UNT®

UNIVERSITY OF NORTH TEXAS®

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://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2330-1643 •http://catalogingandclassificationquarterly.com/ •http://www.emeraldinsight.com/loi/el •http://www.inderscience.com/jhome.php?jcode=ijmso

Metadata-related research is published in:

Hosted in 2018 by UNT, June 4-6, in DFW

conference proceedings, e.g.:	
•Dublin Core Metadata Initiative (DCMI) conference •ASIS&T annual meeting •iSchools conference (iConference) •**Joint Conference on Digital Libraries (JCDL)** •Also European equivalent (ECDL or TPDL), Asian equivalent (ICADL) •**ICKM conference** •etc.	http://dcevents.dublincore.org/IntConf https://www.asist.org/events/annual-meeting/ http://www.jcdl.org/ http://www.tpdl.eu/ , https://link.springer.com/conference/icadl **http://kipanet.org/content/13th-international-conference-knowledge-management-ickm-2017**

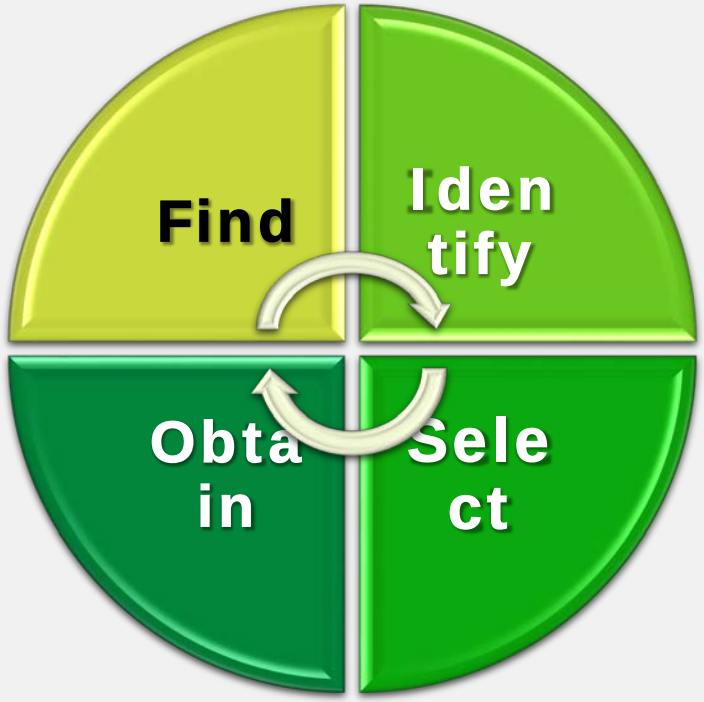
Hosted in 2017 by UNT, October 25-27 in DFW

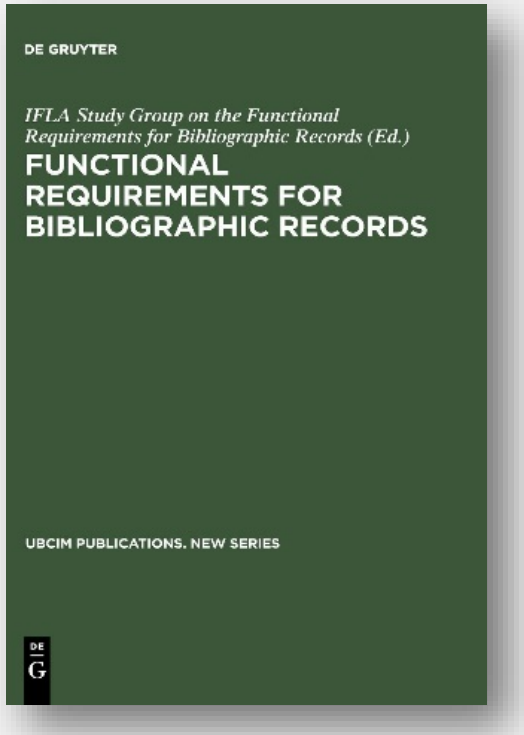
Quality of Metadata is Important

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
- Compactness
- 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

