Information Organization and Metadata Research

Dr. Oksana Zavalina,
Information Science Research Brown Bag presentation
February 1, 2019
Outline

• Introduction
• Research team and publications
• Most recent presentation: ASIS&T 2018 paper
• 2018-2019 IMLS-funded planning project
Information Organization (IO) is Essential for Providing Access to Information (and Data)

Without IO

- humans would not be able to find, identify, select and obtain information and data they need in everyday life, professional and scholarly activities, etc.
- machines would not be able to assist humans in information/data seeking or to make inferences and connect pieces of information and data in a meaningful whole (e.g., Semantic Web)
Metadata as key to IO: main components

1. **Metadata records** adequately representing important to the users attributes of information objects (e.g., journal articles resulting from a research project) and data (e.g., datasets used in that research project)
   - e.g., audience; creator, contributor, publisher etc.; date; format; topical, geographical and temporal subjects; title; type; and many more.

2. **Data content standards** that guide creation of metadata records for various user communities

3. **Data value standards** that provide guidelines and controlled vocabularies for consistent representation of information in metadata elements and enable collocation and disambiguation of results.

4. **Data encoding / transmission standards** that enable sharing, exchanging and reusing metadata
### Metadata-related research is published in:

**journal articles, e.g.:**

- Journal of Library Metadata
- Journal of the Association for Information Science and Technology (JASIS&T)
- Cataloging and Classification Quarterly
- The Electronic Library journal
- International Journal of Metadata, Semantics, and Ontologies
- etc.

- [http://www.tandfonline.com/toc/wjlm20/current](http://www.tandfonline.com/toc/wjlm20/current)
- [http://catalogingandclassificationquarterly.com/](http://catalogingandclassificationquarterly.com/)
- [http://www.emeraldinsight.com/loi/el](http://www.emeraldinsight.com/loi/el)
**Metadata-related research is published in**

<table>
<thead>
<tr>
<th>conference proceedings, e.g.:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Dublin Core Metadata Initiative (DCMI) conference</td>
<td><a href="http://dcevents.dublincore.org/intConf">http://dcevents.dublincore.org/intConf</a></td>
</tr>
<tr>
<td>• ASIS&amp;T annual meeting</td>
<td><a href="https://www.asist.org/events/annual-meeting/">https://www.asist.org/events/annual-meeting/</a></td>
</tr>
<tr>
<td>• iSchools conference (iConference)</td>
<td></td>
</tr>
<tr>
<td>• Joint Conference on Digital Libraries (JCDL)</td>
<td><a href="http://www.jcdl.org/">http://www.jcdl.org/</a></td>
</tr>
<tr>
<td>• Also European equivalent (ECDL or TPDL), Asian equivalent (ICADL)</td>
<td><a href="http://www.tpdl.eu/">http://www.tpdl.eu/</a></td>
</tr>
<tr>
<td>• ICKM conference</td>
<td><a href="https://link.springer.com/conference/icadm">https://link.springer.com/conference/icadm</a></td>
</tr>
<tr>
<td>• etc.</td>
<td></td>
</tr>
</tbody>
</table>
Quality of Metadata is Important

Metadata fitness for the purpose of enabling data/information discovery and reuse through supporting user tasks:

- find
- identify
- select
- obtain
Metadata quality criteria suggested by the literature

• Access
• **Accuracy**
• Availability
• Compatibility
• **Completeness**
• Comprehensiveness
• Content
• **Consistency**
• Cost
• Data Structure
• Ease Of Creation

Most important from the point of view of metadata creators (e.g., Park & Tosaka, 2009)

• Ease Of Use
• Economy
• Flexibility
• Fitness For Use
• Informativeness
• Protocols
• Quantity
• Reliability
• Standard
• Timeliness
• Transfer
• Usability
Metadata change as part of metadata quality assurance

Change in metadata records is encouraged by agencies that facilitate cooperative metadata creation, management and sharing.

