  - Can add local elements extensions, but it can cause difficulty in data sharing across collections and systems.
DC-MODS-MARC Transformation Challenges (Cont.)

- The LC DC-MODS and MODS-MARC stylesheets are impossible to produce adequate or desirable output (in nature), but what can be done to improve the results?
  - On the consortia level
    - DC-MODS
      - PURL (check for PURL, add one if none exist) (FLVC implemented)
      - Local elements: DigiTool pid (FLVC implemented)
      - ETD elements (to be added as a MODS extension)
  - MODS-MARC: Many changes have already been made to the LC stylesheet by FLVC before the UCF library started its migration, e.g.
    - Map MODS issuance to MARC 250, map typeOfResource to 998 77 $b (so will be Mango format facet), map sublocation to 852 $b, punctuate 260 and all main and added entry fields.
    - See: https://github.com/FLVC/FLVC_Islandora_XSLTs/blob/master/MODS3-4_MARC21slim_XSLT1-0_FLVC.xsl

- Read more and check updates at: https://islandora.pubwiki.fcla.edu/wiki/index.php/Main_Page

- Individual libraries can have their local stylesheets.
PRISM Collection at UCF Libraries

- Political & Rights Issues & Social Movements (PRISM) collection: Part of the PALMM (Publication of Archival Library & Museum Materials).

A PRISM record in DigiTool: http://purl.flvc.org/FCLA/DT/671335
The DigiTool DC File

Wrapped in `<xb><mds><md><value>` in the DigiTool generated file

dc:title (1. If subtitle available, will be in the same field)
dc:creator (5. Their relationship unknown, name year and role undifferentiated)
dc:subjects (2. Subdivisions not available. 5 empty subject fields can be disregarded)
dc:description (physical description is not distinguished from general description)
dc:publisher (place and publisher mixed together)
dc:date (unclear what type of date)
dcterms:physicalLocation (impossible to distinguish location and call number)
Issues:
Subtitle is not marked up (when there is a subtitle);
Non-filling characters not marked up;
Name type is not available;
Relationship among names unclear;
Author/name year is not marked up;
Author/name role is not marked up;
When there are more than one creator, it creates problem in MARC display (multiple 1XXs; 1XX/7XX unclear).
Issues:
Place and Publisher cannot be distinguished in <publisher>;
Physical description is under <note>;
Subject subdivisions not available;
Links may need some explanation (for the public);
PhysicalLocation has mixed information.
A MARC display generated from the initially converted MODS (through MODS-MARC stylesheet)

A question: Which way is preferable: a generic MODS-MARC mapping (such as all names mapped to 720, uncontrolled name), or a granular mapping (names mapped to 100/110/111/700/710/711, but with only the majority of the records correctly marked up based on data patterns prior to manual editing)?
Issues (in the initial test):

- Information not mapped to the most accurate field:
  - The main author/creator is mapped to 720, which should go to 100. (LC DC-MODS-MARC stylesheets map all names to 720 to accommodate all types of names, e.g., controlled and uncontrolled, person/corporate/conference)
  - Additional authors/creators and contributors are mapped to 720, which should go to 700. (Impossible for LC DC-MODS stylesheet to distinguish between main entries and added entries)
  - Physical descriptions mapped to note (500 etc.), but not 300. (Physical and other descriptions are not distinguished in DC.)

- Indicator problems:
  - Non filling characters need to be identified. (LC DC-MODS-MARC stylesheets do not take care of non-filling characters)
  - First indicator for 245, 100, 700 need to be corrected. (LC DC-MODS-MARC stylesheets do not deal with the relationship between title and author, main entry and added entries)

- Subfield problems:
  - The publisher place should be mapped to 260 |a, not |b. (LC DC-MODS-MARC stylesheets do not distinguish publication place from the publisher)
  - For all the creators/authors/names, the year range need to be under |d, role needs to be under |c.
  - The stylesheets are unable to distinguish the subtitle (245 |b) from title (245 |a).

- Inadequate information:
  - The collection name needs to be preserved (if possible) in 830 0.
  - The stylesheets are unable to produce a 245 |c. Even we may bypass 245 |c due to the inadequate dc records info, we need to correct 245 |b.
Options to Improve Metadata: XSLT customization

- Can have customized XSLT stylesheets, but it will require an environment to test and run the code on the data source, and preferably, a programmer.