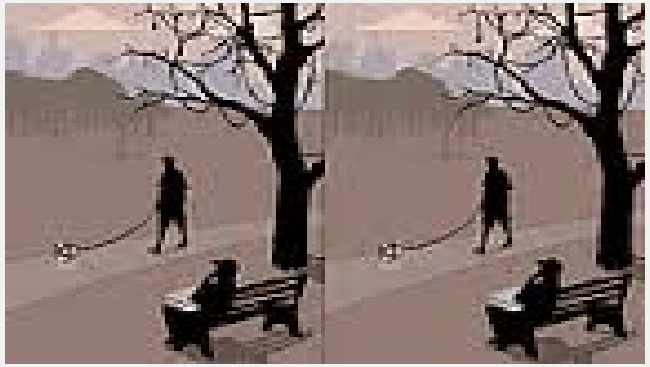
Change in metadata records is encouraged by agencies that facilitate cooperative metadata creation, management and sharing

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

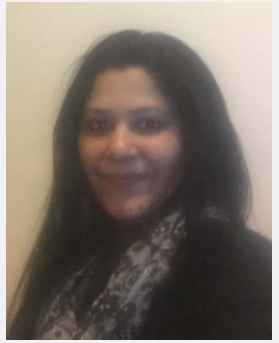
Mark Phillips

Dr. Shadi Shakeri

Hannah Tarver

Slava Zavalin

Dr. Oksana Zavalina





UNT researchers' published contributions to metadata change research

- Tarver, H. Zavalina, O.L., Phillips, M., Alemneh, D., & Shakeri, S. (2014). How descriptive metadata changes in the UNT Libraries' Collections: a case study. *Proceedings of the International Conference and Workshop on Dublin Core and Metadata Applications*, Austin, Texas.
- Zavalina, O.L., Kizhakkethil, P., Alemneh, D., Phillips, M., & Tarver, H.S. (2014). Metadata changes: meeting the evolving requirements. *Proceedings of the 10th International Conference on Knowledge Management*. Antalya, Turkey, November 24-26, 2014.
- Zavalina, O.L., & Kizhakkethil, P. (2015). Exploration of metadata change in a digital repository. *Proceedings of the iConference 2015*. Newport Beach, California, March 24-27, 2015.
- Zavalina, O.L., Kizhakkethil, P., Alemneh, D., Phillips, M., & Tarver, H.S. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information and Knowledge Management, 14* (1), 1-16.
- Zavalina, O.L., Shakeri, S., & Kizhakkethil, P. (2015). Metadata change in traditional library collections and digital repositories: Exploratory comparative analysis. *Proceedings of the 78th Association for Information Science and Technology Annual Meeting*. Saint Louis, Missouri, November 6-10, 2015.
- Zavalina, O.L., Shakeri, S., & Kizhakkethil, P. (2016). An empirical investigation of change in subject metadata in WorldCat. *Proceedings of the International Federation of Library Associations World Library and Information Congress Satellite Conference "Subject Access: Unlimited Opportunities"*, Columbus, Ohio, August 11-12, 2016.
- Zavalina, O.L., Zavalin, V., & Miksa, S. D. (2016). Quality over time: A longitudinal quantitative analysis of metadata change in RDA-based MARC Bibliographic Records Representing Video Resources. *Proceedings of the 79th Association for Information Science and Technology Annual Meeting*. Copenhagen, Denmark, October 14-18, 2016. International.
- Zavalina, O.L., Zavalin, V., Shakeri, S., & Kizhakkethil, P. (2016). Developing an empirically-based framework of metadata change and exploring relation between metadata change and metadata quality in MARC library metadata. *Procedia Computer Science, 99*, 50-63.
- Zavalina, O.L., Phillips, M., & Tarver, H. (2017). Quality assurance and evaluation of change for patent metadata. *Proceedings of the 80th Association for Information Science and Technology Annual Meeting*.
- Zavalina, O.L., Shakeri, S., Kizhakkethil, P., & Phillips, M.E. (2018). Uncovering hidden insights for information management: Examination and modelling of change in digital collection metadata. In (Eds.), *Lecture Notes in Computer Science*. New York: Springer.
- Zavalina, O., L., & Zavalin V. (2018). Evaluation of metadata change in authority data over time: An effect of a standard evolution. *Proceedings of the Association for Information Science and Technology.*

