Identifying Metadata Quality Issues Across Cultures

Julie Shi, Mike Nason, Marco Tullney, and Juan Pablo Alperin

Metadata are crucial for discovery and access by providing 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 quality issues, political acts to assert identity, or strategic choices to maximize visibility. In this context, we sought to understand how metadata quality, consistency, and completeness impact individuals and communities. Reviewing a sample of records, we identified and classified issues stemming from how metadata and communities press up against each other to intentionally reflect (or not) cultural meanings.

Introduction

Metadata are crucial to the dissemination and communication of research. As descriptors of "potentially informative object[s]" (Pomerantz, 2015, p. 26), metadata provide contextual, technical, and administrative information that facilitate the discovery, retrieval, and preservation of scholarly outputs. When created and maintained according to shared standards, metadata allow connections and relationships to be established between research and researchers, as well as across geographic, temporal, and discursive spaces (Gartner, 2016). These shared standards also enable metadata sharing through automated ingest and harvesting between platforms and services (Zeng & Qin, 2016), increasing the reach and, arguably, the use and impact of research.

Metadata are also technical, and "technological constraints should never be an excuse to diminish someone's personhood, or inaccurately reflect their identity" (Coalition Publica Metadata Working Group, 2021, p. 22). Subjective in nature, metadata elements constitute sites of tension and struggle between resource constraints, sociocultural representations, and standardized systems. Formal and informal interventions in these contested spaces may then be dismissed as metadata quality issues or be recognized as political acts to assert aspects of cultural identity or strategic curatorial choices to maximize opportunities for discoverability and visibility in research platforms and services.

---

* *Julie Shi is Digital Preservation Librarian at Scholars Portal, email: juli.shi@utoronto.ca; Mike Nason is Crossref/ Metadata Publishing Specialist with Public Knowledge Project & the Open Scholarship & Publishing Librarian at University of New Brunswick, email: mnason@unb.ca; Marco Tullney is Head of Publishing Services at Technische Informationsbibliothek, email: marco.tullney@tib.eu; and Juan Pablo Alperin is Co-Scientific Director with the Public Knowledge Project & Associate Professor at Simon Fraser University, email: juan@alperin.ca.*

These tensions are simultaneously made invisible and problematic by the broader knowledge landscape in which metadata standards and values operate: a landscape that is overwhelmingly structured around the English language and Western publishing practices, despite the decidedly global and multilingual nature of scholarship (Khanna et al., 2022; Library Publishing Coalition, 2018). In such an environment, norms that are defined according to the needs and concerns of these twin hegemonies become systemic constraints for those not represented by them. Whether in metadata or other aspects of this landscape, deviations from normalized practice are at risk of being dismissed as issues of proficiency and quality.

In this context, and as members of organizations that create systems for managing scholarly metadata and as research users of this data, we were interested in understanding how metadata quality, consistency, and completeness impact individuals and communities. Specifically, we sought to identify the ways in which identities are erased or obscured in metadata.

Treating metadata records as informational objects in their own right, we take the position that metadata may be accurate and of high quality "only if it does not forcibly out or harm the person in the record" (Shiraishi, 2019, p. 192). We recognize the limitations of such a definition, as risks of harm vary by context. Working from a sample of records known to have erroneous, incomplete, or otherwise technically imperfect metadata, this project therefore set out to identify and classify the metadata quality issues stemming from how metadata and communities press up against each other to intentionally reflect (or not) cultural meanings.

Alongside this definition of quality, we define cultural issues as those issues that impact, or have the potential to impact, the representation of identities, roles, intentions, and other factors specific to social, regional, disciplinary, or publishing cultures. This scope attempts to distinguish between issues that relate to identity expressions and those introduced due to aesthetic choices or disciplinary practices, to focus on the ways in which individuals and communities actively seek to convey meaning. Issues found in such standardized fields as ISSNs and page numbers are considered safely out of scope.

Beginning with a review of the literature on metadata quality and a description of our methodology, this article goes on to provide an overview of the various metadata quality issues we identified and the categories we developed to better understand them. We conclude by discussing the implications of our findings and describing future work we intend to undertake.

Literature Review

Undertaking a study of metadata quality begins with understanding that "metadata quality is a multidimensional concept" which requires defining "what we mean by 'good' or 'bad' quality" (Zeng & Qin 2016, pp. 319 & 322). The possible range of metadata issues that can be identified will depend on how quality is defined. In the library community, the consensus is that quality metadata work accounts for user expectations to facilitate resource discovery and use (Bruce & Hillmann, 2004; Cataloging Ethics Steering Committee, 2020; PIE-J Working Group, 2013; Pomerantz, 2015).

Mapping the key user tasks defined in the IFLA Functional Requirements for Bibliographic Records (FRBR) model—finding, identifying, selecting, and obtaining information—to characteristics of metadata, Bruce & Hillmann (2004) determined six dimensions along which metadata quality could be defined. In addition to the completeness and accuracy of information in the record, they note that records should include elements and controlled vocabularies that "the community would reasonably expect to find" and that are "consistent with standard

definitions and concepts used in the subject or related domains" (p. 245). This metadata should also be provided alongside resources in a timely and accessible manner.

Bruce & Hillmann (2004) measure metadata quality according to its "fitness for use" (Zaveri et al., 2012, p. 2) for fulfilling user tasks. Addressing usability more concretely, Yasser (2011) reports incorrect values, incorrect elements, missing information, information loss, and inconsistent value representation as the most common metadata issues degrading the "utility of metadata records" (p. 60). A 2013 NISO working report provided recommendations for presenting and identifying e-journals. Common metadata issues identified include missing information about title changes and publisher history, incorrect citations and URLs, and inconsistent publication information.

This focus on utility extends beyond human users to machines as well. Studies exploring issues in quality have largely addressed the impacts of poor metadata on data aggregation, resource discovery and access, and interface functionality (Bruce & Hillmann, 2004; Malički & Alperin, 2020; Woodley, 2016; Yasser, 2011; Zaveri et al., 2012). These studies work toward goals for metadata sharing and interoperability, for which tools and processes for automated data exchange also introduce tensions, errors, and erasures in metadata (Heery & Patel, 2000; Jaffe, 2020; Zeng, 2018).

The literature tends to overlook the ways in which metadata "contribute to a story we are telling about ourselves as individuals, as organizations, and as a community" (Jaffe, 2020, p. 441). This is despite a general recognition of the "subjective nature of metadata practice" (p. 2), which is inflected by culture and context, biases and structural problems embedded in metadata systems and tools, and the power dynamics and politics of naming and description (Farnel, 2018). Király et al. (2019) propose metrics for evaluating the multilingual dimensions of metadata in the Europeana digital cultural heritage platform, however, the framework is limited to technical and functional aspects of metadata.

Most studies that do address sociocultural themes largely attend to cataloging standards, schemas, and vocabularies, including issues around the representation of non-English languages and non-Roman scripts, non-White and/or non-Western contexts, Indigenous knowledges and worldviews, and gender and sexuality, among other issues (Adler, 2017; Berman, 1971; Billey et al., 2014; Billings et al., 2017; Duarte & Belarde-Lewis, 2015; Ducheva & Pennington, 2019; Farnel et al., 2017; Mahmoud & Al-Sarraj, 2018; Matusiak et al., 2015; Olson, 2002; Rigby, 2015).

Far fewer studies engage with the sociocultural dimensions and consequences of metadata quality issues introduced during the publishing process. In 2021, the Equity and Metadata subgroup of the Coalition Publica Metadata Working Group in Canada reported on barriers to equitable and inclusive publication metadata, raising a critical question: "So perhaps we need to consider not just the practices around metadata but with whom lies the 'power to name' or ascribe metadata. Perhaps accountability in metadata needs to be considered as well?" (p. 15).

