# Authority Control *for* Digital Collections


Authority Control Interest Group
June 30, 2013

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

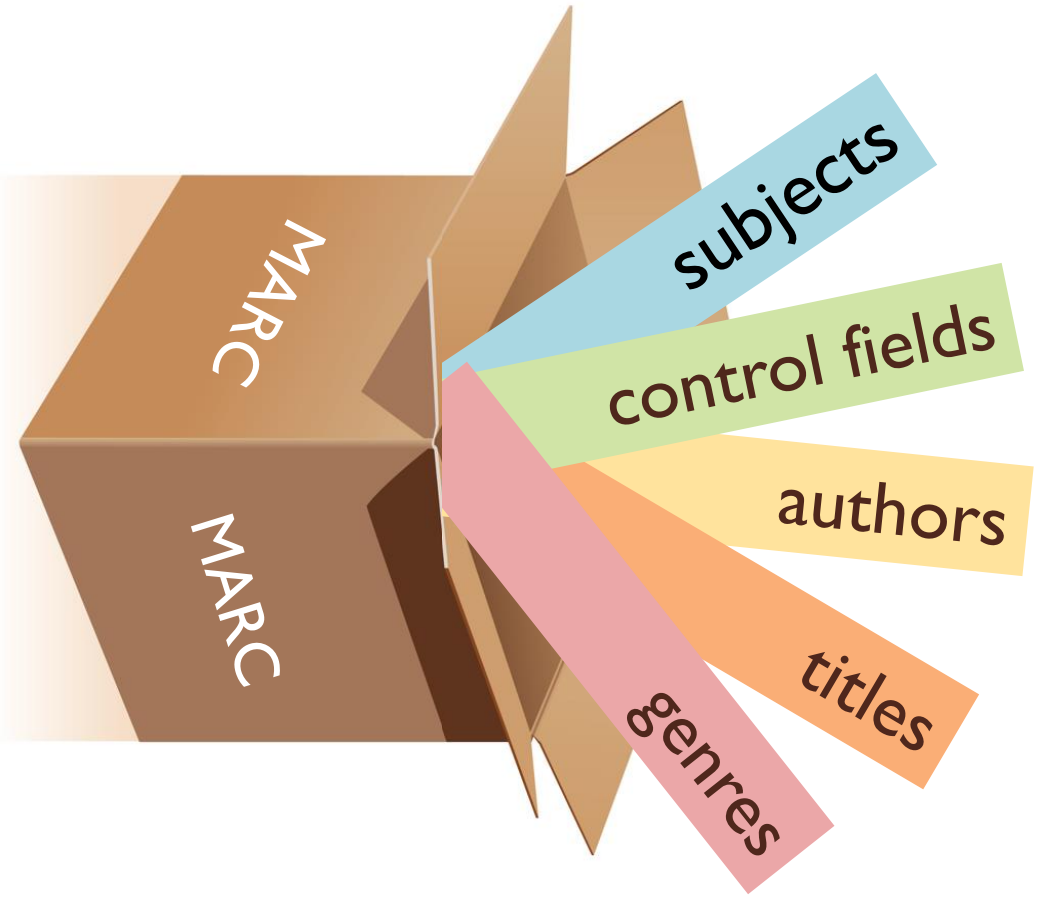

# Raw MARC

00491nam 22001571
4500001001400000003000800014005001700022
0080041000390400012000801000031000922450
0380012326000100016130000180017150501270
0189650001700316ovld00000000 1UtOrBLW201
21220145806.0121220s1991 xx eng
u aUtOrBLW1 aMercer, Leigh,d1893-1977.12aA
man, a plan, a canal -- Panama! c1991. a303 p.
;ccm. aDo geese see God? -- Murder for a jar of
red rum -- Some men interpret nine memos --
Never odd or even. 0aPalindromes.

# Raw MARC

00491nam 2200157l
4500001001400000003000800014005001700022
008004100039040001200080100003100092245 0
038001232600010000161300018000171505012 70
0189650001700316ovld00000000 1UtOrBLW201
212201458 06.0121220s1991 xx eng
u aUtOrBLW1 aMercer, Leigh,d1893-1977.12aA
man, a plan, a canal -- Panama! c1991. a303 p.
;ccm. aDo geese see God? -- Murder for a jar of
red rum -- Some men interpret nine memos --
Never odd or even. 0aPalindromes.

