Metadata Quality Research: *Introduction*

Dr. Oksana L. Zavalina

(Oksana.Zavalina@unt.edu)

Research Brown Bag presentation

October 11, 2024

Presentation Outline

- Introduction to metadata and its quality
- "How-to" of metadata quality assessments
- Some published metadata quality research of UNT-affiliated researchers: dissertations, journal articles, conference papers
- AI as a factor in the growing importance of metadata quality research
- How to develop knowledge & skills needed for doing metadata quality research right (e.g., meaningfully interpret the findings)?

-- Public - No restrictions on sharing --

Metadata is Essential for Providing Access to Information (and Data)

Without adequate metadata, machines would be unable to

- assist humans in information/data seeking
- make inferences and connect pieces of information and data in a meaningful whole (e.g., Semantic Web)

Without adequate metadata, humans would be unable to effectively discover and reuse information and data

- ***find***, ***identify***, ***select*** and ***obtain*** information & data they need in professional and scholarly activities, everyday life, health care, etc.;
- ***explore*** relations between information/data resources, etc.

-- Public - No restrictions on sharing --

Metadata Types & Subtypes

Include but are not limited to:

- **Bibliographic** *(represents information/data resources)*:
  - Descriptive (the most widely used with hundreds of millions of unique metadata records in digital repositories, WorldCat, etc.
  - Administrative
  - Technical
- **Authority** *(represents other entities that information/data resources are related to):*
  - *Persons, organizations, projects*
  - *Topics, genres, & other ABOUTness & ISness terms*
  - *Places, time periods*

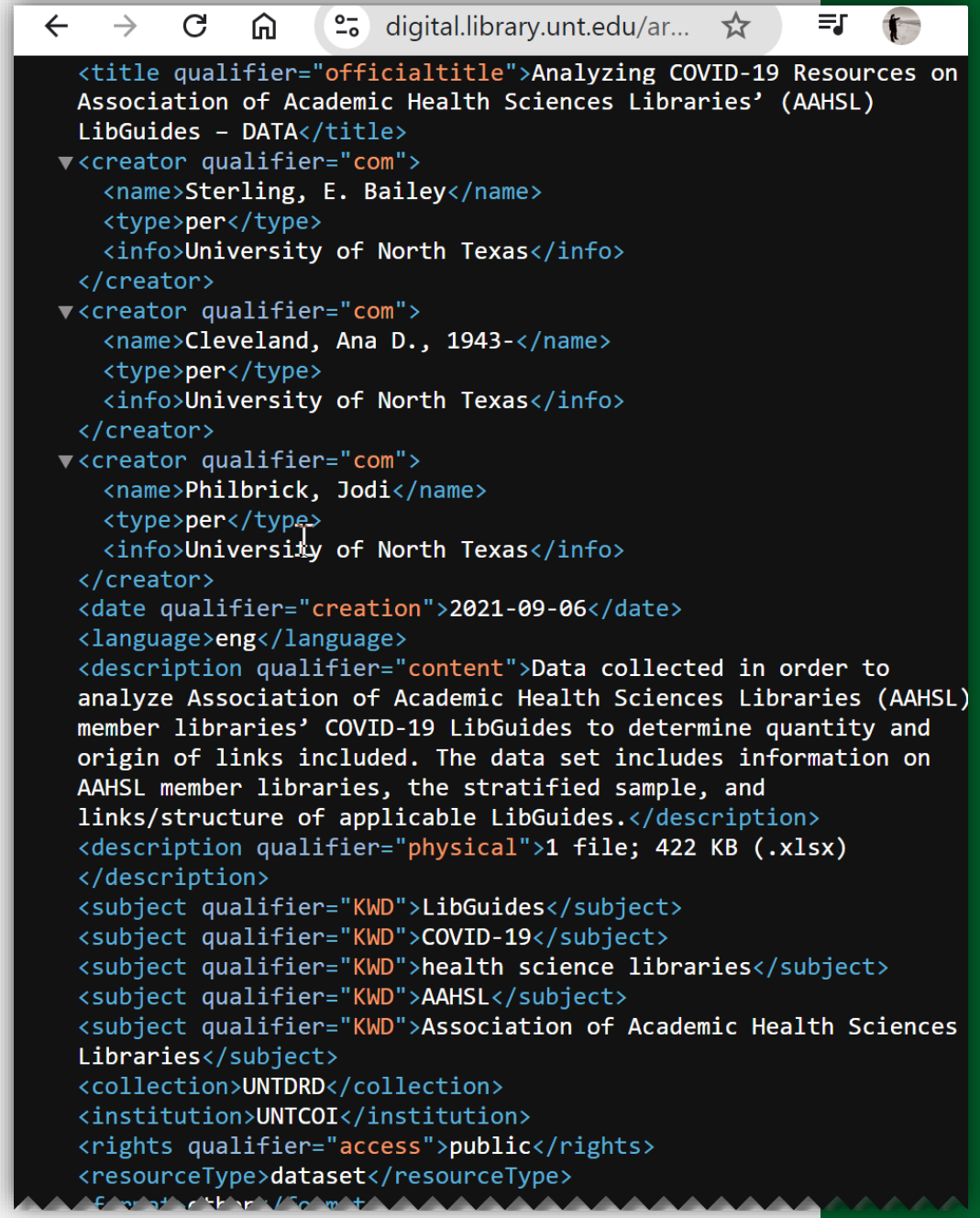
Descriptive Metadata Records

Structured machine-readable documents that **describe** data and various information resources, for example:

- a dataset collected and analyzed in a research project
- a journal article resulting from this research project

Records consist of **metadata fields** that **represent attributes important for discovery**:

- title; date(s) of creation, modification, acceptance, etc.; creator(s), contributor(s), publisher(s), rights holder(s); subjects (topical, geographical and temporal *ABOUTness*); format, type (*ISness*); audience, access and reuse rights, etc.



-- Public - No restrictions on sharing --

Metadata Standards

1. **Data content standards** that guide creation of metadata records for various user communities
   - **CCO** (museums), **DACS** (archives), **RDA** (libraries & beyond), etc.
2. **Data value standards** that provide guidelines and controlled vocabularies for consistent representation in metadata records and enable collocation and disambiguation of results.
   - **DDI SP** (dataset sampling procedure terms), **LCGFT** (genre terms), **MESH** (medical subject terms), **OLAC LSV** (linguistic subject terms), **TGN** (geographic names), **UDC** (classification codes), **ULAN** (names of art-related persons organizations, projects) & many more
3. **Data encoding / transmission standards** that enable sharing, exchanging and reusing metadata records:
   - Metadata schemes (**Dublin Core**, **MARC**, & many more),
   - Interchange formats/protocols (**JSON**, **OAI-PMH**, **RDF/XML**, **Z39.50**, etc.)

-- Public - No restrictions on sharing --

Quality of Metadata

The higher the quality of metadata records is, the more functional metadata is in:

- serving the very **goals** of metadata creation: providing the means for discovery and reuse of information and data
- supporting the *[human and machine]* **user tasks**

User Tasks and Metadata Support for them

3.3 User Tasks Definitions

Table 3.2 Definitions of User Tasks

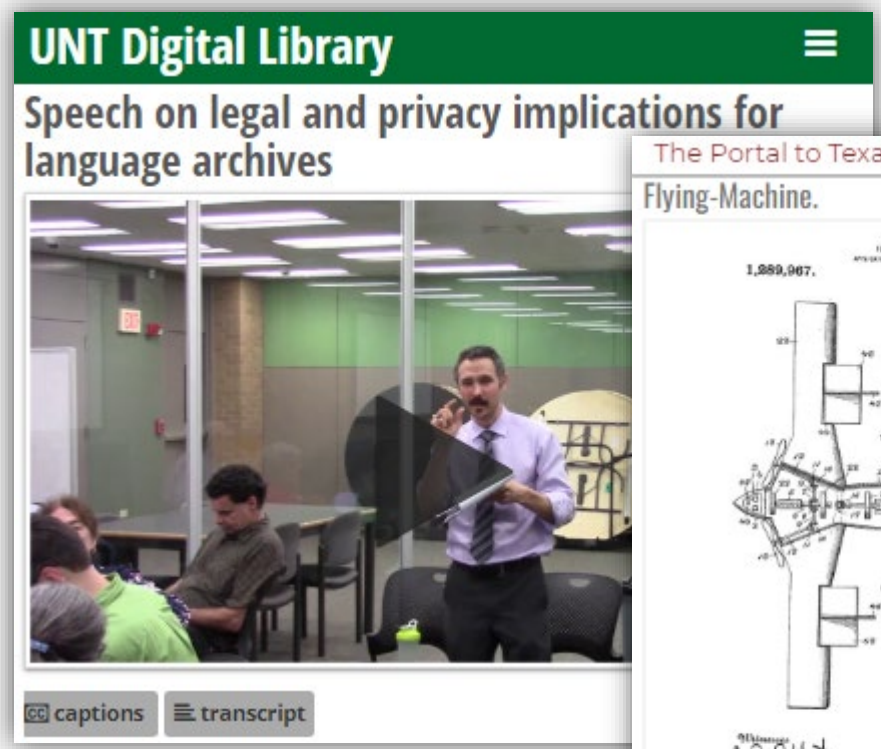
Task	Definition	Comment
Find	To bring together information about one or more resources of interest by searching on any relevant criteria	The *find* task is about searching. The user's goal is to bring together one or more instances of entities as the result of a search. The user may search using an attribute or relationship of an entity, or any combination of attributes and/or relationships. To facilitate this task, the information system seeks to **enable effective searching** by offering appropriate search elements or functionality.
Identify	To clearly understand the nature of the resources found and to distinguish between similar resources	The user's goal in the *identify* task is to confirm that the instance of the entity described corresponds to the instance sought, or to distinguish between two or more instances with similar characteristics. In "unknown item" searches, the user also seeks to recognize the basic characteristics of the resources presented. To facilitate this task, the information system seeks to **clearly describe the resources** it covers. The description should be recognizable to the user and easily interpreted.
Select	To determine the suitability of the resources found, and to be enabled to either accept or reject specific resources	The *select* task is about reacting to possible options. The user's goal is to make choices, from among the resources presented, about which of them to pursue further. The user's secondary requirements or limitations may involve aspects of content, intended audience, etc. To facilitate this task, the information system needs to **allow/support relevance judgements** by providing sufficient appropriate information about the resources found to allow the user to make this determination and act on it.
Obtain	To access the content of the resource	The user's goal in the *obtain* task is to move from consulting a surrogate to actually interacting with the library resources selected. To fulfill this task, the information system needs to either provide direct links to online information, or location information for physical resources, as well as any instructions and access information required to complete the transaction or any restrictions on access.
Explore	To discover resources using the relationships between them and thus place the resources in a context	The *explore* task is the most open-ended of the user tasks. The user may be browsing, relating one resource to another, making unexpected connections, or getting familiar with the resources available for future use. The *explore* task acknowledges the importance of serendipity in information seeking. To facilitate this task the information system seeks to **support discovery** by making relationships explicit, by providing contextual information and navigation functionality.