Multilingualism and Metadata

Language choices open or foreclose on opportunities to represent cultural meaning and identity across scholarly communications spaces. The role of English as *lingua franca* in academic and research spaces has been discussed and debated for decades (Canagarajah, 2002; Crystal, 2012; Turner, 2018). For instance, a shared language can foster communication and collaboration (Alhasnawi, 2021). Yet, scholars from a range of backgrounds point to the psychological,

economic, social, and other burdens that English-language preferences and requirements place on those who do not know English as a first language, or at all (Tomuschat, 2017; Alamri, 2021; Balula & Leão, 2021; Pho & Tran, 2016; Ge, 2015; Santos and Da Silva, 2016; Curry & Lillis, 2010). The language used to create metadata is then a political choice (Rigby, 2015).

From a usability standpoint, accurate multilingual metadata provides critical access to important resources for legal, cultural, and political purposes and also promotes understanding of regional cultures and histories (Mahmoud & Al-Sarraj, 2018; Matusiak et al., 2015). Zeng & Qin (2016) note that authors often provide "multiple local versions" (p. 142) of metadata values for titles, authors, keywords, and glossaries through inline and external parallel metadata. These localized versions refer to translations and references to multilingual glosses that allow authors to capture metadata values in both English and the original language of the materials being described.

Creating consistent multilingual metadata, whether automatically or manually, is a resource intensive process. It requires significant technical development and maintenance and human resources to establish, implement, and maintain (Matusiak et al., 2015; Soglasnova, 2018). They also require systems to be encoded and designed appropriately for communities and researchers to benefit from multilingual metadata and access critical information (Mahmoud & Al-Sarraj, 2018; Rigby, 2015; Shiraishi et al., 2021). This is especially true for languages that are not rooted in the Roman alphabet and have a directionality other than left to right.

In all cases, the appearance and functionality of multilingual metadata in user interfaces is contingent on the quality of language metadata and interface design. Missing or improper language codes and interface designs that fail to account for linguistic differences can prevent metadata in certain languages from being input and render content unintelligible and features unusable (W3C, 2022). Font properties and encoding issues may also prevent the display of characters with diacritics and ligatures used in Roman scripts and Romanizations (e.g., Dartmouth Library Metadata Services, n.d.).

The lack of standardized and widely adopted Romanization schemes for many languages itself results in errors and inconsistencies: localized standards may be developed and used in isolation; when multiple schemes exist like this, guidance may be referenced and applied inconsistently (Park, 2007); or Romanized forms may be decided on independent of any guidance. Moreover, the choice to record Romanizations only may preclude access to resources by users unfamiliar with such schemes or who would transcribe or transliterate differently (Rigby, 2015). This raises further ethical questions about who metadata caters to when rendered only in translation, transcription, or transliteration.

Names and Metadata

Assessing the quality of name forms and expression in MARC library records, Wisser (2014) identified common errors in encoding, typography, content, and format. Issues included variations in the ways that dates, geographic qualifiers, name parts, and abbreviations and initials are included (or not) and represented. Improper encodings and recordings that misrepresented the nature of the value (e.g., a corporate name encoded as a personal one), as well as misspellings and punctuation errors, were also noted.

Yet, the quality of name forms in metadata should not be measured solely by the well-formedness of these values for data exchange and bibliometric analysis. For members of the trans and gender non-binary community, for example, naming and surfacing previous/other

names may in fact produce harm. Best practices published by The Trans Metadata Collective (2022) include a section on recording former names, which opens with "Respect the wishes of the author regarding the use of their former name(s)" and goes on to recommend prioritizing the privacy and safety of the individual during metadata creation (p. 19). Several groups also recommend that journals respect retroactive name change requests in recognition of these harms (Coalition Publica Metadata Working Group, 2021; Committee on Publication Ethics, 2021).

As noted by the Coalition Publica Metadata Working Group (2021), individuals may also carry alternate or multiple names due to marriage and divorce, official government purposes, the use of stage names and/or pseudonyms, and myriad other reasons. While certain features of the ANSI/NISO Z39.96 JATS: Journal Article Tag Suite standard for journal publishing, including the alternative-name field and name-style attribute, allow for more robust name records, the Working Group notes that "'alternative name' is limited in scope… and 'name-style' is limited to Western, Eastern, Given-only, and Islensk (Icelandic) configurations" (p. 19).

Names and naming conventions are also deeply entwined in epistemic traditions and linguistic and cultural histories, and "writing personal names in forms other than [an author's] native languages is essentially a type of translation" (Kim & Cho, 2013, p. 88). As such, when a name is Romanized, nuances and differences in naming conventions can result in errors and information loss.

Methods

We constructed a purposeful sample of 427 records drawn from the Crossref API. Crossref is a non-profit organization that stores over 120 million metadata records from their over 15,000 members (primarily publishers). Our sample was not drawn randomly, since our goal was to learn about the types of metadata quality issues that exist. We hypothesized that records with at least one known issue, and additional randomly chosen records from the same publication by the same publisher would be more likely to yield cases where identity, language, and culture would appear as problematic records for our analysis.

As such, we used the expertise in our research team and from staff at Crossref to identify specific records and Crossref members whose data was known or suspected to have at least one metadata quality issue (e.g., titles in two languages included in a single field). The selected problematic records came from 51 DOI prefixes (typically corresponding to either a publication or a publisher) and were chosen without regard for the manuscript management or publishing platform used by the publisher. We then used the Crossref API to randomly select additional records from the same prefix. An additional three randomly chosen records were selected from 17 DOI prefixes from journals known to use the manuscript management and publishing platform Open Journal Systems (OJS). The choice to sample from OJS-based publishers stemmed from our own familiarity with the platform (with which several of the authors are affiliated), the documented international and multilingual reach (Khanna et al., 2022), and the previous work on its metadata quality, cited earlier (Nason et al., 2021). The seed list of publishers and the code used to extract related records is available online (Shi et al., 2023).

In the sample, 394 records (92%) correspond to research outputs by academic, industry, and government organizations, including journal articles, book chapters, book reviews, conference proceedings, and protocols. The remaining 33 records (8%) describe front and back matter (e.g.,

tables of contents, indexes), notices and communications, journals, journal issues and sections of journal issues, advertisements, and retractions (see Appendix A). As well, 140 records (33%) are associated with multilingual venues, including those that publish only titles, abstracts, and/or keywords in multiple languages and those that also publish full-text in multiple languages.

For each item, the JSON-formatted record (returned by the Crossref API) and the published document (at the URL pointed to by the DOI) were analyzed in tandem to enable us to consider issues present in the metadata as well as issues stemming from discrepancies between the published document and the record. Comparisons were also made with the item landing page and the container,* where further information, such as languages accepted for publication, were necessary. Issues were also investigated within and between records to determine isolated areas of concern and larger patterns. This approach is affirmed by Zeng & Qin (2016), who state that "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" (p. 322).

An initial analysis was completed on a subsample of 61 records to identify the metadata elements in which relevant issues were more likely to appear. After sorting records by DOI prefix, every seventh record in the dataset was selected for this scoping work to ensure an array of publishers were represented. When values were present, a close reading of the value was conducted alongside a comparison of the value with the corresponding information in the published document. The published document was also assessed to locate information absent from the metadata.

The potential political significance of cultural issues was noted and considered when issues could be read as deliberate interventions and/or for which specific motivations may be conjectured (correctly or not). Political significance may be specific to particular instances of an issue, all issues of a certain type, or may apply to a range of issue types.

From the initial analysis, the elements in Table 1 were found to be most pertinent to cultural identity and meaning. Metadata were categorized as either belonging to the work itself (i.e., item level, the contributors (i.e., person level), or the journal or other venue (i.e., container level). These categories provided support for considering the possible range of relevant issues.

