Authority Control for Digital Collections

Jeremy Myntti
Head of Cataloging & Metadata Services

Nate Cothran
Vice President, Automation Services
Why Authority Control?

- Vocabulary control
- Consistency
- Standardization
- Updating mechanism
Inconsistency

Gosiute Indians
Goshute Indians
Navajo Indians
Navaho Indians
Salt Lake
Salt Lake City
Salt Lake City (Utah)
Bishop, Dail Stapley
Bishop, Dale Stapely
Bishop, Dale Stapley

Beckwith, Frank A. (1876-1951)
Beckwith, Frank Asahel (1876-1951)
Beckwith, Frank A.
Beckwith, Frank A. (1876-1951)
Beckwith, Frank Asahel (1876-1951)
Beckwith, Frank Asahel, 1876-1951

Hansen, Henrie
Hansen, Henry
Hansen, Henry Daniel, 1896-1979
Partnership

J. Willard Marriott Library
The University of Utah

Backstage Library Works
aDo geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos - Never odd or even. 0aPalindromes.
A man, a plan, a canal -- Panama!  c1991.  a303 p.; ccm.

Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even. 0aPalindromes.
001 ovld000000001
...
100 1 $a Mercer, Leigh, $d 1893-1977.
245 12 $a A man, a plan, a canal -- Panama!
...
505 $a Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even.
650 0 $a Palindromes.
MARC as a Container

Portable

Entrenched Tools
MARC Display

001 ovld000000001
...
100 1 $a Mercer, Leigh, $d 1893-1977.
245 12 $a A man, a plan, a canal -- Panama!
...
505 $a Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even.
650 0 $a Palindromes.
MARC XML

<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://www.loc.gov/MARC21/slim">
  <record>
    <leader>00491nam 22001571 4500</leader>
    <controlfield tag="001">ovld0000000001</controlfield>
    ...
    <datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Mercer, Leigh</subfield>
      <subfield code="d">1893-1977.</subfield>
    </datafield>
    <datafield tag="245" ind1="1" ind2="2">
      <subfield code="a">A man, a plan, a canal -- Panama!</subfield>
    </datafield>
    ...
    <datafield tag="505" ind1=" " ind2=" ">
      <subfield code="a">Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even.</subfield>
    </datafield>
    <datafield tag="650" ind1=" " ind2="0">
      <subfield code="a">Palindromes.</subfield>
    </datafield>
  </record>
</collection>
XML Defined

- XML stands for eXtensible Markup Language
- XML is similar to HTML, but also different
- XML was designed to carry data, not display data
- XML tags are not pre-defined; you must define tags
- XML is designed to be self-descriptive
- XML is a W3C recommendation
XML Invented

- There are no defined XML tags (title = <title>)
- Tags in HTML are defined (<p> = paragraph)
- Anyone can design their own XML schema
- Common library-centric XML schemas include:
  - MODS, MARC XML, EAD, Dublin Core, ONIX, XC
  - METS Alto, TEI, BIBFRAME / RDF, CONTENTdm
  - All of these either follow their own schema or borrow elements from each other
XML Structure, Tree

- Increases in complexity
  - Root
  - Branches
  - Leaves
XML Structure, Family

- Parents > Children
- Children : Siblings
- Children > Attributes
- Attributes : Values
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://www.loc.gov/MARC21/slim">
<record>
<leader>00491nam 22001571 4500</leader>
<controlfield tag="001">ovld000000001</controlfield>
...
<datafield tag="100" ind1="1" ind2=" ">
<subfield code="a">Mercer, Leigh,</subfield>
<subfield code="d">1893-1977. </subfield></datafield>
<datafield tag="245" ind1="1" ind2="2">
<subfield code="a">A man, a plan, a canal -- Panama!</subfield></datafield>
...
<datafield tag="505" ind1=" " ind2=" ">
<subfield code="a">Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even. </subfield></datafield>
<datafield tag="650" ind1=" " ind2="0">
<subfield code="a">Palindromes.</subfield></datafield>
</record>
</collection>
MARC XML, Parent / Child

<?xml version="1.0" encoding="UTF-8"?>
<root xmlns="http://www.loc.gov/MARC21/slim">
  <parent>
    <child>00491nam 22001571 4500</child>
    <child att="001">ovld000000001</child>
    ...
    <child att="100" att="1" att=" ">
      <child att="a">Mercer, Leigh, </child>
      <child att="d">1893-1977.</child>
    </child>
    <child att="245" att="1" att="2">
      <child att="a">A man, a plan, a canal -- Panama!</child>
    </child>
    ...
    <child att="505" att=" " att=" ">
      <child att="a">Do geese see God? -- Murder for a jar of red rum -- Some men interpret nine memos -- Never odd or even.</child>
    </child>
    <child att="650" att=" " att="0">
      <child att="a">Palindromes.</child>
    </child>
  </parent>
</root>
Methodology