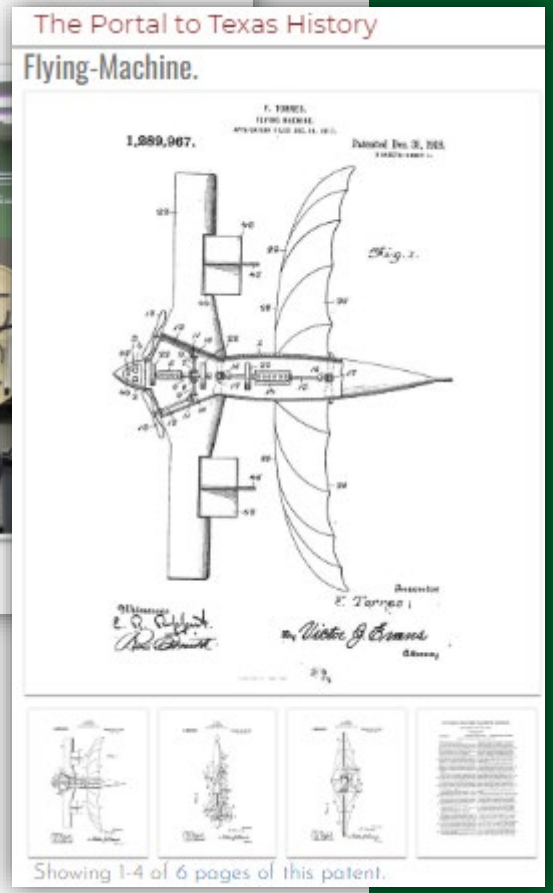
https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

Key Information Sources for Evaluation (*and Creation*) of a Metadata Record: 1

The information resource that the metadata record describes

- Detailed examination when feasible (e.g., for photographs, posters, conference papers, short videos/audios, patents, etc.)
- Cursory examination when detailed examination is not feasible (e.g., more extensive resources such as research monographs, longer audio recordings or video recordings)
  - Table of contents, abstract/summary, information on disk container, preview of an AV resource, transcript, etc.



