Metadata for Everyone: Identifying Metadata Quality Issues Across Cultures

Julie Shi
Presenter

Dennis Donathan II
Contributor

Abstract

Metadata is crucial to the dissemination and communication of research. Quality metadata facilitates discovery and access and provides contextual, technical, and administrative information in a standard form. Yet metadata are also sites of tension between sociocultural representations, resource constraints, and standardized systems. Formal and informal interventions may be interpreted as metadata quality issues, political acts to assert identity, or strategic curatorial choices to maximize discoverability and visibility. This presentation documents the work of Public Knowledge Project (PKP) and Crossref on the Metadata for Everyone project to understand how metadata quality, consistency, and completeness impact individuals and communities. Working from a sample of records known to have erroneous, incomplete, or otherwise imperfect metadata, we set out to identify and classify issues stemming from how metadata and communities press up against each other to intentionally reflect (or not) cultural meanings.

Keywords: metadata quality issues, scholarly publishing, cultural representation, identity

Introduction

Treating metadata records as informational objects in their own right, the Metadata for Everyone project examined metadata for scholarly publications to identify issues related to cultural identities and meanings and their associated implications. This paper begins by introducing the project goals and summarizing our scope and methods. We then discuss the issues and categories identified. We conclude with key takeaways, limitations, and next steps for our project.

Context

Metadata is generally understood to provide key bibliographic and domain information and to connect that resource to the larger knowledge ecosystem. At the same time, we may think about metadata as "contributing to a story we are telling about ourselves as individuals, as organizations, and as a community."[1] In this framing, metadata is inextricably tied to the people who create it and the people, places, and things that it describes. Through description and context, metadata functions as a narrative device to assert the social contexts and genealogies of our knowledge. In this sense, metadata is storytelling and identity building.

In both respects, metadata has the potential to make local research, communities, and stories visible and accessible for wider audiences in our global scholarly ecosystem. The key phrase here is "has the potential," because creating metadata of sufficient quality to enhance discoverability and access and to represent all our stories as we want them to be represented takes time and labor. For some metadata creators, this investment is much larger than for others.

This challenge is, in part, because many of our technical systems and standards are built around the English language and Western ideas of knowledge and scholarly practice.[2,3,4,5,6,7] This can look like English-language metadata requirements for indexing, geographic lists based on standardized lists that do not include Indigenous nations

and contested regions, or user interfaces that struggle with non-Latin scripts and diacritics.[8,9,10,11]

Whose stories do our current metadata systems and standards permit? Who is allowed to contribute to and shape the stories that are told about themselves and their knowledge?

Metadata for Everyone

As organizations that provide systems for managing scholarly metadata, the Public Knowledge Project (PKP) and Crossref sought to understand the ways in which individuals and communities actively seek to convey meaning and express identity through metadata. We were interested in identifying who is left out of metadata, even when standards are perfectly applied; where metadata systems and standards are falling short; and the unique ways that individuals have used metadata to assert or retain their identity in response to these shortcomings. We also wanted to determine how well our current metadata systems reflected our global communities of users.

The project involved three stages, beginning with an initial discovery phase between PKP and Crossref to establish scope and priorities and identify data sources. The Crossref application programming interface (API) was queried to retrieve a sample of records from the identified sources, and the 427 records pulled were analyzed to surface and categorize culturally-related issues in the first phase. The categories derived in this phase continued to inform work in the second phase to programmatically measure metadata quality based on the presence of these issues on a much larger scale.

Identifying Cultural Quality Issues

Common definitions of metadata quality in the literature attend to dimensions like the informational completeness and accuracy of the record and its conformance to standards and controlled vocabularies for use by

humans and machines alike.[12,13,14] Such issues are referred to from here on as those of "general quality." By contrast, the cultural quality of metadata is measured against the possibility that an aspect of the metadata could cause harm or disservice to a person or group and their work.[15]

Specifically, cultural issues are defined as 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. This definition is intentionally broad because "culture" can be interpreted in numerous ways and, with each interpretation, it inflects and impacts how knowledge is understood and research is conducted differently.

To identify issues of cultural quality, we embarked on a qualitative study of the sampled records. Where "a comparison between the surrogate and the original item is absolutely necessary" when evaluating metadata, each record was close read alongside the published item it described—usually a PDF for a journal article, the webpage for accessing that item, and contextual information about the larger work.[16] Close reading involved noting which fields or values were present or absent, as well as examining the content and form of values provided. Differences in content and form between the record, item, webpage, and work were also recorded.