- DC-MODS and MODS-MARC stylesheets Adjustments (Some possible ideas)
  - Adjust the main entry and added entry (1XX, 7XX) logic. Map the first name to 1XX (Element Positioning), other names to 7XX. However how do you know if it’s a person (100/700), an organization (110/710) or a conference (111/711)?
  - Apply data patterns and automatically distinguish and mark up personal, corporate and conference name types (may cover the majority of the situations, but impossible to make all records right. Need review);
  - Generate mark-up code for name year and role based on data patterns for year and role terms;
  - Generate mark-up code for subtitles based on whether a ":" exists in the title;
  - Generate mark-up code for publication place based on whether a ":" exists in the publisher field;
Options to Improve Metadata: XSLT customization (Cont.)

- To identify **field relationships** and have the right indicators for 100 and 245 based on "if an author exists."
- Other adjustments: mark up **non-filling characters**...
- **Need comprehensive review of the results!**
- Some of these adjustments cannot produce one hundred percent correct results (because not all data can fall into a pattern); it might not be realistic to apply them to a national or consortia level stylesheet. They seem to be more feasible on a library collection level. However, what do you think?

- **MARC-MODS, MODS-MARC stylesheets**
  - MARC-MODS conversion produces much better results (e.g. for the Florida Heritage collection in DigiTool). However, a question on local subjects was raised because MARC-MODS XSLT does not take local subjects (690).

- **Subject mapping change (by FLVC)**
  - **MARC-MODS**: Mapped local 690 to `<subject><topic>`
  - Added subject authority "sears“: 6XX _8 mapped to `<subject authority="sears">`
  - **MODS-MARC**: `<subject authority="sears">` mapped to 6XX _8
Options to Improve Metadata: MODS Editing

- Post-transformation MODS Editing
- FLVC recommended Notepad++ for MODS editing

- **Notepad++**
  - Free text and source code editor for Windows
    - Download Executable zip file and install

- **Some characteristics:**
  - Set different languages (XML, HTML, CSS, PHP, Java...)
  - Syntax color coding
  - Tree view of the files and folders
  - Enable tag autocompletion (in set preference)
  - Can open, save and close multiple files at one time (save all, close all...)
  - Plug in: download plug in, such as ftp, XML tools
  - Editing: Find, replacement, can use regular expressions, macros...
  - Check xml for well-formedness, validate against xml schema

- PRISM collection: Edited and marked up 847 MODS XML records using Notepad++ which were generated from Dublin Core (DC).
MODS Editing: Distinguish between the Main Entry and Added Entries

- When there are more than one creator (\textlt{<name><role><roleTerm> with type="text" value "creator," or, <name><role><roleTerm> with type="code" value "cre"}), there can be multiple 100s. However 100 is not repeatable.

- Only keep the first creator marked up with "creator" or "cre" in MODS file, and change the other name types to "contributor." Or just leave the other names’ types blank (depending on the logic in the stylesheet).

Find what: \textlt{</roleTerm>}

When there are more than "1 hit" (1 creator) in a record, keep the first creator "creator," change the others to "contributor" or leave them un-marked-up.
MODS Editing: Identify Pattern and Mark Up Data (Year)

- **Year pattern:** `yyyy-yyyy; yyyy- yyyy; yyyy-; yyyy; b. yyyy`
  (Find pattern by searching `<namePart>` in Notepad++ and sort the results in a spreadsheet)

- Mark-up editing, for example:
  
  ```xml
  <name>
      <namePart>De Leon, Daniel 1852-1914</namePart>
      ...
  </name>
  
  Replace with:
  
  ```xml
  <name>
      <namePart>De Leon, Daniel</namePart>
      <namePart type="date">1852-1914</namePart>
      ...
  </name>
  ```
MODS Editing: Identify Pattern and Mark Up Data (Year, Cont.)

- Use Regular Expression in Notepad++ to perform find/replacement function and mark up the year;
- ( ) is used to tag a match. Tagged matches can be referred to as \1, \2 etc.

Find What: ([0-9]+)-([0-9]+)</namePart>
Replace with: </namePart><namePart type="date">\1-\2</namePart>
MODS Editing: Identify Pattern and Mark Up Data (Year, Cont.)

- `<namePart>Vail, Charles H. (Charles Henry) b. 1866</namePart>`
  Replaced With:
  `<namePart>Vail, Charles H. (Charles Henry)</namePart><namePart type="date">b. 1866</namePart>`

Find What: `b. ([0-9]+)`
Replace with: `</namePart><namePart type="date">b. \1</namePart>`
MODS Editing: Identify Pattern and Mark Up Data (Role Terms)

- **Check pattern** (click "Find All in All Opened Documents")

Find What:  