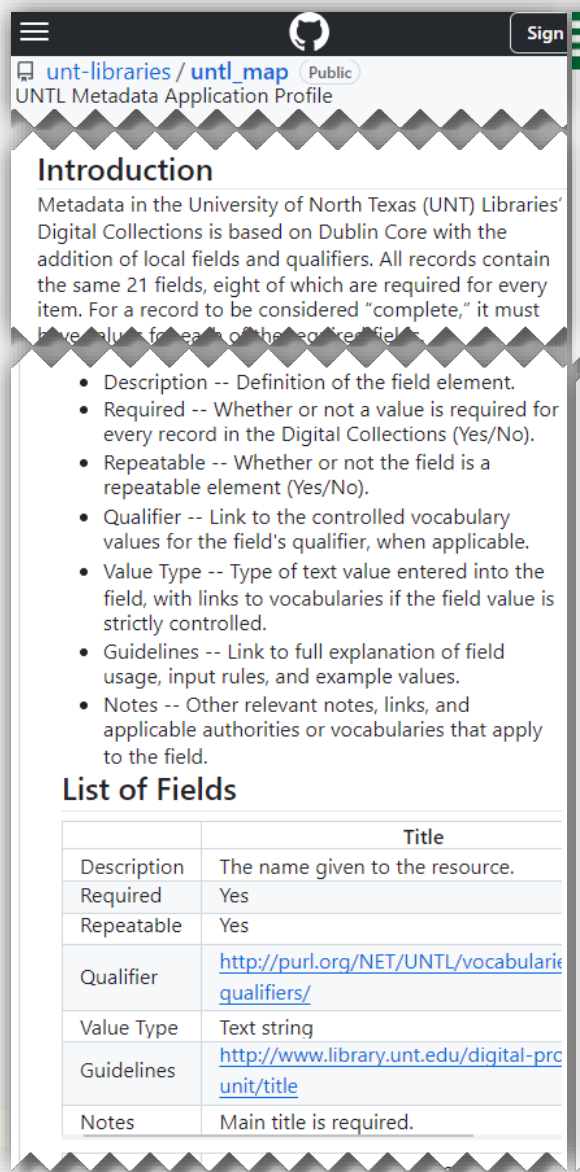


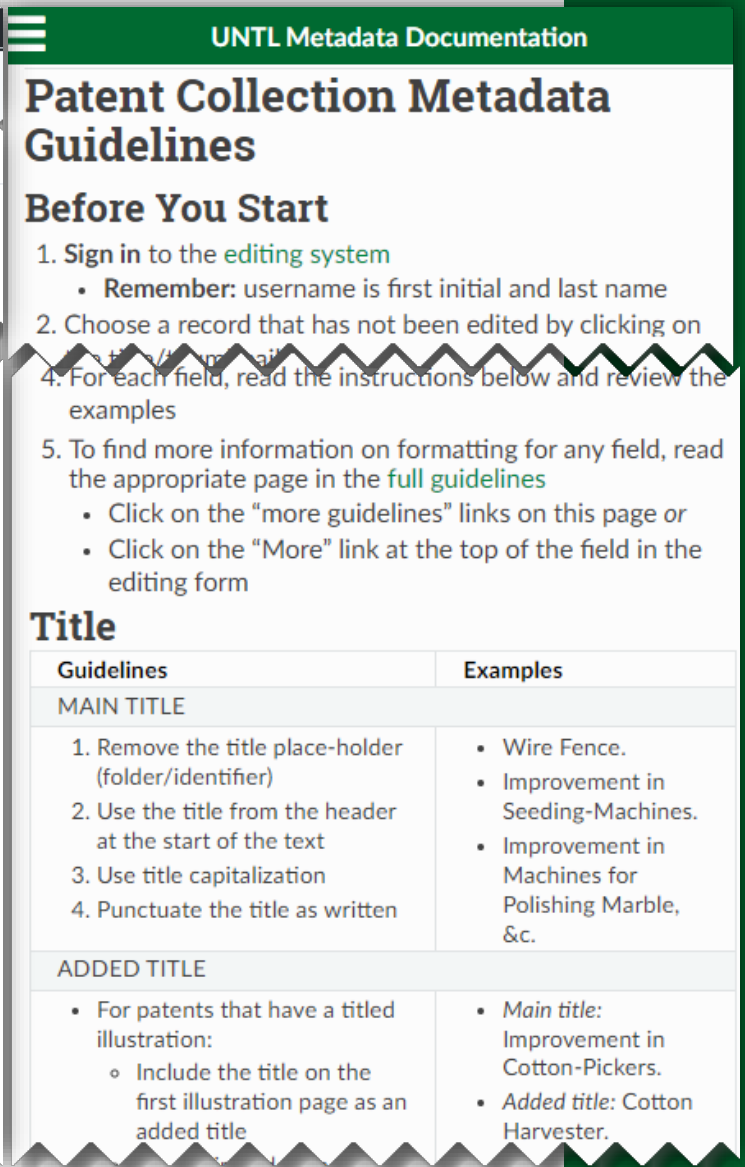
-- Public - No restrictions on sharing --

Key Information Sources for Evaluation (*and Creation*) of a Metadata Record: 2

Documentation of a metadata scheme used:

- Structure: list of metadata *elements* (and if applicable: their *subelements*, element *attributes*)
- Semantics specifications (definitions of metadata elements)
- Syntax specifications (encoding of metadata elements)
- Metadata creation guidelines:
  - **general** (for entire repository that the resource and its metadata record are from)
  - **specific** (for the collection of certain kinds of resources)





-- Public - No restrictions on sharing --

Metadata Quality Evaluation: Criteria (aka Quality Measurement Indicators)

Influential Bruce & Hillmann (2004) framework of metadata quality includes 7 criteria:

1. Completeness
2. Accuracy
3. Provenance
4. Conformance to Expectations
5. Logical Consistency/ Coherence
6. Timeliness
7. Accessibility

Refined by Zeng & Qin (2022) as CCCD Quality Measurement Indicators:

1. **C**ompleteness
2. **C**orrectness
3. **C**onsistency
4. **D**uplication analysis

in part based on earlier findings that *Completeness*, *Accuracy*, and *Consistency* are the most important criteria (e.g., **Park, 2009**)

-- Public - No restrictions on sharing --

Accuracy/Correctness

- **What (if any) descriptive metadata fields in the record:**
  a. contain the data that **misrepresents** this resource?
  b. contain **misspellings** or typographical errors in the data values?
  c. contain the **misplaced** data that according to the general and collection-specific metadata guidelines should have been entered in another field (specify that other field)?
  d. Contain the **mis-formatted** data -- use the formatting of the data value that is different from the formatting suggested for this field by the general and collection-specific metadata guidelines?