We initially reviewed sixty-one records to identify fields more likely to relate to culture and identity. We found nine relevant fields: 1) abstract, 2) item title, 3) given name, 4) family name, 5) institutional affiliation, 6) publisher name, 7) title of work, 8) language/s accepted in the larger work, and 9) the subject headings applied to the larger work. Although we did not actively review for issues of general quality, these were also noted when found.

Findings

Across the 427 records and nine fields of focus, 4,859 issues of both general and cultural quality were found. Of these, 90 percent impact or have the potential to impact cultural meanings and identity. Among the cultural issues, thirty-two unique issues were identified and were organized into five main forms (see Table 1).

Table 1. List of thirty-two identified issues, organized by their five main forms

Form	Issue
Value absent	value absent
	translation absent
	value in original language absent
	language attribute absent
	language style absent: romanization only
	language style absent: romanization absent
	VoR license terms absent
	author/s absent
	not all authors listed
	ORCIDs absent
	not all persons listed
	absent for all authors
	absent for all editors
	not all publishers listed
	related orgs absent
	location absent
	subtitle absent
Value does not match information in the item	outdated
	registered URL out of date
	registered URL invalid
	value in record does not match information on container website
	inaccurate
Value does not match the parameters of the field	affiliations presented as authors
	multiple languages in single field
	multiple values in single field
	original-title used incorrectly: includes value in original language but item is not a translation
	original-title used incorrectly: value repeated
	all authors listed as first
	first author not identified
	input in all caps
	additional persons listed
Issues with completeness of the value	value incomplete
	only provides initial/s
	acronym only
Incorrectly input	Several types of errors

Issues were not equally common, with eight unique types comprising 75 percent of all general and cultural issues found; among the cultural issues specifically, they amounted to 83 percent. Of these eight, seven were of the "value absent" form, indicating a glaring absence in this sample that has cultural implications. Due to the nonrandom nature of the sample, the frequency of each unique issue

is less significant than the categories found. Counts are noted for transparency.

To better understand what is at stake along cultural lines, we classified the thirty-two issues to derive five key categories related to cultural meanings and identities:

- Language: Issues relate to the languages and scripts of values and/or the way in which they are identified using language and style attributes.
- Contribution: Issues relate to the acknowledgment of contributors to the creation and publication of the item and its contents.
- Naming: Issues relate to the recording of individual and organizational names in accordance with linguistic and cultural conventions.
- Status: Issues relate to stylistic and content-based interventions to capture the status, seniority, or prestige of individuals or institutions.
- Geography: Issues are caused by the absence or partial representation of physical location and its social and cultural associations.

These categories are conceptual and issues often fit into one or more categories depending on their context. Table 2 provides annotated examples of issues for each category.

Discussion

With issues across five categories of language, naming, status, contribution, and geography, and cutting across questions of consistency, completeness, and quality, it is apparent that metadata is not for everyone. The five key types, thirty-two unique issues, and five main forms represent just some of the ways that our metadata systems and standards are failing users.

Table 2. Annotated examples of issues found, organized by category

Example	Issues and reasoning
Language	
Record publisher: "Japanese Society for Pharmacoepidemiology" title: "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: "Japanese Journal of Pharmacoepidemiology/Yakuzai ekigaku" language: "en"	*Issues (fields)*: Value in original language absent (publisher, container-title; Multiple languages in single field (container-title); Inaccurate (language)) *Reasoning*: On the item, the item title, abstract, and full text are available in Japanese only, and the publisher and container title in Japanese and English. Although the language of the record is noted as "en," the record includes a mix of Japanese script, romanized Japanese, and English.
Contribution	
Item landing page Reviewed Work: Zarte Liebe fesselt mich. Das Liederbuch der Fürstin Sophie Erdmuthe von Nassau-Saarbrücken by Ludwig Harig, Wendelin Müller-Blattau Review by: Ulla Enßlin *Record* author-1: given: "Ulla" family: "Enßlin" author-2: given: "Ludwig" family: "Harig" author-3: given: "Wendelin" family: "Müller-Blattau" author-4: given: "Ulla" family: "Ensslin" author-5: given: "Wendelin" family: "Muller-Blattau"	*Issues (fields)*: Additional persons listed (author-2, author-3); Incorrectly input: repeated values (author-4, author-5) *Reasoning*: This item is a book review. Authors of the work reviewed are listed in the record alongside the reviewer (author-1). Two author names (author-1, author-3) are also repeated to provide the names with German and English letters. This however suggests that there are more contributors than there actually are.
Naming	
Item Viola Syukrina E Janrosl, dan Yuliadi *Record* author: given: "Yuliadi" family: "Yuliadi"	*Issues (fields)*: Incorrectly input: repeated values (author, all) *Reasoning*: The second author's name in the item is given with only one name part "Yuliadi." In the record, this name appears in both the given and family name fields to suggest that their name is "Yuliadi Yuliadi." In Southeast Asian countries such as Indonesia, where this author is from, an individual's full name may have only one part. Given and family name fields are often set as "required," forcing these individuals to repeat their names or input filler text to advance in the interface.