# MARC as a Container




Portable

Entrenched

Tools

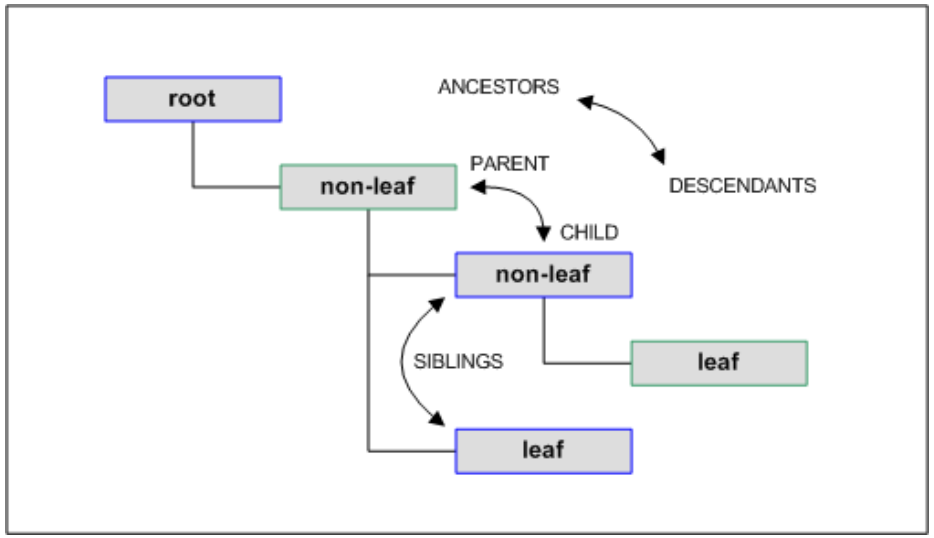
# XML Defined

- 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, Family




- Parents > Children
- Children : Siblings
- Children > Attributes
- Attributes : Values

# 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">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

<creata>Ogburn, Joyce L.</creata>

<title>Acquiring minds want to know: digital scholarship</title>

<date>2003</date>

<descri>A new form of scholarship has emerged in recent years named "digital scholarship.“</descri>

<type>text;</type>

<subjec>Born digital; Libraries; Electronic publishing</subjec>

<subjea>Libraries and electronic publishing; Scholarly electronic publishing </subjea>

# 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
  - <subject>Smith, John, 1947 Apr. 16-; United States--History--20th century; Navajo Indians--History; Rites and ceremonies; Religion;</subject>

- Variants
  - <subject-keyword>Ceremonies; Cult; Cultus; Ecclesiastical rites and ceremonies; Religious ceremonies; Religious rites; Rites of passage; Traditions;</subject-keyword>