Completeness

- How many descriptive metadata fields – available for describing this type of resource in this metadata scheme according to metadata guidelines – are used in this record?
- What is the total number of all descriptive metadata field instances used in this record?
- What (if any) applicable descriptive metadata fields:
  a. are not included in the record?
  b. fail to include additional instances when applicable for representing this resource?
  c. contain incomplete data value (e.g., overly brief for adequate representation of this resource)?

Consistency

- Are the controlled vocabulary terms used? Provide specifics for the failure to use controlled vocabulary terms where applicable.
- Are the data values of the same kind – e.g., personal names, dates, etc. – entered in the same format across fields of the record? Provide specifics for inconsistencies observed.

Questions to answer in Step 1 of metadata quality analysis: Record-level evaluation

Example metadata record (patent) evaluation

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<metadata>
  <title qualifier="officialtitle">Spring Bed-Bottom.</title>
  <title qualifier="addedtitle">Bed Bottom</title>
  <creator qualifier="inv">
    <name>Meriwether, William Hunter</name>
    <type>per</type>
    <info>Wm. H. Meriwether, of New Braunfels, Texas.</info>
  </creator>
  <contributor qualifier="wit">
    <name>Campbell, T.</name>
    <type>per</type>
  </contributor>
  <contributor qualifier="wit">
    <name>Gritzner, M. C.</name>
    <type>per</type>
  </contributor>
  <publisher>
    <name>United States. Patent Office.</name>
    <location>[Washington D.C.]</location>
  </publisher>
  <date qualifier="creation">1854-08-08</date>
```

?

17 descriptive fields *(top-level)*; 24 instances *(e.g., 4 of Subject)*