Item-level metadata corresponds directly to the article page and PDF (when available) returned by the DOI. Person-level metadata describes the entities responsible for the creation of the item, which are typically individuals but can include groups or organizations. A "General" heading was also added to account for person-level issues that did not map directly to the three fields, such as the absence of some or all author names. Metadata at the container level relates to the nature, scope, and maintenance of the larger entity in which the item is found,

TABLE 1
Metadata fields of interest by item, person, and container

Item level	Person level	Container level
Abstract	Given Name	Publisher
Title	Family Name	Title
	Affiliation	Language
	General	Subject

* In this paper, "landing page" indicates the webpage or record for the item that is provided by the publisher or creator. "Container" reflects the language in the Crossref schema and refers to the publisher's platform for the larger work, such as a book or journal.

most often a journal or book in this sample. Issues in the "Subject" field were only noted for series and serials, as subject headings are not applied to books in the Crossref schema.

The "reference" element group for works cited in the published document were excluded from review to ensure a manageable dataset. A separate analysis could be conducted to specifically examine the presence of this element group, and of how issues of cited researchers and their works are represented in metadata records and the reference lists of published documents (Arastoopoor and Ahmadinasab, 2019, pp. 225-226).

It should be noted that this review is not intended to be exhaustive, and findings speak only to those records included in the sample. Cultural issues surfaced are limited to those noticeable to the reviewer and do not necessarily reflect accurately or fully the motivations of the individuals and organizations creating the metadata. Investigating the actual intentions of metadata creators is also out of scope of this work.

Results

This approach allowed us to identify 32 unique issues that took on five main forms (see Table 2). In total, we found 4,859 specific issues (an average of 11.4 issues per record). These issues were not all equally common, with eight comprising 75% (3,644) of the issues found. However, given the non-random sample used for this study, the number of each unique issue is less significant than the categories of issues found and their descriptions. As such, in the remainder of this section, the number of times an issue was identified is noted for transparency, but the focus is placed on the proposed organization and description of the issues themselves.

TABLE 2
List of 32 identified issues and their definitions, organized by their 5 main forms

Form	Issue	Sub-issue	Definition
Value absent			Value is absent from the record, including if the field itself is absent or the field is present but contains a "[]" or similar value. "Value absent" is both a form and a unique issue.
	translation absent		Translations are absent, when (1) items provide translations, (2) containers include multilingual content, or (3) publishers are based in areas where the language of the record is not a main or official language.
	value in original language absent		Value is not given in the original language or script and only a transliteration or English translation is provided.
	language attribute absent		Language of the value is not identified by an attribute, when (1) multiple languages appear in record, (2) journal publishes in multiple languages, (3) multiple language forms appear in a record (e.g., original and transliteration), (4) field is repeated in different languages, or (5) value is transliterated from a language other than the language of the record.
	language style absent	Romanization only	Value in original script is absent, when values that may be rendered in non-Roman scripts in their original language. Use is based on best guesses at times.

TABLE 2
List of 32 identified issues and their definitions, organized by their 5 main forms

Form	Issue	Sub-issue	Definition
	language style absent	Romanization absent	Name in original script only, when records use a mix of transliteration, translation and original script.
	VoR license terms absent		License terms for the version of record (VoR) is not in the record, but licenses for other purposes are (e.g., text and data mining).
	author/s absent		All authors of the item are absent from the record.
	not all authors listed		Some authors of the item are absent from the record.
	ORCIDs absent		ORCIDs included in the item are not included in the record.
	not all persons listed		Contributors other than the author/s are identified in the item but not in the record.
	absent for all authors		No affiliations are provided for any authors.
	absent for all editors		Affiliations are absent for all editors, when editors are listed in the record.
	not all publishers listed		Co-publishers listed on the item or container site are not represented in the record.
	related orgs absent		Organizations other than publishers (such as rightsholder, content manager, or other parties with responsibilities like content hosting) are listed on the item or container site but not in the record.
	location absent		Location of the publisher is absent from the record.
	subtitle absent		The subtitle of the container or item title is absent. Recorded only for the subsample due to common mis-recording of this value.
Value in record does not match with information in the item			Identified discrepancies between information in the record and information on the item itself, its landing page, or the container site.
	outdated		Only the previous title of the container is in the record.
	registered URL out of date		DOI does not resolve but the item can be found through other means (e.g., Google Scholar).
	registered URL invalid		DOI does not resolve and the item cannot be found easily through other channels.
	value in record does not match information on container website		Information in record is incongruent with information on container website.

TABLE 2
List of 32 identified issues and their definitions, organized by their 5 main forms

Form	Issue	Sub-issue	Definition
	inaccurate		Language and/or subject/s noted in the record either incorrectly or inadequately represent that of the item or container.
Value does not match with the parameters of the field			Format or contents of the value does not conform to metadata schema or best practices.
	affiliations presented as authors		Affiliations recorded in a separate author-name element group, instead of within the associated author-name element group.
	multiple languages in single field		A single field contains information in more than one language or language form.
	multiple values in single field		More than one value is presented in a single field.
	original-title used incorrectly	includes value in original language but item is not a translation	Item title in original language input in original-title field but item is not a published translation. Per the schema, original-title is reserved for the title in its original language when the item is a published translation.
	original-title used incorrectly	value repeated	Value input in the title field is repeated in the original-title field, which is reserved for the title in its original language when the item is a published translation.
	all authors listed as first		All authors listed as "first" in the sequence field.
	first author not identified		All authors listed as "additional" in the sequence field.
	input in all caps		A title or person name is input in all caps.
	additional persons listed		Persons other than the authors of the item are included in the record.
Lack of completeness of the value			Issues within the contents of the value.
	value incomplete		Words or characters are missing from the value or are rendered improperly in the value, such as omitting characters with diacritics either by dropping the character entirely or entering its equivalent in the English alphabet.
	only provides initial/s		Only the first letter of the name is provided. Initials may be represented as X.Y. or X. Y. or XY or X Y or X-Y or X.-Y., etc.

TABLE 2
List of 32 identified issues and their definitions, organized by their 5 main forms

Form	Issue	Sub-issue	Definition
	acronym only		Value is entered as an acronym only. The acronym may be based on an organization name in the original language or in translation.
Incorrectly input	Several types of errors (see Figure 1 for examples)		Indicates that (1) information that does not belong in the field is present, or (2) a value is present but information is missing. Issues may be cultural or general.

Categories of Issues

In addition to wanting to identify unique metadata quality issues and their forms, our project sought to determine which issues pertained to cultural meaning and identity and which related to general quality. In some instances, however, the same type of metadata issue could fall under either category, or even both simultaneously. Still, we felt it useful to group issues into categories that could be used when discussing the cultural context from which issues arise. In making such categorizations, we acknowledge that distinctions are often difficult to discern without familiarity with specific regional, disciplinary, and publishing cultures from where the metadata emerged. As such, the following categories are only one interpretation of the possible themes and areas of tension that could be helpful in identifying metadata issues that pertain to cultural identity.

Through the analysis and description of the 32 unique issues, we were able to identify five common categories that would often reflect individual identities or other cultural characteristics: 1) language, 2) contributors, 3) names, 4) status, and 5) geography. These are described in more detail with examples of key issues in Table 3. Due to the complexity of identified issues, certain issues correspond to multiple categories depending on their nature and context. Appendix B provides a full mapping, with examples, of the 32 issues to the categories.

Within each category, we further identified key issues that, in our assessment, deserved special attention based on two factors: 1) the potential impacts of issues that may be deliberately introduced to assert cultural meanings or identity or to strategically present outputs for internationalization and increased visibility, and 2) the feasibility of automating an alert or solution to identify or resolve issues.