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

UNT EST. 1890

asis&t 81st ANNUAL MEETING NOVEMBER 10-14, 2018 VANCOUVER, CANADA

Authority data = controlled vocabularies *in the library community*

Mostly in MARC21 format

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.

Example of an authority record (corporate name, from LC NAF)

The Library of Congress > Linked Data Service

From Library of Congress Name Author

Details | Visualization

Association for Information

URI(s)
- http://id.loc.gov/authorities/names/no20140250

Instance Of
- MADS/RDF CorporateName
- 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

Variants
- ASIS & T (Organization)

Additional Information
- http://id.loc.gov/rwo/agents/no2014025096
- **Associated Locale**
  United States
- **Associated Locale**
  Silver Spring (Md.)

Associated Language
eng

Field of Activity
- Information science
- Information technology
- Information technology

asis&t
Association for Information Science and Technology

Related Terms
- American Society for Information Science and Technology

Exact Matching Concepts from Other Schemes
- http://viaf.org/viaf/sourceID/LC%7Cno2014025096#skos:Concept

Sources
- found: OCLC 853619590: Journal of the Association for Information Science and Technology (a American Society for Information Science and Technology)
- found: ASIS&T website, viewed December 10, 2013 : (Association for Information Science an Information Science and Technology; their headquaters are in Silver Spring, Maryland; the or of information technology)

Change Notes
- 2014-02-25: new
- 2014-02-27: revised

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

This authority record is RDA-based

RDA
RESOURCE DESCRIPTION & ACCESS

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)***

Related Work (1)

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)