```
<date>1854-08-08</date>
<language>eng</language>
<description qualifier="content">Patent for the improvement of bedsteads using a spring bottom-construction with zig-zag wire, including illustration.</description>
<description qualifier="physical">[1], 1 p. : ill. ; 23 cm.</description>
<subject qualifier="LCSH">Patents -- Texas.</subject>
<subject qualifier="UNTL-BS">Science and Technology</subject>
<subject qualifier="UNTL-BS">Social Life and Customs - Furnishings - Furniture</subject>
<subject qualifier="KWD">beds</subject>
<primarySource>1</primarySource>
<coverage qualifier="placeName">United States - Texas - Comal County - New Braunfels</coverage>
<collection>TXPT</collection>
<institution>UNTGD</institution>
<rights qualifier="license">pd</rights>
<resourceType>text_patent</resourceType>
<format>text</format>
<identifier qualifier="LOCAL-CONT-NO">11484</identifier>
<note qualifier="nonDisplay">comment: Descriptive metadata template by htarver 2011-07-19.</note>
```

!

!

!

Info subfield missing in 2 instances of Contributor field for witnesses but included in Creator

Completeness issue in **Date**: missing qualifier="**accepted**"

Consistency in subject representation: 1 **UNTL-BS** heading and 1 **keyword** but no **LCSH** added by metadata editor(s)

Accuracy/Correctness issue in **Identifier** (should be PAT-NO)

- Is the format of data values of the same kind – personal names, dates, etc. – **consistent** across the sample?
- **What % of records fail to:**
  - Use controlled vocabulary data values for subject representation, etc.?
  - Include descriptive metadata field(s) applicable for representing the resource?
  - Include additional instances of a repeatable descriptive metadata field when applicable for representing the resource?
- **What % of metadata records contain data values that:**
  - Are **incomplete** (e.g., overly brief for adequate representation)
  - **Misrepresent** the resource?
  - Include **misspellings or typographical errors**?
  - Are **misplaced** (used in the wrong field) according to the metadata guidelines?
  - Are **mis-formatted** (do not comply to formatting requirements in the metadata guidelines)?
- **What is the average total number of all descriptive metadata:**
  - **fields** – available for describing these resources in the metadata scheme according to metadata guidelines – used per record?
  - **field instances** used per record?

Questions to answer in metadata quality evaluation Step 2: Comparative analysis

Metadata change is part of metadata quality assurance (and evaluation)

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

To keep up with "environmental" changes (Thornburg & Oskins, 2007), including:

- Growth in certain types/formats and subject matter of materials in repositories
- Changes in the content & location of fluid materials (e.g., websites)
- Changing goals of hosting & contributing institutions
- **Evolution of national and international metadata standards:**
  - Data content standards and metadata element sets
  - Data value standards: classification systems & controlled vocabularies.

-- Public - No restrictions on sharing --

UNT dissertations focused on metadata quality (& accessible in full text)

- Aljalahmah, S. H. (2021). *The Status of the Organization of Knowledge in Cultural Heritage Institutions in Arabian Gulf Countries*. **https://digital.library.unt.edu/ark:/67531/metadc1833519/**
- Hasenyager, R. L., Jr. (2015). *Convenience to the Cataloger or Convenience to the User?: An Exploratory Study of Catalogers' Judgment*. **https://digital.library.unt.edu/ark:/67531/metadc799476/**
- Phillips, M. E. (2020). *Exploring the Use of Metadata Record Graphs for Metadata Assessment*. **https://digital.library.unt.edu/ark:/67531/metadc1707350/**
- Snow, K. (2011). *A Study of the Perception of Cataloging Quality Among Catalogers in Academic Libraries*. **https://digital.library.unt.edu/ark:/67531/metadc103394/**
- Zavalin, V. I. (2020). *Exploration of RDA-Based MARC21 Subject Metadata in WorldCat Database and Its Readiness to Support Linked Data Functionality*. **https://digital.library.unt.edu/ark:/67531/metadc1707353/**

-- Public - No restrictions on sharing --

Some examples of peer-reviewed publication venues for metadata quality research

journals, e.g.:	conference proceedings, e.g.:
• *Journal of Library Metadata* • *Journal of the Association for Information Science and Technology (JASIS&T)* • *Cataloging and Classification Quarterly* • *The Electronic Library* • *International Journal of Metadata, Semantics, and Ontologies* • *Library Resources and Technical Services* • … OTHER.	• Dublin Core Metadata Initiative (DCMI) conference • ASIS&T annual meeting • iSchools conference (iConference) • Joint Conference on Digital Libraries (JCDL) • International Conference on Knowledge Management • … OTHER.

-- Public - No restrictions on sharing --

Recently published metadata quality research of UNT-affiliated researchers: 2017-2024 book chapters & journal articles

Aljalahmah, S.H., & Zavalina, O.L. (2024). Student-created Dublin Core metadata representing Arabic language eBooks: Comparison of individual and group work outcomes. *Journal of Education for Library and Information Science (JELIS), 65*(3), 325-344.

Aljalahmah, S.H., & Zavalina, O.L. (2024). Audiovisual resources metadata: Analysis of records originating from novice metadata creators in Kuwait. *Journal of Library Metadata (JLM), 24*(3), 189-214.

Phillips, M., Zavalina, O.L., & Tarver, H. (2020). Exploring the utility of metadata record graphs and network analysis for metadata quality evaluation and augmentation. *International Journal of Metadata, Semantics, and Ontologies, 14*(2), 112-124.

Zavalin, V., Zavalina, O.L., & Miksa, S.D. (2021). Exploration of subject representation and support of Linked Data in recently created library metadata: Examination of most widely held WorldCat bibliographic records. *Library Resources and Technical Services, 65*(4), 154-165.

Zavalina, O.L., & Burke, M. (2021). Assessing skill-building in metadata instruction: Quality evaluation of Dublin Core metadata records created by graduate students. *Journal of Education for Library and Information Science, 62*(4), 423-442.