TABLE 3
Defined categories with key issues

Category	Definition	Specific Key Issues
Language	Issues are in relation to the languages and scripts of values and/or the way in which they are identified using language and style attributes.	• Translation absent • Value in original language absent • Language attribute absent • Multiple languages in single field • Language style absent • Inaccurate (for Language and Subject only)

TABLE 3
Defined categories with key issues

Category	Definition	Specific Key Issues
Contribution	Issues relate to the acknowledgment of contributors to the creation and publication of the item and its contents including, but not limited to, co-authors, funders, and co-publishers.	• Author/s absent (if all authors are absent) • Not all authors listed (if some authors are absent)
Naming	Issues relate to the recording of individual and organizational names in accordance with linguistic and cultural conventions. For Individuals, these can relate to full names and name parts, naming conventions, scripts, or Romanizations. For their affiliations or publishers associated with the work, these might relate to the use of acronyms and abbreviations.	• Incorrectly input (for Given and Family Names, Affiliation, and Publisher only) • Only provides initial/s (for Given and Family Names only) • Acronym only (for Affiliation and Publisher only)
Status	Issues relate to stylistic and content-based interventions to capture the status, seniority, or prestige of individuals or institutions.	• Use of honorifics in name fields • All authors listed as first • First author not identified • Input in all caps • Absent for all authors (for Affiliation only) • Affiliations presented as authors
Geography	Issues are caused by the absence or partial representation of physical location and its social and cultural associations.	• Location absent (for Publisher only) • Absent for all authors (for Affiliation only)

Examples of Issues

Using the categories above, we identified 4,387 (90%) of the 4,859 issues in our sample that could be linked to culture or identity. This corresponded to an average of 10.3 cultural issues per record with the potential impact of metadata quality, consistency, and completeness on individuals and communities across cultures is significant.

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
Language	
DOI 10.32598/jmsp.6.4.686 *Item* Item title, abstract, author names and affiliations, and journal title are provided in Persian and English. The full text is in Persian only.	Value in original language absent (all): According to this journal's policies, the full text of an article is published in Farsi/Persian only. Abstracts are published in Farsi and English, and bibliographies are published in English only. Given that Farsi is the primary language of this journal, the absence of Farsi in the record is significant.

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
Record Item-title: "The Impact of Institutional Quality and Exchange Market Pressure on Foreign Direct Investment : A Cross Countries Study" Author-1: Given-name: "Bahareh" Family-name: "Mofavezi" Author-2: Given-name: "Zohreh" Family-name: "Tabataba'i-Nasab" Author-3: Given-name: "Seyed Yahya" Family-name: "Abtahi" Container-title: "Quarterly Journal of The Macro and Strategic Policies"	
DOI 10.15750/chss..54.201411.007 *Item* Item title, abstract, and author information as well as container title and publisher are available in Korean and English. The full text is in Korean only. *Record* Author-1: family-name: "김성수" … Container-title: "CHUL HAK SA SANG - Journal of Philosophical Ideas" … Language: "en"	Assuming the language "en" is used to indicate the language of the record: Multiple languages in single field (Container title): In a single field, the container title is presented in Romanized Korean and English translation, where Romanization and translation are considered distinct language forms.* Language attribute absent (author-1 family name): The language of the record is set as English and a Romanization of the author's name is provided in the original item, however the record includes the author's name in Korean script only. Input in all caps (container-title): The Romanized journal title is set in all caps while the translated English title is set in regular case. It is assumed that this is related to the common Romanization practice of using all caps for the family name in Romanized Chinese, Japanese, and Korean names in all caps to distinguish name parts.

* From the scope notes and examples in the JATS Tag Library for the attribute @xml:lang, it is unclear what language should be assigned to a value when Latin scripts are used to record non-Latin languages (e.g., transliteration, Romanization, etc.): on the one hand, "Language-Script-Region: xml:lang="sr-Latn-RS" (Serbian written using the Latin script as used in Serbia)," but on the other hand, "Romanized Japanese name referred to as an "English" name" (NCBI & NLM 2021).

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
DOI 10.12681/jode.9694 *Record* Publisher: "National Documentation Centre (EKT)"	Multiple languages in single field (Publisher): The publisher's name is recorded in English translation. This is followed by an acronym in parentheses that is based on the publisher's name in Greek—Εθνικό Κέντρο Τεκμηρίωσης. Such use of multiple languages in one field may lead to confusion downstream.
DOI 10.1055/s-0038-1628298 *Item* Item title is included in original German only, however the item abstract is provided in the original German and translated English. *Item landing page* Item title and abstract are given in both original German and English translation. *Record* abstract: "<jats:title>Zusammenfassung </ jats:title><jats:p>Die Therapie der…" item-title: "Das Problem der Osteitis bei der Periprothetischen Gelenkinfektion"	Value in record does not match information on container website (all): An English translation of the item title that is provided on the item landing page is not given in the item itself or the record. Translation absent (all): English translations on the item landing page are not present in the record.
Contribution	
DOI 10.2307/3595240 *Item* *Zarte Liebe fesselt mich. Das Liederbuch der Fürstin Sophie Erdmuthe von Nassau-Saarbrücken. Teiledition mit Nachdichtungen von Ludwig Harig*. Hg. von Wendelin Müller-Blattau. Saarbrücken: Institut für Landeskunde im Saarland, 2001 (Veröffentlichungen des Instituts für Landeskunde im Saarland 39). 111 S., mus. Not., Abb., Tab., Reg.; Faks.-Beil.: 34 S., mus. Not., ISBN 3-923877. … Ulla Enfilin, Berlin	Additional persons listed (author-2, author-3): This item is a book review. Authors of the work reviewed are listed in the record alongside the reviewer (author-1). Incorrectly input: repeated values (author-4, author-5): Two author names (author-1, author-3) are repeated, which suggests that there are more contributors related to this work than there actually are.

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
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"	
DOI 10.12681/jode.9694 *Container* A note on the journal issue cover also states: "A periodical electronic publication of the Scientific Association: Hellenic Network of Open and Distance Education" *Record* Publisher: "National Documentation Centre (EKT)"	Value in record does not match information on container website (publisher): The journal website and journal issue cover reference the Hellenic Network of Open and Distance Education. Neither the translated English name nor the original Greek acronym in the publisher field refer to this network.
Naming	
DOI 10.15750/chss..54.201411.007 *Item* Author name is included in the original Korean as well as in Romanized Korean as "Kim, Sungsu." Author affiliation is provided in the original Korean only and includes their title alongside their departmental 철학과 (Philosophy) and university 서울시립대학교 (University of Seoul) affiliations.	Incorrectly input: with given name (family-name): Both family and given names for the author are recorded in the family-name field. As Kim & Cho (2012) note, "the three syllables of a Korean name can be written as all attached or spaced"; names written as attached may result in this kind of issue.

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
Item landing page Author name is provided in the original Korean as well as in Romanized Korean as "Sungsu Kim," depending on the selected language for the interface. The author's affiliation is only provided in the original Korean script at the university level. Record author: family-name: "김성수" affiliation: [] Language: "en"	Language attribute absent (family-name): Where the language of the record is stated as English, a language attribute should be used to signal that the author's name is written in Korean script. It is interesting that two different Romanizations appear in the item and item landing page, but neither are used in the record. Affiliation absent for all authors (affiliation): Neither the departmental nor the university affiliation is included in the record, although they are provided in the item and item landing page. An evaluation of how well a value aligns with linguistic and cultural naming practices requires the presence of a value in the record.
DOI 10.2307/4147866 *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"	Language attribute absent (author, all): This record references one reviewer and two authors of the reviewed book, however five author names are recorded. Two author names in the original German contain characters not present in the English alphabet ("ß" in author-1 and "ü" in author-3), resulting in the repetition of these names in Romanized form using the English alphabet only ("ss" in author-4 and "u" in author-5, respectively). Language attributes are not included to note these linguistic distinctions. This stands in contrast to the "multiple values in single field" issue that is more commonly seen in container and item title fields but appears to stem from the same goal of representing information in multiple languages.
DOI 10.35143/jakb.v12i1.2485 *Item* Viola Syukrina E Janrosl, dan Yuliadi	Incorrectly input: repeated values (author, all): The second author's name in the item is given with only one name part "Yuliadi." In the record, however, this name appears in both the given and family name fields to suggest that their name is "Yuliadi Yuliadi."