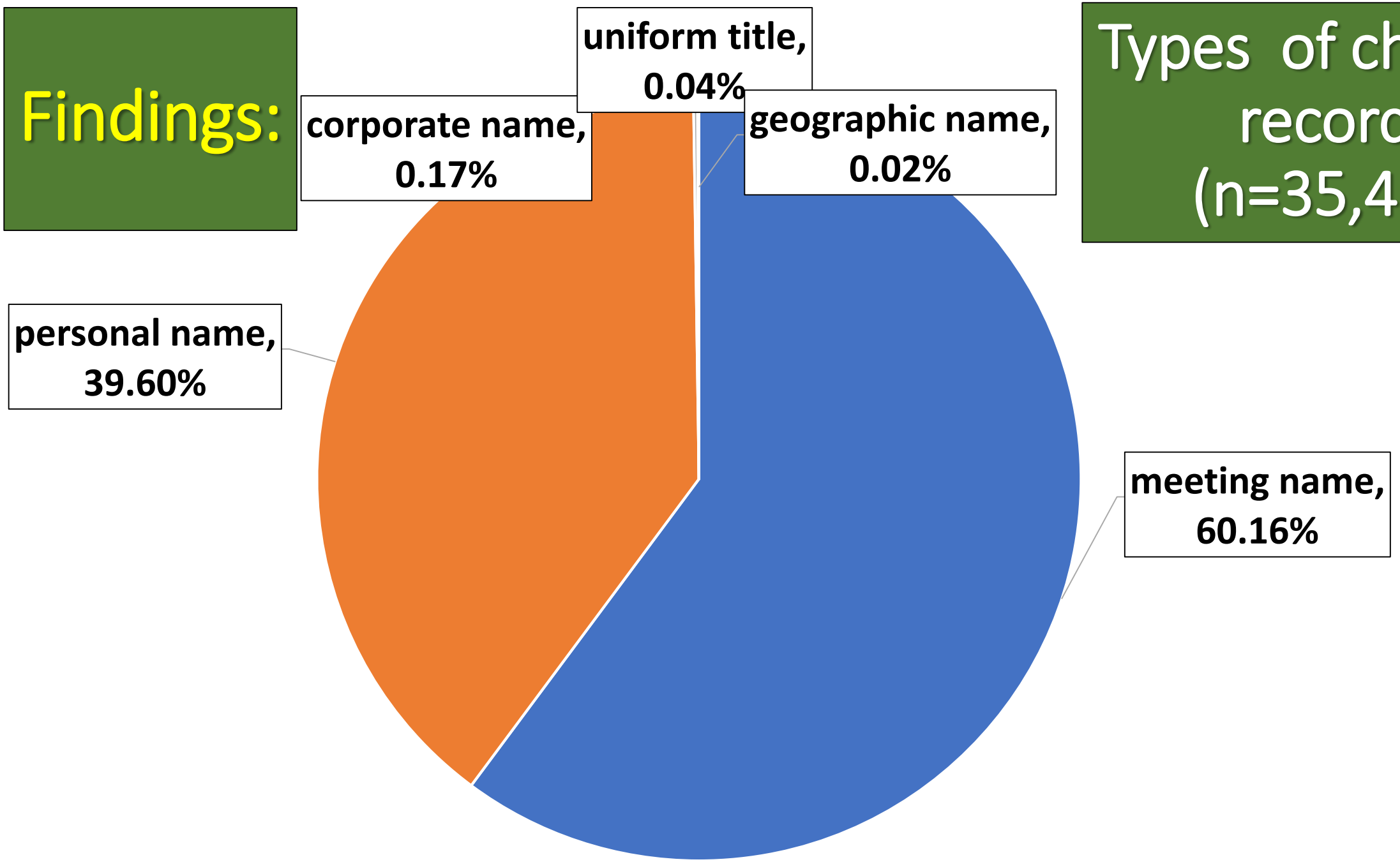
- Intention to apply Big Data analytics approach and collect all RDA-based authority records in NAF as of March 2016: **1.2 M**
- Raw (ADV) search in MARC Edit Z39.50/SRU Client to harvest records
- Software limitations resulted in collecting **408.5 K** records in 2016
  - Large random sample (**34%**)
  - 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)



Example of RDA-based meeting name authority record in MARC

ARN									6904209
Rec stat	c	Entered	20060404	Replaced	20170304073758.0				
Type	z	Upd status	a	Enc lvl	n	Source			
Roman	|	Ref status	a	Mod rec		Name use	a		
Govt agn	|	Auth status	a	Subj	a	Subj use	a		
Series	n	Auth/ref	a	Geo subd	n	Ser use	b		
Ser num	n	Name	n	Subdiv tp	n	Rules	z		