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 G. Chowdhury et al. (Eds.), *Transforming Digital Worlds, Lecture Notes in Computer Science 10766* (pp.645-651). New York: Springer.

Zavalina, O.L., & Zavalin, V. (2017). Identity management analysis: an empirical investigation into the state of library community's authority data conformance to the new standard. In *Knowledge Discovery and Data Design Innovation* (pp. 233-248). Hackensack, NJ: World Scientific.

Recently published metadata quality research of UNT-affiliated researchers: examples of 2020-2024 conference papers

Burke, M., & Zavalina, O.L. (2020). Descriptive richness of free-text metadata: a comparative analysis of three language archives. *Proceedings of the Association for Information Science and Technology, 57*(1)

Roeschley, A., Kim, J., & Zavalina, O.L. (2020). An exploration of contributor-created Description field in participatory archives. *iConference 2020 Proceedings.*

Paterson III, H.J. (2023). OLAC and serials: An appraisal. *Proceedings of the 2nd International Workshop on Digital Language Archives*.

Zavalin, V.I. (2024, in print). Skill-building in subject representation: Assessing learning outcomes through analysis of student-created metadata. *ALISE 2024 Conference Proceedings*.

Zavalin, V.I. (2023). Ukrainian archival metadata in WorldCat: Exploratory analysis. *Proceedings of the 2nd International Workshop on Digital Language Archives.*

Zavalin, V.I., & Zavalina, O. L. (2024, in print). Exploring accuracy, completeness, and consistency of VRA Core 4.0 paintings metadata. In *DCMI 2024 Conference Proceedings*

Zavalin, V.I., & Zavalina, O. L. (2023). Exploration of accuracy, completeness, and consistency in metadata for physical objects in museum collections. *iConference 2023 Proceedings*.

Zavalin, V., Zavalina, O.L., & Safa, R. (2021). Patterns of subject metadata change in MARC21 bibliographic records representing video recordings. *Proceedings of the Association for Information Science and Technology, 58(1).*

Zavalina, O.L. (2021). Notes fields of metadata records generated by the Children's and Young Adults' Cataloguing Program: Exploration of support for general and specific needs of school library users. *Proceedings of the International Association of School Libraries IASL 2021.*

Simple recent example of metadata quality study

iConference 2024
Wisdom Well–being Win–win

How Accurate and Complete is Digital Audio Resources Metadata?
Examination of Metadata Created by Kuwaiti Students

Submission ID: 200

Dr. Saleh Aljalahmah (sh.aljalahmah@paaet.edu.kw) & Dr. Oksana L. Zavalina Oksana.Zavalina@unt.edu)

UNT EST. 1890

Online Audio Resources

- are widely used to learn about important topics (incl. well-being).
- are generated & shared (incl. on the popular social media platforms) in very large and rapidly increasing volumes
- metadata's crucial role in enabling their discovery is long recognized, gains even more importance.

Problem Statement

- Lack of empirical studies.
- No research examined representation of audio materials in student-created metadata, (incl. in Arabian Gulf)

Our Study: Research Questions

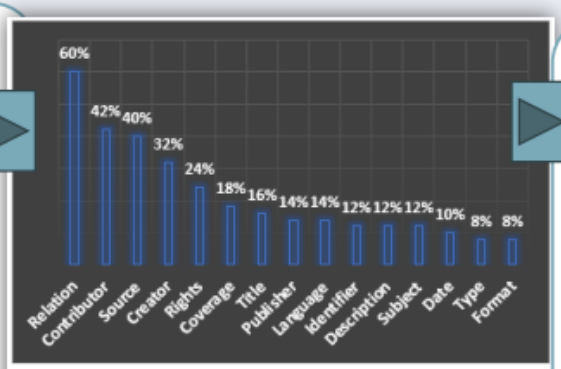
- How accurate are student-created metadata records?
- How complete are student-created metadata records?
- How are the mistakes distributed across the Dublin Core (DCMES 1.1) 15 metadata fields?

Our Study: Data Collection & Analysis

- Collected all digital audio metadata records created by students of the **Kuwait's Public Authority for Advanced Education and Training (PAAET)** Library and Information Sciences Department undergraduate metadata course.
- Analysis focused on 2 criteria of metadata quality: ***accuracy*** and ***completeness*** *(defined by Bruce & Hillman, 2004, definitions adapted & extended by us in Aljalahmah & Zavalina, 2023)*

Highlighted Findings

- **Average no. of errors per metadata record**: 3.26
- **5 most error-prone metadata fields** (had mistakes in 24-60% records):
  - *Relation, Contributor, Source, Creator,* and *Rights*
- **3 least error-prone metadata fields** (8-10% of records):
  - *Format, Type,* and *Date*
- ***Completeness*** **errors**: 1.34/record; observed in all fields but 2 (*Coverage* & *Type*)
- ***Accuracy*** **errors are more common:** observed in all fields, 58.9% of all errors, average of 1.92/record.
- 3 categories (with %) of *accuracy* errors:
  - **misinterpreting the applicability of metadata field** (41.67%)
  - **entering incorrect data value** (42.71%), or
  - **wrong formatting of the data value** (15.62%).



Metadata mistakes distribution in the fields of PAAET student-created Dublin Core metadata records representing an online audio recording (% of records).

Conclusions & Future Research

- Identified DCMES 1.1 metadata fields most and least prone to *accuracy* and *completeness* errors
  - 40% of most error-prone DCMES 1.1 fields (*Relation* & *Source*) are known to be conceptually difficult to apply for beginners
- Relatively small dataset but generalizable to this population (PAAET undergrad metadata students); continue in future semesters
- To obtain a better understanding of the quality of beginner-created metadata that represents online digital audio resources, comparative studies are needed that would examine metadata created:
  - in graduate & undergraduate environments, and in multiple countries.
  - in different metadata schemes beyond Dublin Core.

Steady growth of metadata quality research importance is accelerated by the AI revolution

Artificial Intelligence (AI) & Machine Learning (ML) affect metadata landscape in multiple ways:

In the context of FAIR principles for data in science (& other research fields)

→ the need for evaluating conformance of existing metadata to FAIR principles

As a source of demand for representing AI models with metadata to promote transparency, explainability, accountability, contestability

→ the need for development of: AI model metadata standard, evidence-based quality guidelines