```xml
  <namePart>
  </namePart>
```
MODS Editing: Identify Pattern and Mark Up Data (Role Terms, Cont.)

- Copy the result to a spreadsheet

- Sort the value in the second column to see patterns
MODS Editing: Identify Pattern and Mark Up Data (Role Terms, Cont.)

- **Author Role Terms Pattern**

<table>
<thead>
<tr>
<th>Role terms</th>
<th>Role terms (after being controlled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>defendant</td>
<td>defendant</td>
</tr>
<tr>
<td>illus.</td>
<td>illustrator</td>
</tr>
<tr>
<td>- illus</td>
<td>illustrator</td>
</tr>
<tr>
<td>joint author</td>
<td>author</td>
</tr>
<tr>
<td>ed</td>
<td>editor</td>
</tr>
<tr>
<td>joint ed</td>
<td></td>
</tr>
<tr>
<td>editor</td>
<td></td>
</tr>
<tr>
<td>- editor</td>
<td></td>
</tr>
<tr>
<td>translator</td>
<td>translator</td>
</tr>
<tr>
<td>- translator</td>
<td></td>
</tr>
<tr>
<td>- translator</td>
<td></td>
</tr>
<tr>
<td>tr (at the end)</td>
<td></td>
</tr>
</tbody>
</table>

- Clean up and standardize the terms (find/replace);
- Edit the mark up, for example:

  Replace:
  ```xml
  <name>
    <namePart>McMillan, Hugh translator</namePart>
  </name>
  ```

  With:
  ```xml
  <name>
    <namePart>McMillan, Hugh</namePart><role><roleTerm type="text" authority="marcrelator">translator</roleTerm></role>
  </name>
  ```
MODS Editing: Identify Pattern and Mark Up Data (Role Terms, Cont.)

- In Notepad++
  
  Find what: translator</namePart>
  Replace with: </namePart><role><roleTerm type="text" authority="marcrelator">translator</roleTerm></role>

Please notice that: when there is a role term after "yyyy-yyy," the year could not be marked up in the previous "year" replacement. Can replace it after the role term is marked up.
MODS Editing: Identify Pattern and Mark Up Data (Names)

- Distinguished between personal names, corporate names and conferences;
  - Find: <namePart>, copy the result to a spreadsheet, sort data and identify some patterns;
  - What’s in spreadsheet:
<table>
<thead>
<tr>
<th>Name Pattern (if &lt;namePart&gt; contains)</th>
<th>Add Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Party, Partei, partia, partiï</td>
<td>&lt;name type=&quot;corporate&quot;&gt;</td>
</tr>
<tr>
<td>Committee</td>
<td></td>
</tr>
<tr>
<td>International</td>
<td></td>
</tr>
<tr>
<td>Council</td>
<td></td>
</tr>
<tr>
<td>National</td>
<td></td>
</tr>
<tr>
<td>agency</td>
<td></td>
</tr>
<tr>
<td>Service</td>
<td></td>
</tr>
<tr>
<td>School</td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
</tr>
<tr>
<td>Association</td>
<td></td>
</tr>
<tr>
<td>society</td>
<td></td>
</tr>
<tr>
<td>department</td>
<td></td>
</tr>
<tr>
<td>Mission</td>
<td></td>
</tr>
<tr>
<td>Union</td>
<td></td>
</tr>
<tr>
<td>League</td>
<td></td>
</tr>
<tr>
<td>bund</td>
<td></td>
</tr>
<tr>
<td>scientists of</td>
<td></td>
</tr>
<tr>
<td>Conference</td>
<td>&lt;name type=&quot;conference&quot;&gt;</td>
</tr>
<tr>
<td>Congress</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>&lt;name type=&quot;personal&quot;&gt;</td>
</tr>
</tbody>
</table>

Not comprehensive. Need to check the results!!
MODS Editing: Identify Pattern and Mark Up Data (Names, Cont.)

- Find all `<namePart>` elements which contain "Committee"

Find what: `<name>\n   <namePart>.*Committee.*/</namePart>`
MODS Editing: Identify Pattern and Mark Up Data (Names, Cont.)

Find what: `<name>
    <namePart>(.*Committee.*)</namePart>`
Replace with: `<name type="corporate">
    <namePart>1</namePart>`

(Note: will only work when the corporate name is the first name. May need to manually add: type="corporate" for <name>.)
After checking corporate and conference names, replace all remaining names:

