Metadata for Everyone

Identifying Metadata Quality Issues Across Cultures

Julie Shi – Library Publishing Forum 2023

Is Metadata for Everyone?

Notes on metadata

Notes on metadata in the scholarly ecosystem

Exploring cultural quality in scholarly publishing metadata

Notes on metadata

Metadata (should) make local research visible and accessible for all audiences in our global scholarly ecosystem.

Making metadata that is considered high quality requires considerable time and labor.

Many of our technical systems and standards are built around the English language and Western ideas of knowledge and scholarly practice.

Notes on metadata in the scholarly ecosystem

Researcher Identity

- disambiguating names & collocating work
- expanding research communities

Research Quality and Diversity

- broadening discovery, sharing, and (re)use
- amplifying voices and perspectives

Review, Promotion, Tenure

- research assessment based on metrics and citations (e.g., h-index, impact factors)

Exploring cultural quality in scholarly publishing metadata 🔍

1. Understand who is left out of metadata, even when standards are perfectly applied.

2. Examine how metadata systems capture the global nature of scholarship (or not).

3. Document the various ways that individual or community identity is asserted or retained in metadata.

Exploring cultural quality in scholarly publishing metadata

Discovery
- Establish scope and priorities
- Identify data sources

Phase 1
- Manual review of records to identify issues
- Determine a typology of common cultural issues

Phase 2
- Programmatically measure metadata quality based on identified issues

Identifying Metadata Quality Issues Across Cultures

Considering metadata quality

Defining scope

Close reading the records

Finding fields of focus

Considering metadata quality

> "Metadata **quality** is a multidimensional concept" (Zeng & Qin 2016, 319)

Technical quality
Bruce & Hillmann (2004)

Cultural quality
Shiraishi (2019)

- Completeness
- Accuracy
- Conformance
- Usability

- Potential for harm or disservice to individuals or communities

Defining scope 🔍

🧠 Conceptual scope

- **Cultural issues**: Issues that impact, or have the potential to impact, the representation of identities, roles, intentions, and other factors specific to social, regional, or research cultures

⏳ Practical scope

- **Sample size**: 427 records from 68 DOI prefixes
  - **Publishing with OJS:** 51 records from 17 DOI prefixes

Close reading the records 🔍

> "to examine a metadata record, which can be regarded as a surrogate of an item, **a comparison between the surrogate and the original item** is absolutely necessary" (Zeng & Qin 2016, 322)

 + + +
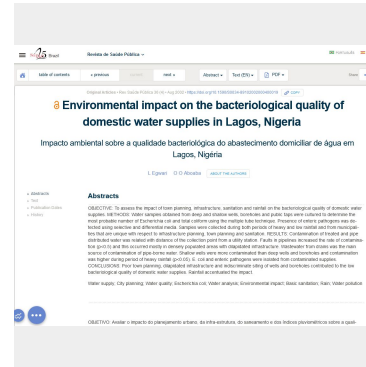
Record   Item PDF   Item landing page   Container

Finding fields of focus
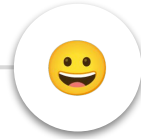
🔍

Initial review of 61 records to determine metadata fields to focus on

📜 Item

- Title
- Abstract

😀 Person

- Given Name
- Family Name
- Affiliation

📕 Container

- Publisher
- Title
- Language
- Subject

Issues of Cultural Quality in Metadata

- Distinguishing issues of technical vs. cultural quality

- Surfacing a range of cultural issues

- Defining a typology of issues

Surfacing a range of cultural issues

5 Main Forms

- Value absent
- Value in record does not match information in item
- Value does not match parameters of field
- In/completeness of value
- Value incorrectly input

32 Unique Issues

- registered URL invalid
- registered URL out of date
- value absent
- translation absent
- value in original language absent
- incorrectly input
- original-title used incorrectly
- lang attribute absent
- multiple languages in single field
- value incomplete
- value in record does not match info on website
- vor license terms absent
- additional persons listed
- author/s absent
- not all authors listed
- all authors listed as first
- first author not identified
- ORCIDs absent
- not all persons listed
- language style absent
- only provides initial/s
- affiliation absent for all authors
- affiliation absent for all editors
- organization as acronym only
- affiliations presented as authors
- multiple values in single field
- location absent
- Inaccurate (language or subject)
- not all publishers listed
- related orgs absent
- container title outdated
- subtitle absent

Distinguishing issues of technical vs. cultural quality

Record
affiliation:
  0:
    name: "a Department
          of Chemistry ,
          Humboldt-
          University ,
          Hessische..."