As a tool considered for generation/augmentation of descriptive metadata for information/data resources

→ the need for quality assessment of this AI-created metadata

FAIR Guiding Principles for scientific data management and stewardship and metadata quality

For Findability, Accessibility, Interoperability, and Reuse of digital assets

- "F1: (Meta)data are assigned globally unique and persistent identifiers
- F2: Data are described with rich metadata
- F3: Metadata clearly and explicitly include the identifier of the data they describe
- F4: (Meta)data are registered or indexed in a searchable resource
- A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
  - A1.1: The protocol is open, free and universally implementable
  - A1.2: The protocol allows for an authentication and authorisation procedure where necessary
- A2: Metadata should be accessible even when the data is no longer available
- I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2: (Meta)data use vocabularies that follow the FAIR principles
- I3: (Meta)data include qualified references to other (meta)data
- R1: (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1: (Meta)data are released with a clear and accessible data usage license
  - R1.2: (Meta)data are associated with detailed provenance
  - R1.3: (Meta)data meet domain-relevant community standards"

Metadata Development for AI Models: 1

Two kinds of metadata needed:

1. Model cards

standardized documentation about the Machine Learning **model** itself

Model card components (Mitchell et al., 2019):

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases

Includes common kinds descriptive metadata

- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Metadata Development for AI Models: 2

Two kinds of metadata needed:

2. Data cards

standardized documentation about the **dataset used** in the Machine Learning model development

Data card template (Pushkarna, Zaldivar, & Kjartansson, 2022): 31 components

Includes common kinds descriptive metadata

(1) The publishers of the dataset and access to them
(2) The funding of the dataset
(3) The access restrictions and policies of the dataset
(4) The wipeout and retention policies of the dataset
(5) The updates, versions, refreshes, additions to the data of the dataset
(6) Detailed breakdowns of features of the dataset
(7) Details about collected attributes which are absent from the dataset or the dataset's documentation
(8) The original upstream sources of the data
(9) The nature (data modality, domain, format, etc.) of the dataset
(10) What typical and outlier examples in the dataset look like
(11) Explanations and motivations for creating the dataset
(12) The intended applications of the dataset
(13) The safety of using the dataset in practice (risks, limitations, and trade-offs)
(14)Expectations around using the dataset with other datasets or tables (feature engineering, joining, etc.)
(15) The maintenance status and version of the dataset
(16) Difference across previous and current versions of the dataset
(17) The data collection process (inclusion, exclusion, filtering criteria)
(18) How the data was cleaned, parsed, and processed (transformations, sampling, etc.)
(19) Data rating in the dataset, process, description and/or impact
(20) Data labeling in the dataset, process, description and/or impact
(21) Data validation in the dataset, process, description and/or impact
(22) The past usage and associated performance of the dataset (eg. models trained)
(23) Adjudication policies and processes related to the dataset (labeler instructions, inter-rater policy, etc.)
(24) Relevant associated regulatory or compliance policies (GDPR, licenses, etc.)
(25) Dataset Infrastructure and/or pipeline implementation
(26) Descriptive statistics of the dataset (mean, standard deviations, etc.)
(27) Any known patterns (correlations, biases, skews) within the dataset
(28) Human attributes (socio-cultural, geopolitical, or economic representation)
(29) Fairness-related evaluations and considerations of the dataset
(30) Definitions and explanations for technical terms used in the Data Card (metrics, industry-specific terms, acronyms)
(31) Domain-specific knowledge required to use the dataset

Use of Generative AI in Descriptive Metadata Creation/Augmentation

So far, several reports published in peer-reviewed venues:

- Mostly experimented with the ChatGPT tool and evaluated generation of data values in **individual fields of metadata records** – those intended for *ABOUTness* representation:
  - subject terms (e.g., Ganadi et al., 2023)
  - classification numbers/codes (e.g., Bodenhamer, 2023)
- Included only 2 analyzes of AI-created **metadata records**:
  - compared with human-created MARC & Dublin Core records for same resources (Brzustowitz, 2023)
  - also, new paper, published 2 weeks go: Taniguchi (2024)

Need many more AI-generated metadata quality studies, where:

- Entire metadata records are generated an analyzed
- Metadata is truly created by AI tools themselves: *no pre-existing publicly available metadata for the same information resource*
- Human metadata experts use their knowledge of metadata standards to meaningfully evaluate the quality

Interested in Doing Metadata Quality Research?
You need metadata background to adequately prepare

These regularly offered UNT graduate courses help develop robust understanding of metadata practices, principles, and standards, obtain necessary practical experience of metadata creation/analysis:

- INFO 5223 Metadata 1
- INFO 5385 Community Language Archiving and Curation for Information Professionals *(team-developed; 2 out of 4 modules have metadata focus, including evaluation)*
- INFO 5210 Cataloging & Classification 1 (Dr. Shawne Miksa)
- INFO 5220 Cataloging & Classification 2
- INFO 5212 Intro to Dewey Decimal Classification (Dr. Shawne Miksa)
- INFO 5224 Metadata 2 *(offered in Spring semesters, includes a 4-week metadata quality module in which students collect and comparatively evaluate 2 sample of metadata, report metadata quality assessment results)*

Time for Roundtable Discussion and Q & A !

-- Public - No restrictions on sharing --