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
Record author: given: "Yuliadi" family: "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.
DOI 10.12681/jode.9694 *Record* Publisher: "National Documentation Centre (EKT)"	Value in original language absent (publisher): The publisher's full name in the original Greek is absent from the record. This absence stands out especially in this record as the item abstract and title and container title are all given in Greek only.
Status	
DOI 10.28933/ajcsa-2017-05-1801 *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"	Input in all caps (author-1, all): 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.
DOI 10.28933/ajcsa-2017-05-1801 *Item* Zubair Ahmad Research Scholar: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan Zawar Hussain Assistant Professor: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan	Not all authors listed (author-1, name and affiliation): The name of the first author is not included in the record, although their title as "Research Scholar" alongside their affiliation is included. Affiliations presented as authors (author-1, author-3): Instead of using the affiliation field for each author, affiliations, as well as titles, are recorded as independent authors of the item (author-1 and author-3).

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
Record author-1: name: "Research Scholar: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan" sequence: "first" affiliation: [] author-2: given: "Zawar" family: "Hussain" sequence: "additional" affiliation: [] author-3: name: "Assistant Professor: Department of Statistics, Quaid-i-Azam University 45320, Islamabad 44000, Pakistan"	
Geography	
DOI 10.15750/chss..54.201411.007 *Item landing page* Publisher is identified, in both Korean and English, as 서울대학교 철학사상연구소 the Institute for Philosophy at Seoul National University. The author's affiliation is noted in Korean only as 서울시립대학교 (University of Seoul). *Record* Publisher: "Institute for Philosophy" Author-1: Affiliation: [] Language: "en"	Value incomplete (publisher): Per the item landing page, the publisher for this journal is a unit within a larger organization. In the absence of this larger organization's name in the record, however, "Institute for Philosophy" carries little contextual information about the publisher and its location, geographic and otherwise. Publisher location absent (publisher-location): Where the publisher-location field could have remedied the incomplete publisher name, whether by mention of Seoul or Korea, the absence of this field further prevents understanding of how and where to locate this publication. Value in original language absent (publisher): The original name of the publisher in Korean is not included in the record. While the inclusion of only the English translation may be because English is stated as the language of the record, this reasoning is weakened by the use of the author's Korean name instead of one of the two Romanizations used in the item and item landing page.

TABLE 4
Examples of issues by category

Example	Issue details Issue (field in example): reasoning
	Affiliation absent for all authors (affiliation): In the same vein as "Publisher location absent" above, the absence of the author's affiliation (and therefore, in this case, their geographic location) also limits understanding of the author's context. In this case, it is possible that the affiliation is not recorded because no English translation is available; only the original Korean is noted in the item or item landing page.
DOI 10.12681/jode.9694 *Record* Publisher: "National Documentation Centre (EKT)"	Location absent (publisher): The publisher-location field is not used and the location of the publisher is not immediately apparent from the value recorded for the publisher. Both the full name and acronym are official names used by the organization, however the absence of the full name in the original Greek may prevent educated guesses about the publisher's location based on language.

As noted earlier, some issues were more prominent than others, with eight issues classified as cultural appearing over 200 times within our non-random sample: 1) value absent, 2) language attribute absent, 3) publisher location absent, 4) affiliation absent for all authors, 5) language style absent: Romanization only, 6) incorrectly input, 7) value in original language absent, and 8) translation absent. Appendix C contains the full list of issues and the number of occurrences of each, by metadata level and field, in our sample.

Of these eight most common issues, all but one ("incorrectly input") refer to the absence of certain values or attributes from the record, with four correlating to language representation and two related to geographic and institutional location. Depending on the granularity of detail for affiliations, this field may also reflect disciplinary (and to a lesser extent, theoretical) locations.

Over half (n = 728, 54%) of the issues classed as "value absent" relate to rights and licensing information. Another 43% of absent values are in the abstract, language, and subject fields; the absence of a value in the language field is especially significant when multiple languages are present in the item and/or record or when the language of the record is different to that of the item.

Relatedly, when the language of individual values is different from the stated language of the record, a language attribute can be appended to the element. However, "language attribute absent" issues were frequently found in the container title, item title, and given and family name fields. In some of these cases, most notably in the name fields, only Romanizations or translations are provided. This raises further questions about the politics of naming and language, where researchers may choose Romanizations or other names for personal or professional reasons, or may not have a name in a non-Roman script.

In contrast, the "value in original language absent" issue corresponded most often with the publisher and affiliation fields, while "translation absent" occurred frequently with container

and item titles and abstracts; titles and abstracts in both original and translated languages were not included in any of the 140 records from multilingual venues included in the sample. These issues appear equally for container-level subjects; journals that recorded subject headings only provided headings in English, regardless of publication and record language/s. It is unclear if journals are able to apply non-English subject headings. The presence and accuracy of subject headings in records may also vary by publisher size, with smaller or independent journals less likely to assign relevant headings.

Other issues were not always so clearly of cultural significance. The "Incorrectly input" issue, for example, is an umbrella form that covers a variety of issues. Table 5 illustrates some of the issues under this umbrella and how they are designated as being cultural or non-cultural. Where deliberate motivations, such as using sentence case or all capitals to reflect seniority, are suspected, issues are recorded as cultural issues; this issue is noted as "input in all caps" for the item title field. In other cases where capitalization in the record may result from copy-pasting values from the published document, for instance, such issues are noted as "Other" (i.e., non-cultural). The authors recognize that such decisions are subjective.

TABLE 5
Examples of the range of issues of the form "incorrectly input"

Example	**Issue details** **Issue (field in example)**
Cultural issues	
DOI 10.17504/protocols.io.taheib6 *Record* author-1: given: "**Assoc**." family: "**Prof**. Vichien Srimuninnimit" author-2: given: "**Dr**." family: "Areewan Somwangprasert"	Incorrectly input: with titles only (given name) and Incorrectly input: with titles (family name) Definitions: • with titles only: person's title recorded in given name field without given name. • with titles: person's title is recorded in field with given name. Reasoning: recording titles in name fields may suggest the importance of seniority and rank. Suggested citations on the landing page that include these titles reflect downstream consequences.
DOI 10.7705/biomedica.v28i2.101 *Record* publisher: "Instituto Nacional de Salud (**Colombia**)"	"Incorrectly input: with location in parentheses" (publisher) Definition: value includes location, which is not part of the official name/title. Reasoning: including the publisher's location suggests the importance of place to organizational identity. Location is even more significant for organizations with less unique names such as this one. In many cases (as in this one), the publisher-location field is not used.

TABLE 5
Examples of the range of issues of the form "incorrectly input"

Example	Issue details Issue (field in example)
DOI 10.14710/jadu.v2i2.7641 *Record* publisher: "Institute of Research and Community Services Diponegoro University (**LPPM UNDIP**)"	Incorrectly input: with acronym of original lang value (publisher) Definition: value includes an acronym of the organization or container name in the original language. The acronym is not part of the official name or title and it often appears alongside an English translation of the name or title. Reasoning: an acronym of the original name is read as resisting linguistic erasure, providing a familiar access point to the organization's local community, or maintaining a consistent identity across languages over time.
Non-cultural issues	
DOI 10.1080/10587259408027158 *Record* affiliation-1: name: "**a** Department of Chemistry , Humboldt-University […]" affiliation-2: name: "**b** L. Dähne Institute of Organic Chemistry, […]"	Incorrectly input: with footnote marker (affiliation) Definition: numbers or punctuation marks (e.g., asterisk) for footnotes included incorrectly in field, with or without text of footnote. Reasoning: footnote marker likely included by accident due to copy-paste style of data entry.
DOI 10.15530/urtec-2017-2670073 *Record* author-1: given: "**null**" family: "**null**"	Incorrectly input: as null (given and family name) Definition: value entered as "null" and without actual value. Similar issues with "none," "not provided," and punctuation marks like "—" and ".." Reasoning: where "null" appears in multiple fields in the record, the issue is likely to be the result of an issue related to automated metadata creation or because the item does not have a dedicated author (e.g., editorials, full volumes, etc.).
DOI 10.1055/b-0037-147455 *Record* title: "6.4 Vorgehen bei äußeren Laryngozelen"	Incorrectly input: with chapter and section numbering (title) Definition: chapter and/or section number included with title; however they are not part of the title itself. Reasoning: chapter and section numbering possibly included by accident due to copy-paste style of data entry or a lack of other appropriate elements in the user interface.