```
<name>
  <namePart>

With

<name type="personal">
  <namePart>
```

Need to review all the names!!
Make edits when needed.

MODS Editing: Identify Pattern and Mark Up Data (Names, Cont.)
MODS Editing: Identify Pattern and Mark Up Data (Subtitle)

- Find all titles with subtitles (find “:” in title field).

Find what: `<title>.*:.*</title>`

It returns 225 hits.
For example:

Replace
<titleInfo>
<title>Socialism: a paper read before the Albany Press Club "Socialist Night"</title>
</titleInfo>

With:
<titleInfo>
<title>Socialism: <subtitle> a paper read before the Albany Press Club "Socialist Night"</subtitle>
</title>
</titleInfo>

May leave out “:” in title.

Find what: <title>(.*):(.*)</title>
Replace with:
<title>\1:<subtitle>\2</subtitle></title>
For example:

- Is this a war for freedom?  
- A key to survival  
- An alternative to war  
- The theory of the Cuban Revolution  

Find what: <title>The  
Replace with:  
<title><nonSort>The </nonSort>
MODS Editing: Identify Pattern and Mark Up Data (Publication Place)

- Identify publication place pattern

  Find what:
  <publisher>(.*):(.*)</publisher>

- Publisher place pattern, for example:
  "New York
  [New York
  [New York]
  New York City
  New York, [N.Y.]
  New York, N.Y.
  New York, NY
  (all need to change to New York, N. Y.)
For example:

Replace

<publisher>New York : New York Labor News</publisher>

with

<publisher><place><placeTerm type="text">New York</placeTerm></place> New York Labor News</publisher>

Find what:
<publisher>(.*):(.*)</publisher>

Replace with:
<publisher><place><placeTerm type="text">\1</placeTerm></place> New York, N.Y.</publisher>

Need to standardize and clean the place terms before the mark-up replacement, e.g. change "New York" to "New York, N.Y."
MODS Editing: Identify Pattern and Mark Up Data (Subject)

- Mark up topical, temporal, geographic, genre subdivisions for subjects, and personal/corporate/conference name subjects
- Subdivision patterns: Use test set to find pattern (don’t mess with the “edit-in-progress” set)

Find what: `<topic>(.*)--(.*)</topic>`
Replace with `<topic>--\2</topic>`

Then find "<topic>--(.*)</topic>" to get subdivisions only, copy to a spreadsheet to identify patterns.
Total 1682 subdivisions.  
May find and replace/mark up some popular terms.

Find some popular terms: 
Subdivisions in column A, 
In B1, add =COUNTIF($A:$A,A1) 
Copy down 
Then sort by column B to find the times a word appears.

Top occurred subdivisions:
- United States 200
- Soviet Union 90
- Socialism 69
- Communism 52
- History 35
- Politics and government 34
- Communist Party of the United States of America 25
- Disarmament 20
- Congresses 19
- Marx, Karl 1818-1883 18
- Peace 17
- 1917-1945 15
MODS Editing: Identify Pattern and Mark Up Data (Subject, cont.)

○ Examples

Replace:
<subject>
  <topic>Labor -- United States</topic>
</subject>
With:
<subject><topic>Labor</topic><geographic>United States</geographic></subject>

Replace:
<subject>
  <topic>Labor unions -- Great Britain -- History</topic>
</subject>
With:
<subject><topic>Labor unions</topic><geographic>Great Britain</geographic><topic>History</topic></subject>

Replace:
<subject>
  <topic>Soviet Union -- Economic conditions -- 1917-1945</topic>
</subject>
With:
<subject><geographic>Soviet Union</geographic><topic>Economic conditions</topic><temporal>1917-1945</temporal></subject>

Some batch replacement can be done to top occurred subdivisions, but comprehensive review of all subdivisions and manual editing will be needed!!
MODS Editing: Related Item, Collection Title and Physical Description

- **Add related item, e.g.**
  ```xml
  <relatedItem type="original">
    <location>
    </location>
    <titleInfo>
      <title>title.</title>
    </titleInfo>
  </relatedItem>
  