# Pre-CONTENTdm Processing

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

- Temporary container elements are added

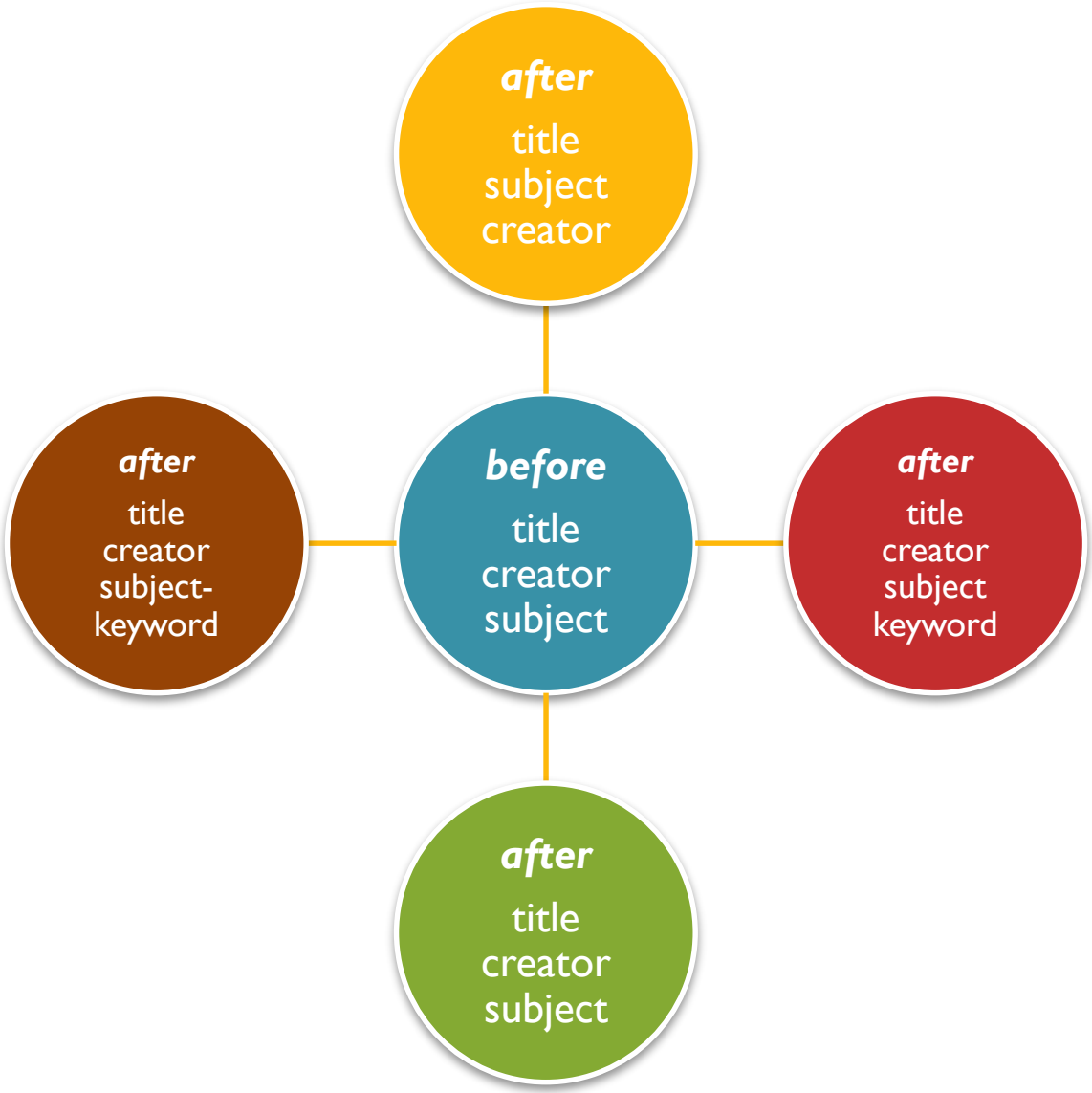
# Post-CONTENTdm Processing

<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)

<transc>...
immunofluorescence at sites of actin-membrane< f" " • - t . "U . * • ?vl />/> . * .. O' ' .< .*V#•* * •• * r , . *  . y l •* ...</transc>

- System tries to assign < … > as discrete child of transc

# Other Issues

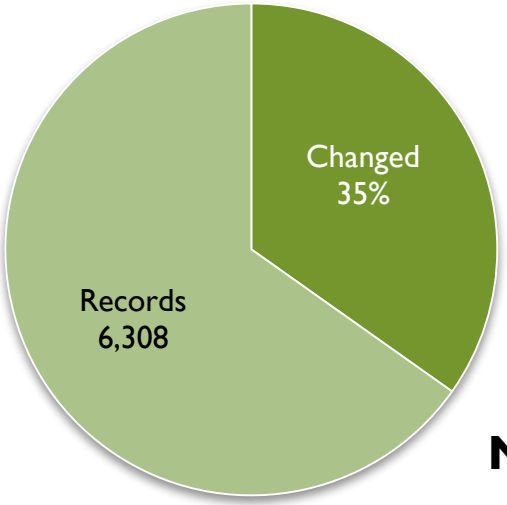
- Bad OCR data
- <descr> or <transc> field elements
- Errant angle brackets: < or >
- CDATA workaround (pre-processing)

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

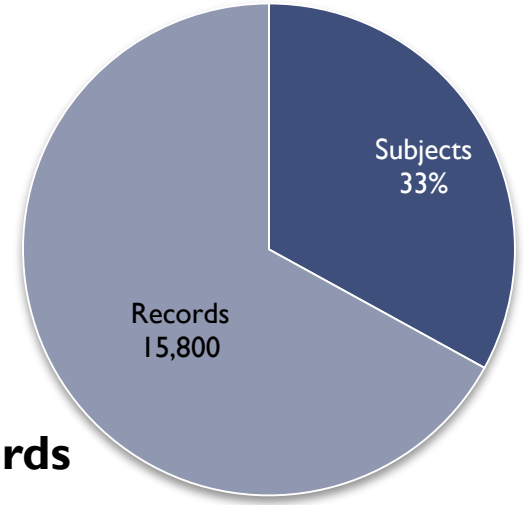
<![CDATA[<transc>... immunofluorescence at sites of actin-membrane< f" " • - t . "U . * • ?vl />/> . * .. O' ' .< .*V#•* * •• * r , . *  . y l •* ...</transc>]]>