TABLE 5
Examples of the range of issues of the form "incorrectly input"

Example	Issue details Issue (field in example)
Issues that are not clearly cultural or non-cultural	
DOI-1 10.24114/konseling.v19i2.30476 *Record (1a)* Title: "Citra Diri Penyandang Tunanetra terhadap Diskriminasi dari Lingkungan Sosial" *Item (1b)* CITRA DIRI PENYANDANG TUNANETRA TERHADAP DISKRIMINASI DARI LINGKUNGAN SOSIAL Widya Lestari1 Riski Fitlya2 Program Studi Psikologi, Universitas Muhammadiyah Pontianak1,2 *DOI-2* 10.24114/konseling.v19i2.30439 *Record (2a)* Title: "META ANALISIS GRATITUDE INTERVENTION PADA WELL-BEING" *Item (2b)* META ANALISIS GRATITUDE INTERVENTION PADA WELL-BEING *Levina Wicaksono* *Universitas Surabaya, Fakultas Psikologi, Magister Psikologi Profesi*	Incorrectly input: input in all caps (title-2) Definition: item title for the second article is input in all caps. Reasoning: In the first article, the authors appear to be non-faculty members and the item title is recorded in regular sentence case in the record. By contrast, in the second article, the author is a faculty member and the item title is recorded in all caps in the record. It is possible that capitalization choices are based on the seniority of the author, however it is just as possible that this stems from inconsistent practice. Further analyses of other records from this journal would be needed to determine if a pattern emerges and the issue leans more toward cultural or non-cultural.

Discussion

While many of the identified issues may, in fact, be due to poor metadata practice, it is apparent from the findings that the potential cultural motivations behind their presence in the metadata cannot be ignored. Measured against the possibility of harm to the individuals and communities most affected by a resource, there is clearly a need to consider metadata while engaging in broader conversations about the effects of homogenizing standards and equitable participation in research. The consequences of providing bibliographic information in English only for an article that is published wholly in another language, as is the case in some instances in our sample (e.g., Table 4, example 1 under "Language"), are not trivial and cut across these broader conversations.

Intentional or not, deviations from standards and so-called "best practices" for metadata entry affect the representations of cultural meanings and identities in substantive ways and should not be preemptively dismissed as input errors or problems with quality. While certain

issues may be more significant than others, they all create the possibility of confusion and, in aggregate, reduce trust in the reliability of metadata for conveying meanings and identities. The issues and the questions they raise require further research and consultation with stakeholder groups in scholarly publishing as well as with regional and disciplinary communities to ascertain if and how communities are variously impacted.

Specific to the categories identified in this review, consultation with publishers, editors, authors, and other creators of metadata is needed to confirm the nature and scope of issues (as technical or cultural, and intentional or accidental). While our analysis was able to determine the breadth of issues that have a cultural dimension, more work is needed to understand the reason why the issues exist, including metadata creators' intentions when inputting or recording data in these ways. Such discussions would also need to identify current and desired uses and functionalities of metadata, and to determine how tools and infrastructure can be adjusted or created to enable quality metadata creation and transmission.

In the absence of established good practices for multilingual metadata creation, community engagement would also provide critical insights for policies, recommendations, and guidance that address issues related to the Language category. The COAR Task Force on Supporting Multilingualism and non-English Content in Repositories (2022) confirms and addresses the issue of missing language attributes, recommending that repositories "include a tag in the language metadata field that identifies the language of the resource, and a tag that identifies the language of the metadata" in all records. These tags inform how systems parse and index content, which means that proper tagging will result in more accurate and effective discovery and indexing services. More consistent tagging should therefore be coupled with improvements to multilingual indexing in scholarly systems.

Training and guidance materials may also help increase awareness, understanding, and use of elements and attributes available in schemas and standards. For instance, the @xml:lang Language attribute in the JATS schema allows subtags for defining the language, script, and regional variant used for the content of an element (NCBI & NLM, 2021). Their adoption would enhance records that contain a mixture of values in translation, transliteration, and original scripts (such as example 2 under "Language" in Table 4) by indicating the various languages present; they may also help prevent issues such as the inclusion of multiple languages in a single field. Lapeyre & Usdin (2011) provide detailed guidance on the JATS elements and attributes that can be used to create records that are reflective of multilingual content.

Our view of articles with issues related to publishing in a language other than English or in multiple languages (which may or may not include English) suggests that some editors may struggle to produce metadata that reflects the diversity of their contributors and their linguistic practice, and/or to locate the sufficient financial, human, and technical resources required to translate and process metadata. It may very well be that, in areas where resources are particularly constrained, the presence of translated titles and abstracts in metadata depends on the ability and/or willingness of authors to provide their own translations.

For some journals seeking more plural representation, policies or recommendations have been developed to support representing a more holistic range of languages, conventions, and practices. Some strategies are: requiring titles, abstracts, and keywords be provided in the language of the manuscript as well as the publisher's national language and for affiliation names to be given in their national language (Revista, n.d., sec. Language and study areas); committing to publish author names in Chinese, Japanese, and Korean alongside English vari-

ants and providing technical guidance for doing so (AIP Publishing, n.d., sec. Guidelines for Using Chinese Japanese, and Korean Names); or suggesting that authors "provide a second abstract in their native language or the language relevant to the country in which the research was conducted" (British Ecological Society, n.d., sec. Manuscript Specifications).

These approaches need not be mutually exclusive; however, they may depend on the affordances and restrictions of schemas and interfaces for inputting and displaying metadata. Publishing tools and solutions in place should first be tested to ensure that metadata entered into the system can be transmitted and displayed accurately along both technical and cultural lines. The utility and impact of such strategies may also depend upon where additional language versions are published: in the journal platform and/or in the article PDF, for instance. Journal publishing services might also explore linked data methods to support multilingualism and cross-linked name references in publication metadata (Niininen et al., 2017; El-Sherbini, 2018; Hardesty & Nolan, 2021). Fields already exist for persistent ORCiD identifiers for researcher profiles, which can be utilized for linked data initiatives.

Certain issues may be unique to those assuming an English-first approach with the goal of increased indexing and discoverability. For items providing titles and abstracts in multiple languages, metadata records may only include the English version regardless of the language of the text itself. This approach could also result in publisher names, journal titles, and institutional affiliations appearing in English translation and/or transliteration only, regardless of the accepted language/s for publication or the original language of names and titles (e.g., Table 4, example 1 under "Geography"). Such a strategy may be indicative of the influence of prominent indexing services on the construction of metadata (Arastoopoor & Ahmadinasab, 2019, 223). To be considered for inclusion in Clarivate's Web of Science citation database, for instance, journals must provide titles and abstracts in English and bibliographic information in Roman script, regardless of the language of publication (e.g., Clarivate, n.d.).