- Extracting data from CONTENTdm
  ◦ Stop updates to collection and make it *read-only*
  ◦ Make copy of *desc.all* metadata file for backup
  ◦ Run *desc.all* file through AC processing
  ◦ Replace *desc.all* file on CONTENTdm server
  ◦ Run the full collection index
  ◦ Remove *read-only* status from collection

- Ogburn, Joyce L.
- *Acquiring minds want to know: digital scholarship*
- 2003
- A new form of scholarship has emerged in recent years named "digital scholarship."
- *Born digital; Libraries; Electronic publishing*
- *Libraries and electronic publishing; Scholarly electronic publishing*
Normalized Headings

MARC
600 $a Smith, John, $d 1947 Apr. 16-
651 $a United States $x History $y 20th century.

600 $ SMITH, JOHN $ 1947 APR 16
651 $ UNITED STATES $ HISTORY $ 20TH CENTURY

XML
<subject>Smith, John, 1947 Apr. 16-;
United States--History--20th century;</subject>

<subject>SMITH, JOHN, 1947 APR 16;
UNITED STATES--HISTORY--20TH CENTURY;</subject>
MARC & XML, Normalized

- **MARC**
  - 651 $a United States $x History $y 20th century.
  - 651 $ UNITED STATES $ HISTORY $ 20TH CENTURY

- **XML**
  - United States--History--20th century;
  - 650 $a United States $x History $x 20th century
  - 650 $ UNITED STATES $ HISTORY $ 20TH CENTURY
MARC & XML Examples

- MARC
  - 600 $a Smith, John, $d 1947 Apr. 16-
  - 651 $a United States $x History $y 20th century.
  - 650 $a Navajo Indians $x History.
  - 650 $a Religious ceremonies.
  - 650 $a Religion.

- XML
  - Smith, John, 1947 Apr. 16-; United States--History--20th century; Navajo Indians--History; Religious ceremonies; Religion;
MARC & XML Conversions

- **MARC**
  - 600 $a Smith, John, $d 1947 Apr. 16-
  - 651 $a United States $x History $y 20th century.
  - 650 $a Navajo Indians $x History.
  - 650 $a Religious ceremonies.
  - 650 $a Religion.

- **XML**
  - 600 $a Smith, John, $d 1947 Apr. 16-
  - 650 $a United States $x History $x 20th century
  - 650 $a Navajo Indians $x History
  - 650 $a Religious ceremonies
  - 650 $a Religion
XML Authority Matching

- **XML**
  - 600 $a Smith, John, $d 1947 Apr. 16-
  - 650 $a United States $x History $x 20th century
  - 650 $a Navajo Indians $x History
  - 650 $a Religious ceremonies
  - 650 $a Religion

- **Authority**
  - 150 $a Rites and ceremonies
  - 450 $a Ceremonies
  - 450 $a Cult
  - 450 $a Cultus
  - 450 $a Ecclesiastical rites and ceremonies
  - 450 $a Religious ceremonies
  - 450 $a Religious rites
  - 450 $a Rites of passage
  - 450 $a Traditions
XML Authority Control

- XML
  - \(<\text{subject}>\text{Smith, John, 1947 Apr. 16-; United States--History--20th century; Navajo Indians--History; Rites and ceremonies; Religion;}\</\text{subject}>\)

- Variants
  - \(<\text{subject-keyword}>\text{Ceremonies; Cult; Cultus; Ecclesiastical rites and ceremonies; Religious ceremonies; Religious rites; Rites of passage; Traditions;}\</\text{subject-keyword}>\)
MARC vs XML Tags

- **MARC**
  - 600, 610, 611, 630, 650, 651, 655

- **XML**
  - `<subject>`, `<subjeca>`, `<subjec-1>`, `<subject-title>`, `<topics>`, `<entry>`, `<creata>`, `<s01>`, `<genre>`, etc
Pre-CONTENTdm Processing

<records>
  <record>
    <title>…</title>
  </record>
  <record>
    <title>…</title>
  </record>
  <record>
    <title>…</title>
  </record>
</records>