# Statistics for USpace

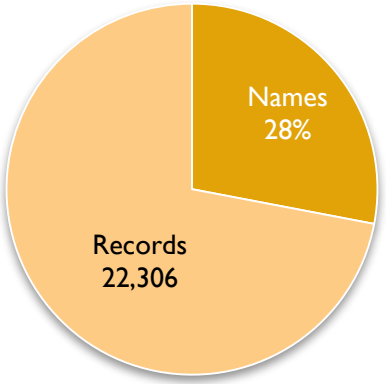
## Changed Records




## Matched Records




## Matched Records



# 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

- 1980; 1981; 1982; 1983;
- 2013-05-26;

# Reports

## Partially Matched Subjects

| Total | Records | Type | Partially Matched | Remainder of Heading |
|---|---|---|---|---|
| 1 | record | subjec | Gene targeting | --History |
| 1 | record | subjec | Religion | --Health aspects |

## Unmatched Names

| Total | Records | Type | Heading |
|---|---|---|---|
| 1 | record | other | Ahrens, Jim |
| 1 | record | other | Aide, T. Mitchell |
| 47 | records | creata | Ailion, David Charles |
| 1 | record | other | Ailion, Michael |
| 1 | record | other | Airola, Marc B. |
| 1 | record | other | Aizawa, Shin-Ichi |
| 1 | record | other | Aizprua, Rafael |
| 2 | records | other | Akella Venkatesh |

## Split Subjects

| Total | Records | Type | Heading |
|---|---|---|---|
| 3 | records | subjec | Beauty |
| 3 | records | subjec | Aesthetics |
| 3 | records | subjec | Art--Philosophy |
| 3 | records | subjec | Beauty, Personal |
| 2 | records | subjec | Biomedical research |
| 2 | records | subjec | Medicine--Research |
| 2 | records | subjec | Biology--Research |
| 2 | records | subjec | Birth rate |
| 2 | records | subjec | Childbirth--Statistics |
| 2 | records | subjec | Fertility, Human--Statistics |

## Updated Subjects

| Total | Records | Type | Heading |
|---|---|---|---|
| 1 | record | subjec | Abstract machines |
| 1 | record | subjec | Machine theory |
| 1 | record | subjec | ACE Inhibitors |
| 1 | record | subjec | Angiotensin converting enzyme--Inhibitors |
| 1 | record | subjec | Ache |
| 1 | record | subjec | Pain |

# Near Match Report

Levenshtein-Distance Algorithm

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

Near Matches

| Bib ID | *Type* | UNM | Unmatched Heading | %-1 | Tag-1 | Near Match-1 | Authority-1 |
|---|---|---|---|---|---|---|---|
| ir-main, 13197, ir-ma | *other* | 700 | Anderson, Richard C. E. | 90.0% | 100 | Anderson, Richard C. | n50021951 |
| ir-main, 1704, ir-mair | *creata* | 700 | Angelucci, Alessandra | 95.0% | 100 | Angelucci, Alessandro | n2012066779 |
| ir-main, 11558 | *other* | 700 | Annesley, Thomas M. | 88.2% | 100 | Annesley, Thomas | n2002152176 |
| ir-main, 566 | *other* | 700 | Arenkeil, Benjamin R. | 89.5% | 100 | Arenkiel, Benjamin R. | nr2004033027 |
| ir-main, 1637, ir-mair | *other* | 700 | Armstrong, Phillip A. | 89.5% | 100 | Armstrong, Phillip | n96107480 |
| ir-main, 16069 | *creata* | 700 | Ashton, Alan Conway | 75.0% | 100 | Ashton, Alan Conway, 1942- | no2007080328 |
| ir-main, 16084, ir-ma | *creata* | 700 | Aspinwall, Lisa | 87.5% | 100 | Aspinwall, Lisa G. | n2002101864 |
| ir-main, 1934 | *other* | 700 | Atomic force microscope | 95.7% | 150 | Atomic force microscopy | sh94008704 |
| ir-main, 4316 | *other* | 700 | Atwood, R. L. | 90.0% | 100 | Atwood, R. A. | n82034533 |
| ir-main, 9241 | *other* | 700 | Aubourt, E. | 88.9% | 100 | Aubourg, E. | no00039482 |
| uspace, 17683 | *other* | 700 | Austin, C. A. | 80.0% | 100 | Austin, C. | n79116889 |

# Future Plans: Ongoing AC via Linked Data

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

LIBRARY OF CONGRESS | ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS | Search | Search Loc.gov | GO

The Library of Congress > Linked Data Service > LC Name Authority File

# Ogburn, Joyce L.

From **Library of Congress Name Authority File**

Details | Visualization

## 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**
- found: North American Serials Interest Group. Conference (4th : 1989 : Scripps College). The serials partnership, 1989: CIP t.p. (Joyce L. Ogburn) galley (MSLS, MA; acquisitions librarian, Penn St. Univ., University Park, Pa.)