Table 2. (Continued)

Example	Issues and reasoning
Status *Item* DR. IRAM MANZOOR Associate Professor Mr. F. S. Azeez Bukhari 4th Year MBBS *Record* author-1: given-name: "IRAM" family-name: "MANZOOR" author-2: given-name: "Azeez" family-name: "Bukhari"	*Issues (fields)*: Input in all caps (author-1, all) *Reasoning*: In the original item, the names of professors and associate professors are entered in all caps, while the names of students ("4th Year MBBS") are in regular case. This formatting distinction is replicated in the metadata record, although faculty and student titles are not included.
Geography *Record* publisher: "Elsevier BV" author-1: affiliation: [] author-n: affiliation: []	*Issues (fields)*: Location absent (publisher); Affiliation absent for all authors (affiliation, all) *Reasoning*: The publisher-location field is not used and the location of the publisher is not immediately apparent from the name of the publisher. In the case of multinational publishers, their location is less meaningful. Geographic context could be provided instead through author affiliations. However, those too are absent.

Although many of the identified issues may in fact result from poor metadata practice, we should not pre-emptively assume that all issues are ones of general quality. It is just as possible that they result from deliberate acts to assert or retain cultural meanings and aspects of identity because the current systems and standards are failing to reflect user realities and require significant resources to generate rich and inclusive metadata.

As well, deviations from metadata standards and best practices can affect how cultural meanings and identities are represented. Certain issues may have more substantive impacts than others. In all cases, though, there is potential for confusion and, taken together, increased wariness of how well metadata can convey meanings and identities.

At local and regional levels, more conversations are needed with scholarly, publishing, technical, and metadata communities to

understand and address the ways that metadata renders identities and meanings invisible in relation to the five categories identified in this review. If metadata contributes to stories that individuals and communities seek to tell about themselves, we must ask what stories our current systems and standards are neglecting and how this neglect can be addressed.

This review was not intended to surface all issues in the sample. Those found are specific to a single interpretation of these 427 records. The project also did not reveal every possible issue of cultural quality that could exist in metadata. The ability to recognize issues often depends on one's familiarity with particular social, regional, and research cultures.

Future Developments

The next phase of our project focuses on detecting the identified cases in a sample of 100,000 records to provide a better sense of their prevalence in the scholarly record. With over 5,600 publishers and thirty-seven languages in this sample, this work also looks at intersections between metadata quality, language use, and publisher size to explore who is most affected by issues related to language.

Preliminary analyses suggest that the issues discussed above appear more frequently in multilingual and non-English monolingual records. While instances are found in records created by publishers of all sizes, these records are most often created by the smallest publishers. The smallest publishers working with multilingual and non-English monolingual publications thus bear the highest burden for issues related to language, which often relate to the four other types.

Conclusion

The work of libraries, publishing houses, and technical organizations revolves heavily around the management of information. We nurture

information resources in their creation, development, distribution, acquisition, discovery, access, use, and preservation in the scholarly record for the long term. With so many touch points in the resource life cycle, it is important to ask how our metadata systems and standards contribute to definitions of who or what can be in the scholarly record, and who or what can subsequently be discovered and how.

In the larger scholarly ecosystem, we also hope these findings can support and contribute to broader conversations and struggles around equitable participation and homogenizing standards. By critically reflecting on our current systems and standards in this way, we can tease out such biases and start to unpack the tensions that they can create. From there, we can weave principles of equity into our best practices for metadata creation, journal publishing, and indexing, as well as our development and evaluation processes for technical systems and standards.