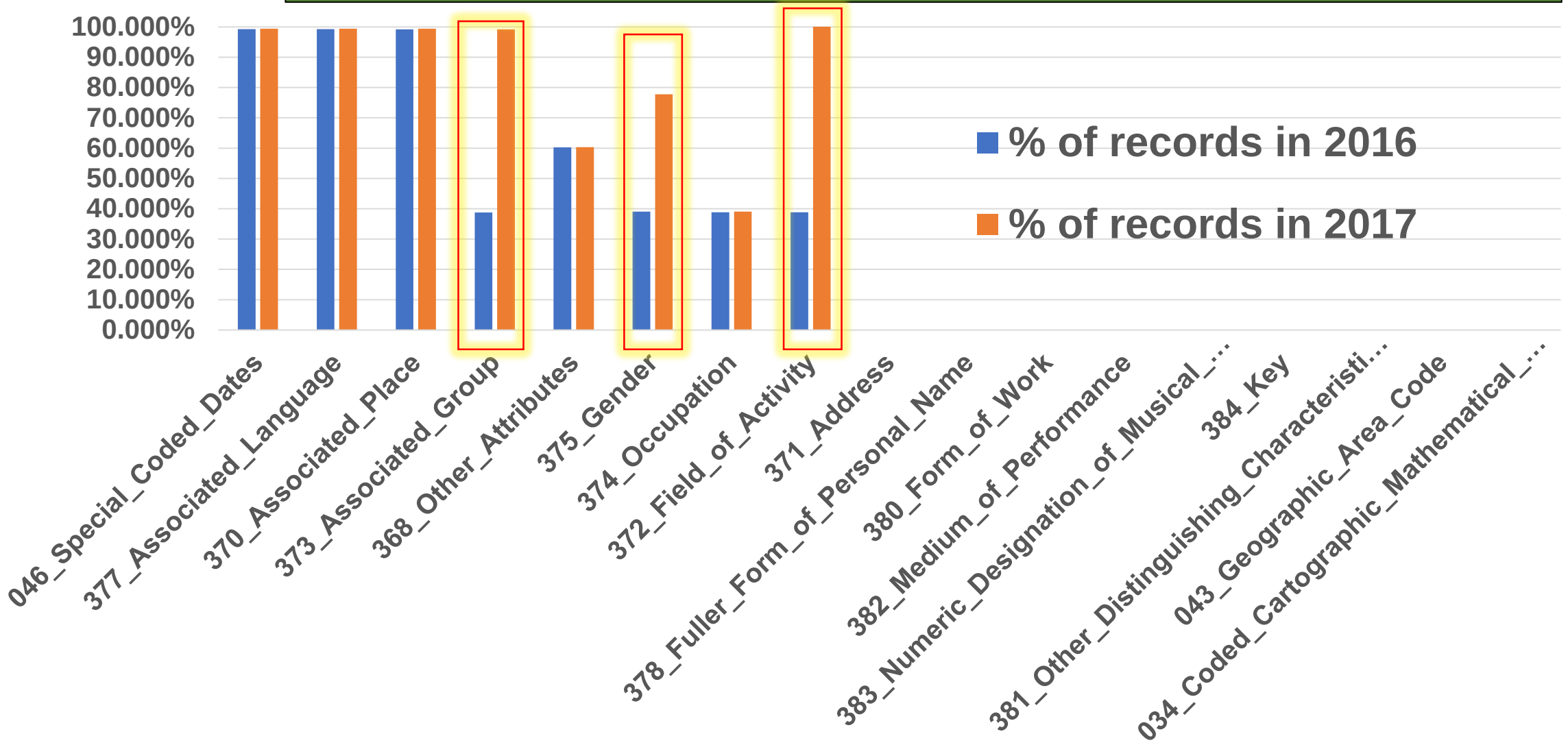
010		nb2006008174
040		Uk ‡b eng ‡e rda ‡c Uk ‡d DLC ‡d IEN ‡d WaU
046		‡s 20120727 ‡t 20120812
111	2_	Olympic Games ‡n (30th : ‡d 2012 : ‡c London, England)
368		Multi-sport event ‡a Sporting event
368		Sports tournaments ‡2 lcsh
370		‡c Great Britain ‡e London (England) ‡2 naf
377		eng
411	2_	Games of the XXX Olympiad ‡d (2012 : ‡c London, England)
411	2_	Games of the 30th Olympiad ‡d (2012 : ‡c London, England)
411	2_	Games of the Olympiad ‡n (30th : ‡d 2012 : ‡c London, England)
411	2_	Olympiad, Games of the ‡n (30th : ‡d 2012 : ‡c London, England)
411	2_	London 2012 (Olympics) ‡d (2012 : ‡c London, England)
411	2_	Summer Olympics ‡n (30th : ‡d 2012 : ‡c London, England)
411	2_	London Olympic Games ‡d (2012 : ‡c London, England)
411	2_	London Summer Olympics ‡d (2012 : ‡c London, England)

RDA-specific MARC fields

670	Olympic games impact study, 2005.
670	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)
670	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)
670	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)
670	Olympic Games 27 July-12 August official London 2012 website, viewed Feb. 17, 2013: ‡b home page (Olympic Games; London 2012 Olympic and Paralympic Games) ‡u http://www.london2012.com/
670	Olympic.org website, Feb. 17, 2013: ‡b Olympic Games > London 2012 (London 2012; Olympic Games; Opening date: 27 July 2012; Closing date: 12 August 2012; Country of the host city: Great Britain; HTML headoer: London 2012 Summer Olympics) ‡u http://www.olympic.org/london-2012-summer-olympics