Many issues in the Naming and Status categories relate to the use of fields to record information that does not align with the defined scope of the field; this may be due to an absence of more appropriate options or lack of clarity around existing ones. Obstacles for authors, journals, and other metadata creators to present names and status information appropriately may appear more immediately in journal publishing and hosting systems and user interfaces, or downstream in indexing and discovery platforms. Elements related to persons and their attributes and scope notes could also be revised or expanded to account for a broader range of naming conventions, and to enable notations of status and/or titles alongside affiliations. Such changes would accommodate cases like the one described in Table 4 by allowing Indonesian authors to input a single or multipart given name with no family name—common name forms in Indonesia—instead of repeating their given name in the family name field to comply with required fields. It could also lead to a decreased presence of titles like "Dr." or "Professor" or the use of capitalization in given and family name or other fields to indicate seniority and status, as the examples in Table 4 and 5 show.

Directions for Future Research

More than providing definitive conclusions about the state of metadata quality, this study raises further questions that warrant the attention of the scholarly community. While our team has been intent on addressing the first of the following questions (Donathan II et al., forthcoming), we call on the community to seek to address the following:

- To what extent are the metadata issues identified in this study present in the scholarly record?
- How does technical infrastructure exacerbate these issues? For instance, are indexing and discovery services capable of handling metadata in different languages well, and are user interfaces designed for non-Roman characters and multidirectionality? How well do systems operate independently and together to enable metadata exchanges that remain culturally attuned?
- Are English translations or Romanizations used intentionally to increase opportunities for indexing and metadata harvesting? How do these choices impact the discoverability and accessibility of content by those working in non-English languages and/or non-Romanized language forms?
- Whether because personal names are closely tied to identity or because Romanizations make professional interactions smoother, when are Romanized names in fact the preferred name of an author? When are Romanized names in fact the only names for an author?
- When affiliations are noted, how often are home institutions recorded as compared to affiliated or partner institutions? What are the consequences of including one or the other, or both?
- What should best or good practices be for journals that accept and publish full-text articles in multiple languages or publish titles and abstracts in multiple languages? If a journal changes its language policy, should metadata be retroactively updated to reflect or make note of this change? Would such updates have meaningful impacts?
- How can standards, best practices, and goals for interoperability be balanced against heterogeneous cultural, epistemic, and resourcing realities?
- Who is metadata being created for, for what purpose/s, and why?

Limitations

As previously stated, this review is the result of one author's interpretation of the sampled records and articles. It is therefore an incomplete picture of the cultural issues present in the sample and across all journal article metadata. Any issues that were overlooked or misinterpreted deserve attention, and efforts should be made to address these in other projects.

Scoped by the elements available in JSON-formatted records, the authors do not fully address issues resulting from the absence of elements—in the schema, data model, or end-user interface—to which values can be assigned, such as keywords, Romanization or transliteration styles, or professional or community titles. Studies to identify elements and standardized values that could be added to metadata schemas and standards to enhance cultural representation would provide further clarity for next steps. Where this research did not involve a rigorous close reading of the associated articles, separate studies may also attend to cultural issues related to the quality of subject analysis as well as relationships between the accuracy of subject analysis and the prevalence of cultural metadata issues.

This review hopes to prompt further investigations into metadata practices and issues specific to given disciplines, cultures, regions, and languages that are not explored in depth here. Likewise, the impact of regional publishing and research norms on metadata creation, the size and resourcing available to publishers, or the cultural downstream effects of the identified issues may be taken up in the future. Focusing largely on academic journal articles in this review, later studies might also examine metadata for other primary and secondary

resource types. Building on the work of Barnett at al. (2010), further studies specific to the ways in which metadata are interpreted downstream by systems and organizations, such as search and cataloging platforms, libraries, and citation management systems, would also be useful.

Conclusion

Viewing metadata as informational objects in their own right encourages us to consider records beyond functional objects requiring technical accuracy to support resource use and discovery. As we build, refine, and expand our publishing infrastructures and resource discovery systems, we must recognize that metadata is not a mechanism created solely to connect end users to resources. Cultural issues should be foregrounded during the review and development of local journal policies, research and publishing practices, technical training, and metadata systems and standards.

Instead, as informational objects, metadata records should be treated as sites in need of critical, intellectual engagement to surface the perspectives and identities embedded and obscured in their creation. In taking up the responsibility of describing a researcher's output in a record, journal editors and publishers also have a responsibility to the researcher to ensure that their contributions and identity are represented as fully as relevant and possible to their work and the communities most affected by it.

Efforts such as the 2019 Helsinki Initiative on Multilingualism in Scholarly Communication, 2021 Coalition Publica Metadata Working Group report, and COAR Task Force on Supporting Multilingualism and non-English Content in Repositories, struck in August 2022, speak to the importance of supporting the dissemination of and access to locally relevant research and nurturing regional publishing infrastructures. Ensuring metadata appropriately and respectfully represent cultural identities and nuances is one step toward that goal.

Acknowledgments

Special thanks to Crossref, especially Jennifer Kemp and Isaac Farley, for their support throughout this project, and Dennis Donathan II for his invaluable contributions to the project team. Thanks also to Coalition Publica for scholarship funding.

References

Adler, M. (2017). *Cruising the library: Perversities in the organization of knowledge*. Fordham University Press.

AIP Publishing. (n.d.). *Author Instructions*. Retrieved March 27, 2023, from https://publishing.aip.org/resources/researchers/author-instructions/#cjk

Alamri, B. (2021). Multilingual scholars' experiences in publishing in the social sciences and humanities. *Journal of Scholarly Publishing*, 52(4), 248–272. https://doi.org/10.3138/jsp.52.4.04

Alhasnawi, S. (2021). English as an academic lingua franca: Discourse hybridity and meaning multiplicity in an international Anglophone HE institution. *Journal of English as a Lingua Franca*, 10(1), 31–58. https://doi.org/10.1515/jelf-2021-2054

Arastoopoor, S., & Ahmadinasab, F. (2019). From personal to corporate and from names to titles: The challenges of Iranian scholars with scientific publications. In J. Sandberg (Ed.), *Ethical questions in name authority control* (pp. 72–98). Library Juice Press.

Balula, A., & Leão, D. (2021). Multilingualism within scholarly communication in SSH. A literature review. JLIS.It, 12(2), 88–98. https://doi.org/10.4403/jlis.it-12672

Barnett, J., Lovins, D., Novak, A., Riley, C., & Suzuki, K. (2010). *Investigating multilingual, multi-script support in Lucene/Solr library applications*. Faculty Digital Archive: NYU Libraries. http://hdl.handle.net/2451/38726

Billey, A., Drabinski, E., & Roberto, K. R. (2014). What's gender got to do with it? A critique of RDA 9.7. *Cataloging & Classification Quarterly*, 52(4), 412-421. https://doi.org/10.1080/01639374.2014.882465

Billings, L., Llamas, N. A., Snyder, B. E., & Sung, Y. (2017). Many languages, many workflows: Mapping and

analyzing technical services processes for East Asian and International Studies materials. *Cataloging & Classification Quarterly*, 55(7-8), 606-629. https://doi.org/10.1080/01639374.2017.1356783

British Ecological Society. (n.d.). *Methods in Ecology and Evolution: Author Guidelines*. Wiley. Retrieved March 27, 2023, from https://besjournals.onlinelibrary.wiley.com/hub/journal/2041210X/author-guidelines

Bruce, T., & Hillmann, D. (2004). *The continuum of metadata quality: Defining, expressing, exploiting*. eCommons. https://hdl.handle.net/1813/7895

Canagarajah, A. S. (2002). *A geopolitics of academic writing*. University of Pittsburgh Press.