- Temporary container elements are added
Post-CONTENTdm Processing

<title>…</title>

<title>…</title>

- Temporary container elements are removed
CDM Before & After
**Other Issues**

- **Bad OCR data**
- `<descr>` or `<transc>` field elements
- Errant angle brackets: `<` or `>`
- CDATA workaround (pre-processing)

```xml
<transc>...
immunofluorescence at sites
of actin-membrane</transc>
```

- System tries to assign `<...>` as discrete child of `<transc>`
Other Issues

- Bad OCR data
- `<descr>` or `<transc>` field elements
- Errant angle brackets: `< or >`
- CDATA workaround (pre-processing)

  - CDATA blocks effectively cancel out all data between them
  - CDATA = character data

  ```xml
  <![CDATA[<transc>...
  immunofluorescence at sites of actin-membrane< f" "  
  t "U .* • ?vl /&gt; .* . O' 
  .< .*V#•••• r , * . y | •  
  ...</transc>]]>
  ```
Statistics for USpace

**Changed Records**
- Changed 35%
  - Records: 6,308

**Matched Records**
- Subjects 33%
  - Records: 15,800

**Matched Records**
- Names 28%
  - Records: 22,306

- Changed Records 35%
- Subjects 33%
- Names 28%

Records: 6,308
Subjects: 15,800
Names: 22,306
Standardization: Missing semicolons

- Smith, John, 1932- United States--History --20th century;

- Smith, John, 1932-; United States--History --20th century;
Standardization: Subdivisions

- United States--History--20th century;
- United States – History – 20th century;
- United States-History-20th century;
- United States-History--20th century;
Standardization: Dates

**Improper**

- 1980-3;
- 20130526;
- May 26, 2013;

**Proper**

- 2013-05-26;
# Reports

## Unmatched Names

<table>
<thead>
<tr>
<th>Total Records</th>
<th>Type</th>
<th>Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 record</td>
<td>other</td>
<td>Ahrens, Jim</td>
</tr>
<tr>
<td>1 record</td>
<td>other</td>
<td>Aide, T. Mitchell</td>
</tr>
<tr>
<td>47 records</td>
<td>creat</td>
<td>Ainion, David Charles</td>
</tr>
<tr>
<td>1 record</td>
<td>other</td>
<td>Aion, Michael</td>
</tr>
<tr>
<td>1 record</td>
<td>other</td>
<td>Airola, Marc B.</td>
</tr>
<tr>
<td>1 record</td>
<td>other</td>
<td>Aizawa, Shin-Ichi</td>
</tr>
<tr>
<td>1 record</td>
<td>other</td>
<td>Aizprua, Rafael</td>
</tr>
<tr>
<td>2 records</td>
<td>other</td>
<td>Akella Venkatesh</td>
</tr>
</tbody>
</table>

## Partially Matched Subjects

<table>
<thead>
<tr>
<th>Total Records</th>
<th>Type</th>
<th>Partially Matched</th>
<th>Remainder of Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 record</td>
<td>subj</td>
<td>Gene targeting</td>
<td>--History</td>
</tr>
<tr>
<td>1 record</td>
<td>subj</td>
<td>Religion</td>
<td>--Health aspects</td>
</tr>
</tbody>
</table>

## Split Subjects

<table>
<thead>
<tr>
<th>Total Records</th>
<th>Type</th>
<th>Heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 records</td>
<td>subj</td>
<td>Beauty</td>
</tr>
<tr>
<td>3 records</td>
<td>subj</td>
<td>Aesthetics</td>
</tr>
<tr>
<td>3 records</td>
<td>subj</td>
<td>Art--Philosophy</td>
</tr>
<tr>
<td>3 records</td>
<td>subj</td>
<td>Beauty, Personal</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Biomedical research</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Medicine--Research</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Biology--Research</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Birth rate</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Childbirth--Statistics</td>
</tr>
<tr>
<td>2 records</td>
<td>subj</td>
<td>Fertility, Human--Statistics</td>
</tr>
</tbody>
</table>
Near Match Report

Levenshtein-Distance Algorithm

\[
\text{lev}_{a,b}(i,j) = \begin{cases} 
\max(i,j) & \text{if } \min(i,j) = 0, \\
\min \left\{ \begin{array}{l}
\text{lev}_{a,b}(i-1,j) + 1 \\
\text{lev}_{a,b}(i,j-1) + 1 \\
\text{lev}_{a,b}(i-1,j-1) + [a_i \neq b_j]
\end{array} \right. & \text{otherwise.}
\end{cases}
\]