Incorrectly input: with
footnote marker

Record
affiliation:
  0:
    name: "Universidade...
          Brazil"
...
affiliation:
  0:
    name: "Instituto...
          Brasil"

Affiliation: country name
in multiple languages

Record
author:
  0:
    given: "Assoc."
    family: "Prof. Vichien
          Srimuninnimit"

Person: name and title
input incorrectly

Defining a typology of issues 🔍

Language

Naming

Contribution

Geography

Seniority

Prestige

Language

The languages and scripts of values and their identification using language and style attributes

Record
publisher: "Japanese Society for Pharmacoepidemiology"

…
title:
  0: "製造販売後調査と安全対策における製薬企業の取り組みと課題"

…
container-title:
  0: "Japanese Journal of Pharmacoepidemiology/Yakuzai ekigaku"

…
language: "en"

Item
Title, abstract, and full text in Japanese only

👋

Naming

Recording of individual and organizational names in accordance with linguistic and cultural conventions

Record
author:
 0:
  family: "이창환"

Full name in Family Name field

Record
author:
 0:
   given: "Hui-bin"
   family: "ZHANG"

Romanized Family Name in all caps

Record
author:
 2:
   given: "Yuliadi"
   family: "Yuliadi"

Name with one part repeated

Record
publisher: "Institute of Research and Community Services Diponegoro University (LPPM UNDIP)"

Different languages for full name and acronym

Contribution

Acknowledgment of contributors to the creation and publication of the item and its contents

Record
No author field present

Item
One or more authors

Record
author:
  0:
    given: "Brajpal"
    family: "Tyagi"

Item
Two authors

Some or all authors absent

Record
assertion:
  5:
    value: "© 2019 Institute of Rock and Soil
            Mechanics, Chinese Academy of Sciences.
            Production and hosting by Elsevier B.V."
    name: "copyright"
    label: "Copyright"

Copyright holder noted (unusual)

Geography

Documentation for geographic locations, which gesture to social, cultural, and linguistic positions

Record
publisher: "IAIN Langsa"

Publisher location absent

Record
publisher: "FapUNIFESP (SciELO)"

Publisher location absent

Record
publisher: "Elsevier BV"
...
author:
  0:
    ...
    affiliation: []
  X:
    ...
    affiliation: []

Pub location and affiliations absent

Seniority

Stylistic and content-based interventions to reflect the position of certain authors

Record	Record
author:	author:
0:	0:
...	...
sequence: "first"	sequence: "additional"
1:	1:
...	...
sequence: "first"	sequence: "additional"

All authors listed as first or additional

Record	Item
0:	
given: "IRAM" family: "MANZOOR" ...	DR. IRAM MANZOOR Associate Professor ...
1:	
given: "Azeez" family: "Bukhari"	Mr. F. S. Azeez Bukhari 4th Year MBBS

Faculty names in all caps, student names in regular case

🏫 Prestige

Significance of institutional representation or reputation

Record
author:
 0:
  name: "Research Scholar: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan"
  sequence: "first"
  affiliation: []

1:
 given: "Zawar"
 family: "Hussain"
 sequence: "additional"
 affiliation: []
2:
 name: "Assistant Professor: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan"

Reflections & Next Steps 🔍

Interpreting the issues in the bigger picture

Understanding the limitations of this review

Next steps for *Metadata for Everyone*

Interpreting the issues in the bigger picture

Many of the identified issues may really just be the result of poor practice. It is just as possible, however, that issues result from deliberate acts.

Metadata "contribute to a story we are telling about ourselves as individuals, as organizations, and as a community" (Jaffe 2020, 441).

More work is needed to understand and address the ways that systems and standards render communities, worldviews, and stories invisible.

Understanding the limitations of this review

This review is not comprehensive. The issues and patterns found are specific to my interpretation of this sample.

The ability to recognize issues and nuances often depends on familiarity with particular social, regional, and research cultures.

Close reading the records alone does not tell us which issues result from poor metadata practice and which result from deliberate interventions.

Next steps for *Metadata for Everyone* 🔍

Phase 2: Programmatically measure metadata quality based on identified issues

🤔 To what extent are the types of metadata issues identified in this study present in the scholarly record?

💻 Identify issues and measure quality and completeness for 100,000 metadata records

🔬 Analyze variations in quality by language of publication, publisher size, and other facets

Thank you to the team! 🔍

Project team:

- Juan Pablo Alperin (PKP / ScholCommLab / SFU)

- Mike Nason (PKP / UNB)

- Marco Tullney (PKP / TIB)

- Dennis Donathan II (PKP)

Lots of support from:

- The wonderful folks at Crossref!

Thank you for listening!

🔍

Comments, questions, conversations?

juli.shi@utoronto.ca

References

Bruce T. R., Hillmann D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. *ALA Editions*, 238-256.

Jaffe, R. (2020). Rethinking metadata's value and how it is evaluated. *Technical Services Quarterly*, 37(4), 432–443. DOI: 10.1080/07317131.2020.1810443

Shiraishi, N. (2019). Accuracy of identity information and name authority records. In J. Sandberg (Ed.), *Ethical Questions in Name Authority Control* (pp. 181-194). Library Juice Press.

Zeng, M. L., & Qin, J. (2016). Metadata quality: Measurement and improvement. In *Metadata* (2nd ed., pp. 317-346). American Library Association.