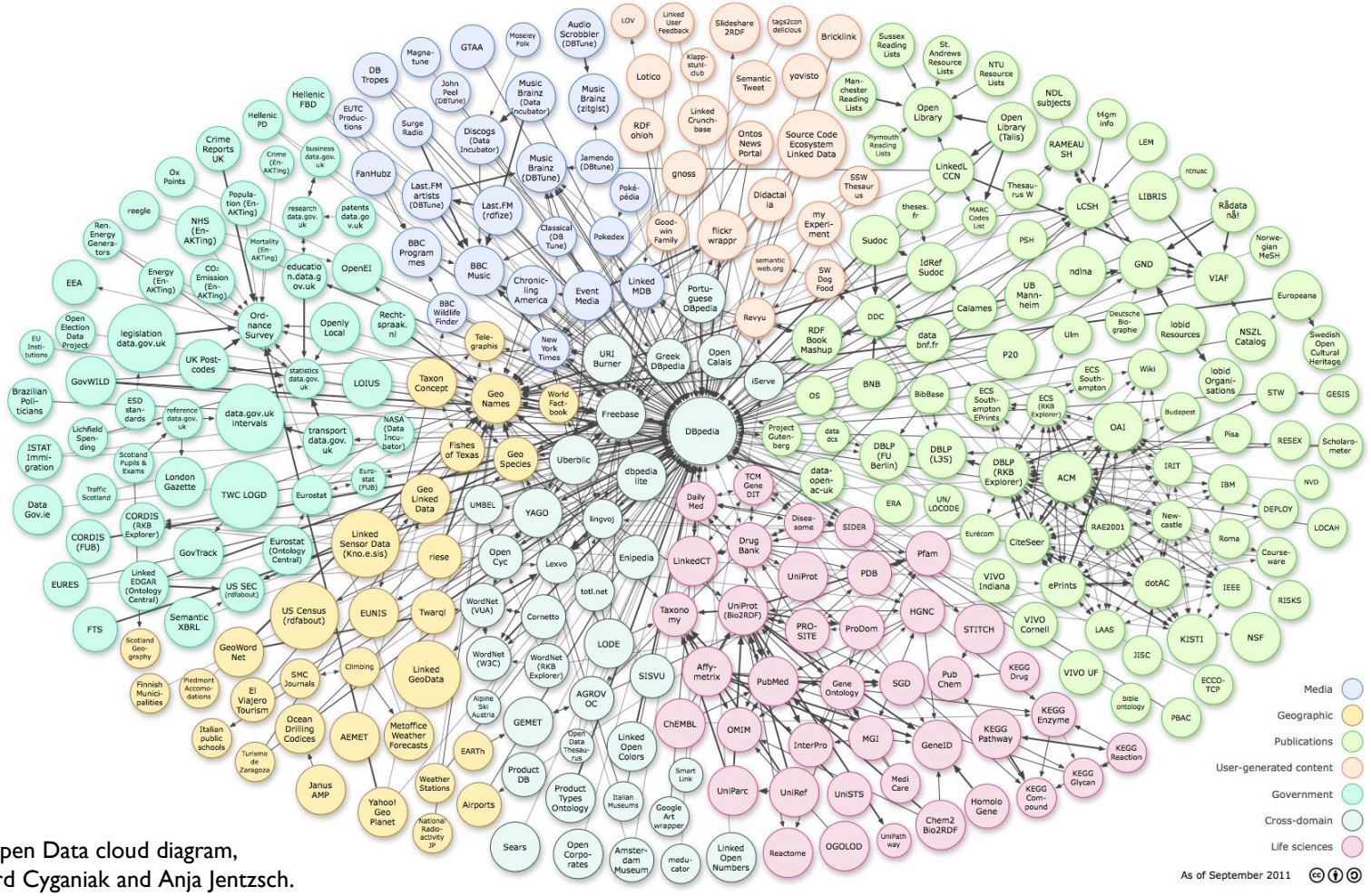
**Change Notes**
- 1989-09-27: new
- 1989-11-28: revised

**Alternate Formats**
- RDF/XML (MADS and SKOS)
- N-Triples (MADS and SKOS)
- JSON (MADS/RDF and SKOS/RDF)

# 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




Linking Open Data cloud diagram,
by Richard Cyganiak and Anja Jentzsch.
http://lod-cloud.net/

# 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

Backstage
LIBRARY WORKS