Near Matches

<table>
<thead>
<tr>
<th>Bib ID</th>
<th>Type</th>
<th>UNM</th>
<th>Unmatched Heading</th>
<th>%-1</th>
<th>Tag-1</th>
<th>Near Match-1</th>
<th>Authority-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ir-main, 13197, ir-ma</td>
<td>other</td>
<td>700</td>
<td>Anderson, Richard C. E.</td>
<td>90.0%</td>
<td>100</td>
<td>Anderson, Richard C.</td>
<td>n50021951</td>
</tr>
<tr>
<td>ir-main, 1704, ir-mair</td>
<td>creata</td>
<td>700</td>
<td>Angelucci, Alessandra</td>
<td>95.0%</td>
<td>100</td>
<td>Angelucci, Alessandro</td>
<td>n2012066779</td>
</tr>
<tr>
<td>ir-main, 11558</td>
<td>other</td>
<td>700</td>
<td>Annesley, Thomas M.</td>
<td>88.2%</td>
<td>100</td>
<td>Annesley, Thomas</td>
<td>n2002152176</td>
</tr>
<tr>
<td>ir-main, 566</td>
<td>other</td>
<td>700</td>
<td>Arenkel, Benjamin R.</td>
<td>89.5%</td>
<td>100</td>
<td>Arenkel, Benjamin R.</td>
<td>nr2004033027</td>
</tr>
<tr>
<td>ir-main, 1637, ir-mair</td>
<td>other</td>
<td>700</td>
<td>Armstrong, Phillip A.</td>
<td>89.5%</td>
<td>100</td>
<td>Armstrong, Phillip</td>
<td>n96107480</td>
</tr>
<tr>
<td>ir-main, 16069</td>
<td>creata</td>
<td>700</td>
<td>Ashton, Alan Conway</td>
<td>75.0%</td>
<td>100</td>
<td>Ashton, Alan Conway, 1942-</td>
<td>no2007080328</td>
</tr>
<tr>
<td>ir-main, 16084, ir-ma</td>
<td>creata</td>
<td>700</td>
<td>Aspinwall, Lisa</td>
<td>87.5%</td>
<td>100</td>
<td>Aspinwall, Lisa G.</td>
<td>n2002101864</td>
</tr>
<tr>
<td>ir-main, 1934</td>
<td>other</td>
<td>700</td>
<td>Atomic force microscope</td>
<td>95.7%</td>
<td>150</td>
<td>Atomic force microscopy</td>
<td>sh94008704</td>
</tr>
<tr>
<td>ir-main, 4316</td>
<td>other</td>
<td>700</td>
<td>Atwood, R. L.</td>
<td>90.0%</td>
<td>100</td>
<td>Atwood, R. A.</td>
<td>n82034533</td>
</tr>
<tr>
<td>ir-main, 9241</td>
<td>other</td>
<td>700</td>
<td>Aubourt, E.</td>
<td>88.9%</td>
<td>100</td>
<td>Aubourg, E.</td>
<td>no00039482</td>
</tr>
<tr>
<td>uspace, 17683</td>
<td>other</td>
<td>700</td>
<td>Austin, C. A.</td>
<td>80.0%</td>
<td>100</td>
<td>Austin, C.</td>
<td>n79116889</td>
</tr>
</tbody>
</table>
## Future Plans: Ongoing AC via Linked Data