To keep up with “environmental” changes, e.g.:

- Growth in certain types/formats and subject matter of materials in repositories
- Changes in the content & location of fluid materials (e.g., websites)
- Goals of hosting & contributing institutions
- KOS: classification systems & controlled vocabularies
- National & international standards for metadata creation.
Change research in computer science does not look into metadata

- Mechanisms for identifying change (e.g., edit distance, Bille, 2005) and file comparison tools for isolating differences between:
  - files
  - texts, strings
  - programs, scripts, applications, ontologies
  - multiple versions of the same entities.
  (e.g., Cheney, 2010; Horwitz, 1990; Noy et al., 2004)
Change research in information science

**Metadata quality** research:
  * suggested the link between metadata change and metadata quality
  * emphasized the need to measure the metadata change and its outcomes for the users

(Stvilia et al., 2004; Stvilia & Gasser, 2008)

**BUT**

Almost no published research identifying and measuring metadata change until recently
  * UNT team is pioneering metadata change research.
UNT researchers contributing to metadata change research since 2014

Dr. Daniel Alemneh
Priya Kizhakkethil
Dr. Shawne Miksa
Mark Phillips
Dr. Shadi Shakeri
Hannah Tarver
Slava Zavalin
Dr. Oksana Zavalina
UNT researchers’ published contributions to metadata change research


Evaluation of Metadata Change in Authority Data over Time: an Effect of a Standard Evolution

Oksana L. Zavalina and Vyacheslav Zavalin
Department of Information Science, University of North Texas
Authority data = controlled vocabularies in the library community

Standardized machine-readable records that describe:

- persons
- institutions
- places
- events
- works, etc.
- relations between them

Extensive and rapidly growing databases, e.g.:

- US Library of Congress
  - Name Authority File (NAF):
    - over 8 M records
    - 22% growth between 2011 & 2014
- Subject Authority File (SAF)
  - Concepts
  - Objects
- Virtual International Authority File (VIAF), etc.

Mostly in MARC21 format
Example of an authority record (corporate name, from LC NAF)

This authority record is RDA-based.
Resource Description and Access (RDA)

• A data content standard for metadata, including authority data in library community

• Developed since 2008, officially implemented in March 2013
  • replaced previous standard AACR

• Intended to greatly improve functionality of authority data:
  • focus on representing important attributes and relations
  • for Linked Data / Semantic Web development

• Introduces a number of new data elements in authority records, e.g.:
  • 35 new MARC fields for name and/or title authority records overall
  • 7 new MARC fields for corporate name name authority data
  • 5 new Linked Data enabling MARC subfields

  e.g., MARC field 377 Associated Language
  e.g., 370 Associated Place
  e.g., #u Universal Resource Identifier (URI)
Research team at UNT has been investigating metadata change in metadata that describes information objects (e.g., bibliographic records) since 2014

- In digital and traditional libraries
- RDA and non-RDA
- MARC21 and beyond

Several quantitative studies attempted to identify and measure change in metadata records in digital libraries that enable metadata versioning (e.g., Tarver, Zavalina & Phillips, 2016; Zavalina, Phillips & Tarver, 2017).

A qualitative research project (e.g., Zavalina et al., 2015, 2016; Zavalina, Shakeri, & Kizhakkethil, 2015; Zavalina, Shakeri, Kizhakkethil, & Phillips, 2018) categorized metadata change in digital library metadata and in traditional library metadata.
Related Work (2)

- Few published studies analyzed authority data in relation to RDA guidelines
  - 2 focused on personal name authority records (*Moulaison, 2015; Thompson, 2016*)
    - either small sample of records or a subset of data elements
  - 1 study (*Kimura, 2015*) focused on authority data created in China, Japan, & Korea
    - over 1M of name authority records of 3 kinds: personal, corporate, and meeting
    - BUT mostly non-RDA-based authority data
  - 1 recent study (Zavalina & Zavalin, 2017) evaluated application of RDA-specific elements in a large sample of RDA-based authority records of 5 kinds: personal, corporate, meeting, geographic and title
Problem Statement / Research focus

• Shortage of research evaluating results of implementation of RDA in authority records
  • across various kinds of authority records
  • for the whole spectrum of data elements, and
  • with large samples or Big Data approach

• NO research evaluating how authority records change over time
Research Questions

• What is the level of application of the new RDA-based data elements of MARC authority records
• How does this level change over time?