  (note: need to edit the related item title information)
  ```

- **Add collection title, e.g.**
  ```xml
  <relatedItem type="series">
    <titleInfo type="uniform">
      <title>PRISM: Political &amp; Rights Issues &amp; Social Movements collection</title>
    </titleInfo>
    <location>
      <url displayLabel="(Link to Collection)" displayLabel="https://ucf.digital.flvc.org/islandora/object/ucf%3Asmc"></url>
    </location>
  </relatedItem>
  ```

- **Physical description in note field:**
  - Change `<note>47 p. ; 22 cm.</note>` to:
    ```xml
    <physicalDescription>
      <extent>47 p. ; 22 cm.</extent>
    </physicalDescription>
    ```
MODS Editing: File Example (Cont.)

- <relatedItem type="original">
  - <location>
    <url note="[Link to Original Online Catalog Record]">http://ucf.catalog.fcla.edu/permalink.jsp?29CF000388014</url>
  </location>
- <relatedItem type="series">
  - <titleInfo type="uniform">
    <title>PRISM: Political & Rights Issues & Social Movements collection</title>
  </titleInfo>
  - <location displayLabel="[Link to PRISM Collection]">https://ucf.digital.flv.org/islandora/object/ucf%3Asmc</location>
</relatedItem>
- <location displayLabel="UCF Libraries Special Collections - 5th Floor -- HX15.V35 no.829">
  <url>http://purl.flvc.org/FCLA/DT/671335</url>
</location>
<accessCondition>All rights to images are held by the respective holding institution. This image is posted publicly for non-profit educational uses, excluding printed publication. For permission to reproduce images and/or for copyright information contact Special Collections and University Archives, University of Central Florida Libraries, (407) 823-2576. http://library.ucf.edu/SpecialCollections/</accessCondition>
<typeOfResource>text</typeOfResource>
A Resulted MARC Display (only to logged in users)
Record Public Display
Metadata Migration Issues

- Is there another way to improve the metadata when migrating from a less granular schema to a more granular one besides customizing the XSLT stylesheet and performing post-conversion pattern based records editing? What are the advantages and disadvantages of these methods?

- How to determine the extensibility a stylesheet can be customized prior to post transformation editing in a project?

- Should data mark-up based on patterns and conditions be dealt with in the stylesheet or the post-transformation editing? Which factors affect the decision?

- Does MARCXML data play an important role in the non-Integrated Library System (ILS) environment? Can auto-generated pseudo MARC records be accepted in a CMS or DAMS? To which degree should the converted MODS and MARC records be edited?

- In mapping local elements to MODS, how to determine whether a MODS extension is considered good practice?

- How much automation will a MODS editing tool allow? How much manual or semi-manual editing work is realistic for librarians? What skills are required to perform the task?
Final Thoughts

- There might not be an easy way to get very good data in migrating from a less granular schema to a more granular one. To which degree and in which method should or can the data be improved may depend on many factors.

- The decisions on metadata migration will be different depending on whether the project is on national, consortia or library level.

- Extensive stylesheet customization seems to be more attainable on the collection level when comparable data is available.

- Balance needs to be sought in pre-transformation stylesheet modification and post-transformation records editing.

- It still seems relevant to share collections in a non-traditional DAMS with a traditional ILS in the current library environment.

- Medium or large scale text markup and encoding presents new challenges and requirements for librarians especially cataloging and metadata librarians.
Project Team and Acknowledgments

- Sai Deng, Metadata Librarian, University of Central Florida, sai.deng@ucf.edu
- Lee Dotson, Digital Initiatives Librarian, University of Central Florida, lee.dotson@ucf.edu
- Mary Page, Associate Director, Collections & Technical Services

Special thanks go to FLVC folks. Inquires can be sent to FLVC Help Desk, help@flvc.org

Thank you!