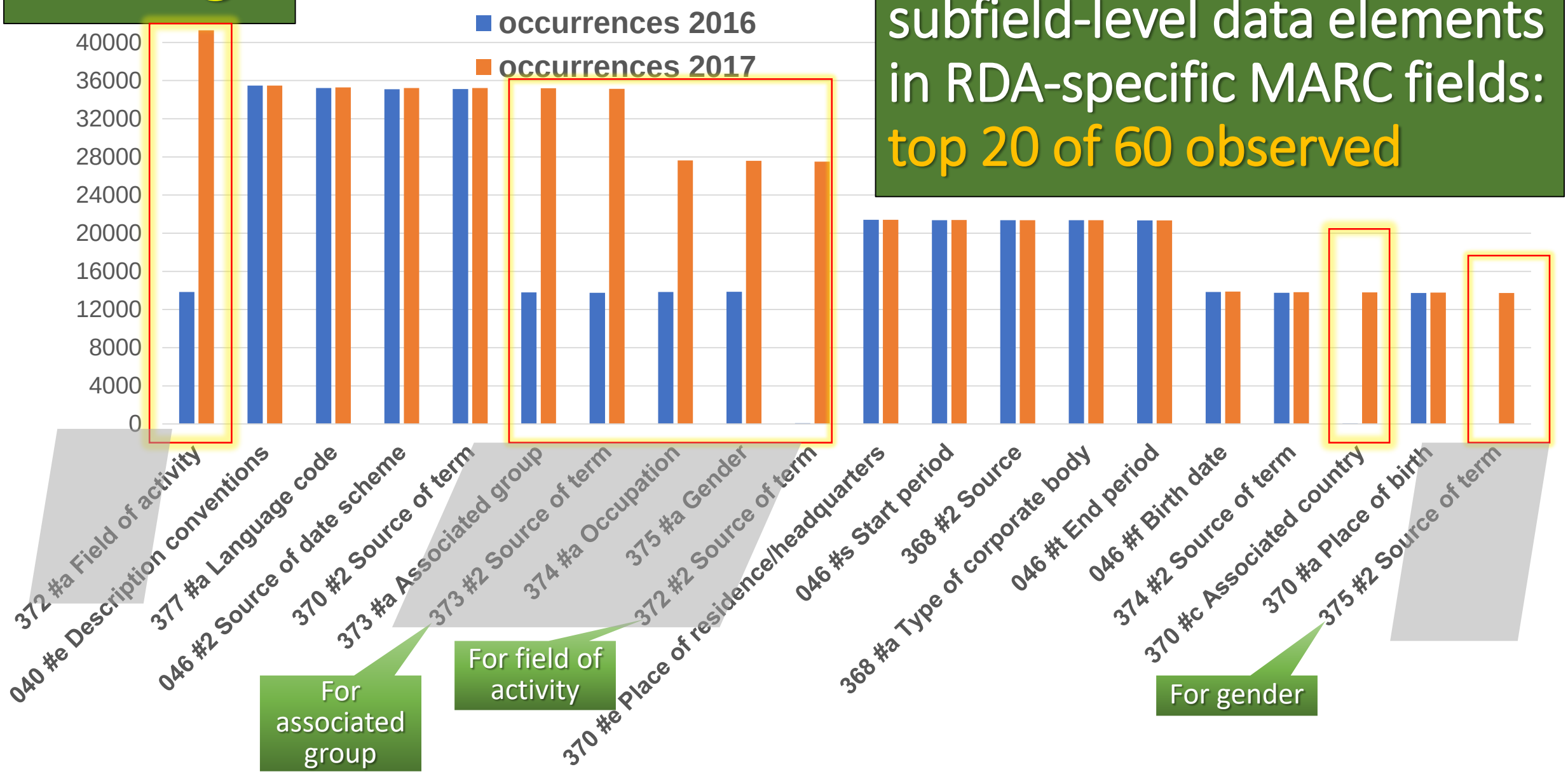
Findings:

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



Findings:

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



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)

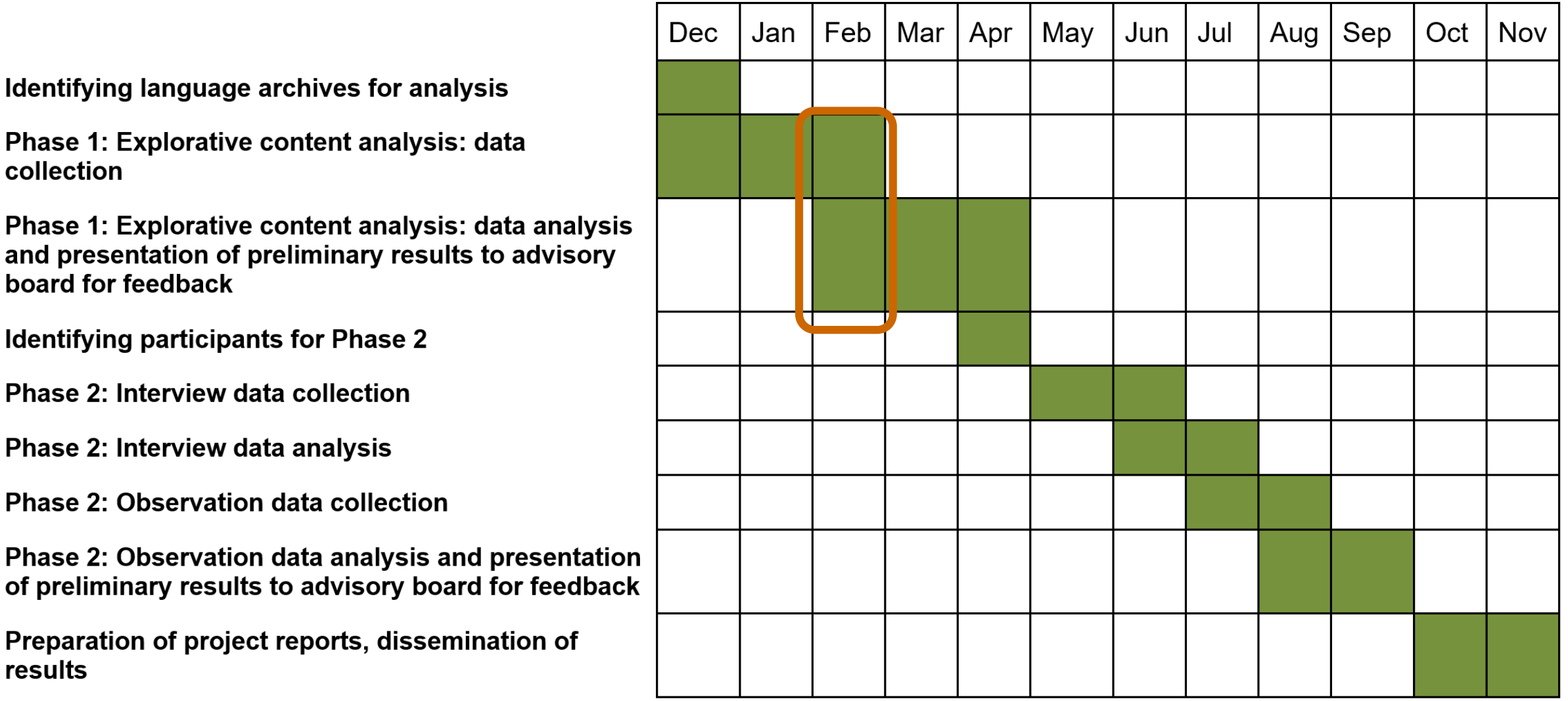
Dr. Shobhana Chelliah
Mark Phillips
Mary Burke
Dr. Oksana Zavalina

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



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.