• How are the Linked Data enabling elements of RDA applied in the existing authority data and
• How this changes over time?
Methods (1)

• Intention to apply Big Data analytics approach and collect all RDA-based authority records in NAF as of March 2016: **1.2 M**

• Raw (ADV) search in MARC Edit Z39.50/SRU Client to harvest records

• Software limitations resulted in collecting **408.5 K** records in 2016
  • Large random sample (**34%**)
  • Representative sample: all 5 types of NAF authority records harvested

• 2 data collection points approx. 22 months apart:
  • early March 2016 and late December 2017
Methods (2)

• Same dataset of 408.5 K authority records (based on unique record IDs) harvested in 2017:
  • All but 26 records (probably deleted from NAF between collection points)
• Identified 35.47 K records that underwent changes between 2016 and 2017 data collection points
  • Based on data in MARC field 005 Date & Time of Latest Transaction
• Quantitative content analysis of the 35.47 K changed authority records
Findings:

Types of changed records (n=35,472)

- meeting name, 60.16%
- personal name, 39.60%
- corporate name, 0.17%
- geographic name, 0.02%
- uniform title, 0.04%
### Example of RDA-based meeting name authority record in MARC

<table>
<thead>
<tr>
<th>ARN</th>
<th>6904209</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec_stat</td>
<td>c</td>
</tr>
<tr>
<td>Entered</td>
<td>20060404</td>
</tr>
<tr>
<td>Replaced</td>
<td>20170304073758.0</td>
</tr>
<tr>
<td>Type</td>
<td>z</td>
</tr>
<tr>
<td>Upd status</td>
<td>a</td>
</tr>
<tr>
<td>Enc lv</td>
<td>n</td>
</tr>
<tr>
<td>Source</td>
<td>Source</td>
</tr>
<tr>
<td>Roman</td>
<td>Ref status</td>
</tr>
<tr>
<td>Auth status</td>
<td>a</td>
</tr>
<tr>
<td>Govt agn</td>
<td>Subj use</td>
</tr>
<tr>
<td>Series</td>
<td>Auth/ref</td>
</tr>
<tr>
<td>Ser num</td>
<td>Name n</td>
</tr>
<tr>
<td>Sub div tp</td>
<td>Sub div tp n</td>
</tr>
<tr>
<td>Rules</td>
<td>Rules</td>
</tr>
</tbody>
</table>

**RDA-specific MARC fields**

<table>
<thead>
<tr>
<th>Field</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>010</td>
<td>nb2006008174</td>
</tr>
<tr>
<td>040</td>
<td>Uk ‡b eng ‡e rda ‡c Uk ‡d DLC ‡d IEN ‡d WaU</td>
</tr>
<tr>
<td>046</td>
<td>t 20120727 ‡t 20120812</td>
</tr>
<tr>
<td>111</td>
<td>Olympic Games ‡n (30th ‡d 2012 : ‡c London, England)</td>
</tr>
<tr>
<td>368</td>
<td>Multi-sport event ‡a Sporting event</td>
</tr>
<tr>
<td>368</td>
<td>Sports tournaments ‡a lcsh</td>
</tr>
<tr>
<td>370</td>
<td>‡c Great Britain ‡c London (England) ‡2 naf</td>
</tr>
<tr>
<td>377</td>
<td>eng</td>
</tr>
</tbody>
</table>