Contributor Notes

Julie Shi is Digital Preservation Librarian, Scholars Portal, University of Toronto Libraries and Ontario Council of University Libraries, Toronto, Ontario.

Dennis Donathan II is Research Associate, Public Knowledge Project.

Notes

1 Rachel Jaffe, "Rethinking Metadata's Value and How It is Evaluated," *Technical Services Quarterly* 37, no. 4 (2020): 432–43, https://doi.org/10.1080/07317131.2020.1810443.

2 Sungwon Kim and Seongyun Cho, "Characteristics of Korean Personal Names," *Journal of the American Society for Information Science and Technology* 64, no. 1 (2013): 86–95, https://doi.org/10.1002/asi.22781.

3 Mahmoud Sayed A. Mahmoud and Maha M. Al-Sarraj, "Bilingual Qatar Digital Library: Benefits and Challenges," in *Maturity and Innovation in Digital Libraries*, eds. Milena Dobreva, Annika Hinze, and Maja Žumer

(Switzerland: Springer Cham, 2018), https://doi.org/10.1007/978-3-030-04257-8_19.

4 Krystyna K. Matusiak, Ling Meng, Ewa Barczyk, and Chia-Jung Shih, "Multilingual Metadata for Cultural Heritage Materials: The Case of the Tse-Tsung Chow Collection of Chinese Scrolls and Fan Paintings," *The Electronic Library* 33, no. 1 (February 2015): 136–51, https://doi.org/10.1108/EL-08-2013-0141.

5 Carol Rigby, "Nunavut Libraries Online Establish Inuit Language Bibliographic Cataloging Standards: Promoting Indigenous Language Using a Commercial ILS," *Cataloging & Classification Quarterly* 53, no. 5–6 (July 2015): 615–39, https://doi.org/10.1080/01639374.2015.1008165.

6 Naomi Shiraishi, Charlene Chou, Liangyu Fu, and Xiuying Zou, "CEAL Task Force for Review of the ERMB Interim Report," *Journal of East Asian Libraries* 2021, no. 173 (2021): 4, https://scholarsarchive.byu.edu/jeal/vol2021/iss173/4.

7 Lana Soglasnova, "Dealing with False Friends to Avoid Errors in Subject Analysis in Slavic Cataloging: An Overview of Resources and Strategies," *Cataloging & Classification Quarterly* 56, no. 5–6 (April 2018): 404–21, https://doi.org/10.1080/01639374.2018.1438551.

8 Clarivate, "Web of Science Journal Evaluation Process and Selection Criteria," *Clarivate*, accessed December 4, 2022, https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/web-of-science/core-collection/editorial-selection-process/editorial-selection-process/.

9 Coalition Publica Metadata Working Group, *Technical Report: Metadata Feedback for Coalition Publica* (Canada: Erudit, 2021), accessed November 24, 2022, https://www.erudit.org/public/documents/CP_Technical_Report.pdf.

10 Dartmouth Library Metadata Services, *Troubleshooting Guide for Diacritics*, accessed November 27, 2022, https://www.dartmouth.edu/library/catmet/cataloging/diacritics-troubleshooting.html.

11 W3C Internationalization Working Group, *Strings on the Web: Language and Direction Metadata* (2022), accessed December 6, 2022, https://www.w3.org/TR/string-meta/.

12 Thomas R. Bruce and Diane I. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*, eds. Diane I. Hillmann and Elaine Westbrooks (Chicago: American Library Association, 2004).

13 Mary S. Woodley, "Metadata Matters: Connecting People and Information," In *Introduction to Metadata*, 3rd ed., ed. Murtha Baca (Los Angeles: Getty Publications, 2016), http://www.getty.edu/publications/intrometadata/metadata-matters/.

14 Chuttur M. Yasser, "An Analysis of Problems in Metadata Records," *Journal of Library Metadata* 11 (2011): 51–62, https://doi.org/10.1080/19386389.2011.570654.

15 Naomi Shiraishi, "Accuracy of Identity Information and Name Authority Records," in *Ethical Questions in Name Authority Control*, ed. Jane Sandberg (Sacramento: Library Juice Press, 2019): 181–94.

16 Marcia Lei Zeng and Jian Qin, "Metadata Quality: Measurement and Improvement," in *Metadata*, 2nd ed. (Chicago: American Library Association, 2016): 317–46.