<table>
<thead>
<tr>
<th>Heading</th>
<th>Uniform Resource Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelson, Arthur C.</td>
<td><a href="http://id.loc.gov/subjects/n88271189">http://id.loc.gov/subjects/n88271189</a></td>
</tr>
<tr>
<td>Rosati, P.</td>
<td><a href="http://id.loc.gov/subjects/n88275175">http://id.loc.gov/subjects/n88275175</a></td>
</tr>
<tr>
<td>Grant, B. Rosemary</td>
<td><a href="http://id.loc.gov/subjects/n88276300">http://id.loc.gov/subjects/n88276300</a></td>
</tr>
<tr>
<td>Mayer, Robert N.</td>
<td><a href="http://id.loc.gov/subjects/n88278123">http://id.loc.gov/subjects/n88278123</a></td>
</tr>
<tr>
<td>Ogburn, Joyce L.</td>
<td><a href="http://id.loc.gov/subjects/n88283447">http://id.loc.gov/subjects/n88283447</a></td>
</tr>
<tr>
<td>Liu, Fei</td>
<td><a href="http://id.loc.gov/subjects/n88291613">http://id.loc.gov/subjects/n88291613</a></td>
</tr>
<tr>
<td>Mathur, S. P.</td>
<td><a href="http://id.loc.gov/subjects/n88293854">http://id.loc.gov/subjects/n88293854</a></td>
</tr>
<tr>
<td>Mitchell, Murray D.</td>
<td><a href="http://id.loc.gov/subjects/n88293998">http://id.loc.gov/subjects/n88293998</a></td>
</tr>
<tr>
<td>Joy, Kenneth I.</td>
<td><a href="http://id.loc.gov/subjects/n88297793">http://id.loc.gov/subjects/n88297793</a></td>
</tr>
<tr>
<td>Shaw, M.</td>
<td><a href="http://id.loc.gov/subjects/n88605835">http://id.loc.gov/subjects/n88605835</a></td>
</tr>
<tr>
<td>Kidd, Thomas</td>
<td><a href="http://id.loc.gov/subjects/n88606392">http://id.loc.gov/subjects/n88606392</a></td>
</tr>
<tr>
<td>Werblin, Frank S.</td>
<td><a href="http://id.loc.gov/subjects/n88606934">http://id.loc.gov/subjects/n88606934</a></td>
</tr>
<tr>
<td>Henry, David J.</td>
<td><a href="http://id.loc.gov/subjects/n88609950">http://id.loc.gov/subjects/n88609950</a></td>
</tr>
<tr>
<td>Jones, David T.</td>
<td><a href="http://id.loc.gov/subjects/n88611445">http://id.loc.gov/subjects/n88611445</a></td>
</tr>
<tr>
<td>Nguyen, T. D.</td>
<td><a href="http://id.loc.gov/subjects/n88611802">http://id.loc.gov/subjects/n88611802</a></td>
</tr>
<tr>
<td>Wright, Richard</td>
<td><a href="http://id.loc.gov/subjects/n88612923">http://id.loc.gov/subjects/n88612923</a></td>
</tr>
<tr>
<td>Anderson, Paul A.</td>
<td><a href="http://id.loc.gov/subjects/n88613440">http://id.loc.gov/subjects/n88613440</a></td>
</tr>
<tr>
<td>Tol, J. van</td>
<td><a href="http://id.loc.gov/subjects/n88618201">http://id.loc.gov/subjects/n88618201</a></td>
</tr>
</tbody>
</table>
Ogburn, Joyce L.

From Library of Congress Name Authority File

Details

Ogburn, Joyce L.

URI(s)
› http://id.loc.gov/authorities/names/n88283447

Instance Of
› MADS/RDF PersonalName
› MADS/RDF Authority
› SKOS Concept

Scheme Membership(s)
› Library of Congress Name Authority File

Collection Membership(s)
› Names Collection - Authorized Headings
› LC Names Collection - General Collection

Exact Matching Concepts from Other Schemes
› http://viaf.org/viaf/sourceID/LC%7Cn+88283447#skos:Concept

Sources

Change Notes
› 1989-09-27: new
› 1989-11-28: revised

Alternate Formats
› RDF/XML (MADS and SKOS)
› N-Triples (MADS and SKOS)
› JSON (MADS/RDF and SKOS/RDF)
› HTML (MADS)
Future Plans: Ongoing AC via Linked Data

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <madsrdf:PersonalName rdf:about="http://id.loc.gov/authorities/names/n88283447">
    <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Authority"/>
    <madsrdf:authoritativeLabel xml:lang="en">Ogburn, Joyce L.</madsrdf:authoritativeLabel>
  </madsrdf:PersonalName>
  ...
</rdf:RDF>
Linked Data: Cloud Diagram

Future Plans: Legacy Collections
Other XML File Structures

**Dublin Core**
- record>recordData>dc:creator
- record>recordData>dc:subject
- record>recordData>dc:identifier

**EAD**
- archdesc.did.repository.corpname
- archdesc.did.repository.controlaccess.persname
- archdesc.did.repository.controlaccess.subject
- archdesc.controlaccess.geogname
- eadheader.eadid
- archdesc.did.unitid

**ONIX**
- Product>Contributor>PersonNameInverted
- Product>Contributor>PersonNameInverted>PersonDate>Date
- Product>Conference>ConferenceName
- Product>Conference>ConferenceDate
- Product>Conference>ConferencePlace
- Product>MainSubject>SubjectHeadingText
- Product>PlaceAsSubject
- Product>CorporateName

**XC**
- xc:entity>xc:creator
- xc:entity>xc:subject
- xc:entity>xc:type
- xc:entity>xc:id
Questions

Jeremy Myntti | jeremy.myntti@utah.edu
Head of Cataloging & Metadata Services

J. Willard Marriott Library
THE UNIVERSITY OF UTAH

Nate Cothran | nate@bslw.com
Vice President, Automation Services