**Olympic games impact study, 2005.**

**Olympic organisation website, 4 Apr. 2006 ‡b London page (London 2012 Games of the XXX Olympiad to be held 27 July to 12 Aug. 2012).**

**Programme procurement in construction, learning from London 2012, 2013: ‡b galley (The Games of the 30th Olympiad held in London during the summer of 2012).**

**Wikipedia, Feb. 17, 2013 ‡b (2012 Summer Olympics: The 2012 Summer Olympics, officially the Games of the XXX Olympiad and commonly known as London 2012, was a major international multi-sport event celebrated in the tradition of the Olympic Games, as governed by the International Olympic Committee (IOC). It took place in London, United Kingdom, from 27 July to 12 August 2012. The first event, the group stage in women's football, began two days earlier, on 25 July.**


Findings:

Changes in occurrences of RDA-specific MARC21 fields: observed 17 out of 35 fields

- % of records in 2016
- % of records in 2017
Findings:

Change in occurrences of subfield-level data elements in RDA-specific MARC fields: top 20 of 60 observed

For field of activity
For associated group
For gender
Conclusions

Begin addressing research gap in the area of implementation of RDA standard in authority data that are crucial for providing adequate access to information

• Lower overall level of editing activity than that observed by previous research for RDA-based bibliographic metadata (Zavalina, Zavalin, & Miksa, 2016)
• Higher editing activity for meeting name and personal name authority data than for three other types of authority records
• Change in application of certain data elements, related to evolution of RDA standard
• Gradual and sometimes drastic increase in the use of elements representing persons, as well as some of the Linked Data-enabling elements
• Despite the observed growth, the level of application of Linked-Data-enabling elements in authority records remains relatively low
Future Research

• Supplement quantitative analysis of a large dataset by in-depth analysis on its samples
  • Focus on data values in fields/subfields
  • Categorization of change beyond addition or deletion of a field/subfield instance

• Comparative analysis of metadata change needed
  • for different kinds of authority data in NAF
  • between records in NAF subject authority records in SAF
  • between authority data and bibliographic data.

• Longitudinal analysis of change in authority records over time
  • Especially application of data elements that provide Linked Data functionality (e.g., #0 Authority record control number or standard number, #2 Source of term, #4 Relationship, #i Relationship information, #u Uniform Resource Identifier, etc.).
IMLS-funded project to support information organization for Linguistics community (2018-2019)

Dr. Shobhana Chelliah
Mark Phillips
Mary Burke
Dr. Oksana Zavalina
LG-87-18-0197
Exploring Methods and Techniques for Facilitating Access to Digital Language Archives

Planning project to identify the gaps between the information organization methods and techniques currently offered in existing language data archives and the needs of actual and potential language data archive users.

Expected to provide necessary background information and preparation for a forthcoming collaborative research project that will aim to extend the usefulness of existing language data archive collections through a user-centered design of systems incorporating the efficient methods and techniques for providing digital access to language data collections at scale.
LG-87-18-0197
Project stages

<table>
<thead>
<tr>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Identifying language archives for analysis

Phase 1: Explorative content analysis: data collection

Phase 1: Explorative content analysis: data analysis and presentation of preliminary results to advisory board for feedback

Identifying participants for Phase 2

Phase 2: Interview data collection

Phase 2: Interview data analysis

Phase 2: Observation data collection

Phase 2: Observation data analysis and presentation of preliminary results to advisory board for feedback

Preparation of project reports, dissemination of results
INTERESTED IN CONTRIBUTING TO METADATA RESEARCH BUT LACK NECESSARY METADATA BACKGROUND?

Complete UNT graduate courses:
INFO 5223 Metadata 1
INFO 5210 RDA 1
INFO 5740 Digital Libraries

Also, more advanced courses are offered:
INFO 5224 Metadata 2
INFO 5220 RDA 2
INFO 5212 DDC
etc.