Cataloging Ethics Steering Committee. (2021). Cataloguing code of ethics. https://docs.google.com/document/d/1IBz7nXQPfr3U1P6Xiar9cLAkzoNX_P9fq7eHvzfSlZ0/edit?usp=sharing

Clarivate. (n.d.). *Web of Science journal evaluation process and selection criteria*. Retrieved December 4, 2022, from https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/web-of-science/core-collection/editorial-selection-process/editorial-selection-process/

Coalition Publica Metadata Working Group. (2021). *Technical report: Metadata feedback for Coalition Publica*. Erudit. https://www.erudit.org/public/documents/CP_Technical_Report.pdf

Crystal, D. (2012). *English as a global language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139196970

Curry, M. J., & Lillis, T. M. (2010). Academic research networks: Accessing resources for English-medium publishing. *English for Specific Purposes*, 29(4), 281–95. https://doi.org/10.1016/j.esp.2010.06.002

Dartmouth Library Metadata Services. (n.d.). *Troubleshooting guide for diacritics*. Retrieved November 27, 2022, from https://www.dartmouth.edu/library/catmet/cataloging/diacritics-troubleshooting.html

Duarte, M. E., & Belarde-Lewis, M. (2015). Imagining: Creating spaces for Indigenous ontologies. *Cataloging & Classification Quarterly*, 53(5-6), 677-702. https://doi.org/10.1080/01639374.2015.1018396

Ducheva, D. P., & Pennington, D. R. (2019). Resource description and access in Europe: Implementations and perceptions. *Journal of Librarianship and Information Science*, 51(2), 387-402. https://doi.org/10.1177/0961000617709060

El-Sherbini, M. (2018). *Improve discoverability of non-Roman materials*. ALA Webinar. Retrieved February 18, 2023, from https://www.ala.org/alcts/confevents/upcoming/webinar/041818

Farnel, S. (2018). Metadata as data: Exploring ethical metadata sharing and access for Indigenous resources through OCAP principles. *Proceedings of the Annual Conference of CAIS / Actes Du congrès Annuel De l'ACSI*. https://doi.org/10.29173/cais974

Farnel, S., Shiri, A., Campbell, S., Cockney, C., Rathi, D., & Stobbs, R. (2017). A community-driven metadata framework for describing cultural resources: The Digital Library North Project. *Cataloging & Classification Quarterly*, 55(5), 289–306. https://doi.org/10.1080/01639374.2017.1312723

Gartner, R. (2016). What metadata is and why it matters. In *Metadata* (pp. 1-13). Springer. https://doi.org /10.1007/978-3-319-40893-4_1

Ge, M. (2015). English writing for international publication in the age of globalization: Practices and perceptions of mainland Chinese academics in the humanities and social sciences. *Publications*, 3(2), 43-64. https://doi.org/10.3390/publications3020043

Hardesty, J. L., & Nolan, A. (2021). Mitigating bias in metadata: A use case using Homosaurus linked data. *Information Technology and Libraries*, 40(3). https://doi.org/10.6017/ital.v40i3.13053

Heery, R., & Patel, M. (2000). Application profiles: Mixing and matching metadata schemas. *Ariadne*, (25). http://www.ariadne.ac.uk/issue25/app-profiles/

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

Khanna, S., Ball, J., Alperin, J. P., & Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, 3(4), 912–930. https://doi.org/10.1162/qss_a_00228

Kim, S., & Cho, S. (2013). Characteristics of Korean personal names. *Journal of the American Society for Information Science and Technology*, 64(1), 86-95. https://doi.org/10.1002/asi.22781

Király, P., Stiller, J., Charles, V., Bailer, W., & Freire, N. (2019). Evaluating data quality in Europeana: Metrics for multilinguality. In E. Garoufallou, F. Sartori, R. Siatri, & M. Zervas (Eds.), *MTSR 2018: Metadata and semantic research. Communications in computer and information science* (Vol. 846). Springer. https://doi.org/10.1007/978-3-030-14401-2_19

Lapeyre D. A., & Usdin, B. T. (2011). Introduction to multi-language documents in NISO JATS. In *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2011*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK62175/

Library Publishing Coalition. (2018). *An ethical framework for library publishing*, version 1.0. Educopia. http://dx.doi.org/10.5703/1288284316777

Mahmoud, M.S.A., & Al-Sarraj, M.M. (2018). Bilingual Qatar Digital Library: Benefits and challenges. In M.

Dobreva, A. Hinze, & M. Žumer (Eds.), *Maturity and Innovation in Digital Libraries. ICADL 2018. Lecture Notes in Computer Science* (Vol. 11279). https://doi.org/10.1007/978-3-030-04257-8_19

Malički, M., & Alperin, J. P. (2020, April 8). *Four recommendations for improving preprint metadata*. Scholarly Communications Lab. https://www.scholcommlab.ca/2020/04/08/preprint-recommendations/

Matusiak, K.K., Meng, L., Barczyk, E., & Shih, C.J. (2015). Multilingual metadata for cultural heritage materials: The case of the Tse-Tsung Chow Collection of Chinese scrolls and fan paintings. *The Electronic Library*, 33(1), 136-151. https://doi.org/10.1108/EL-08-2013-0141

National Center for Biotechnology Information (NCBI) & National Library of Medicine (NLM). (2021). *Attribute: language*. Journal Archiving and Interchange Tag Library NISO JATS Version 1.3 (ANSI/NISO Z39.96-2021). https://jats.nlm.nih.gov/archiving/tag-library/1.3/attribute/xml-lang.html

Niininen, S., Nykyri, S. and Suominen, O. (2017). The future of metadata: Open, linked, and multilingual – the YSO case. *Journal of Documentation*, 73(3), 451-465. https://doi.org/10.1108/JD-06-2016-0084

Olson, H. A. (2001). The power to name: Representation in library catalogs. *Signs*, 26(3), 639–668. http://www.jstor.org/stable/3175535

Park, J-R. (2007). Cross-lingual name and subject access: Mechanisms and issues. *Library Resources and Technical Services*, 51(3), 80-89.

Pho, P. D., & Tran, T. M. P. (2016). Obstacles to scholarly publishing in the social sciences and humanities: A case study of Vietnamese scholars. *Publications*, 4(3), 19. https://doi.org/10.3390/publications4030019

PIE-J Working Group. (2013). *NISO RP-16-2013, PIE-J: The presentation & identification of e-journals*. National Information Standards Organization. https://groups.niso.org/higherlogic/ws/public/download/10368

Pomerantz, J. (2015). Definitions. In *Metadata* (pp. 20-64). MIT Press.

Revista Brasileira de Engenharia Agrícola e Ambiental. (n.d.). *Submissions*. Retrieved March 27, 2023, from https://submission.scielo.br/index.php/rbeaa/about/submissions

Rigby, C. (2015). Nunavut Libraries Online establish Inuit language bibliographic cataloging standards: Promoting Indigenous language using a commercial ILS. *Cataloging & Classification Quarterly*, 53(5-6), 615–639. https://doi.org/10.1080/01639374.2015.1008165

Santos, J. V., & Da Silva, P. N. (2016). Issues with publishing abstracts in English: Challenges for Portuguese linguists' authorial voices. *Publications*, 4(2), 12. https://doi.org/10.3390/publications4020012

Shi, J., Nason, M., Tullney, M., & Alperin, J.P. (2023). Data for: Identifying Metadata Quality Issues Across Cultures (V1) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/GZI7IA

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.

Shiraishi, N., Chou, C., Fu, L., & Zou, X. (2021). CEAL Task Force for Review of the ERMB interim report. *Journal of East Asian Libraries*, *2021*(173), 4. https://scholarsarchive.byu.edu/jeal/vol2021/iss173/4

Soglasnova, L. (2018). Dealing with false friends to avoid errors in subject analysis in Slavic cataloging: An overview of resources and strategies. *Cataloging & Classification Quarterly*, 56(5-6), 404-421. https://doi.org/10.1080/01639374.2018.1438551

Tomuschat, C. (2017). The (hegemonic?) role of the English language. *Nordic Journal of International Law*, 86(2), 196–227. https://doi.org/10.1163/15718107-08602003

The Trans Metadata Collective (2022). *Metadata best practices for trans and gender diverse resources* (1.5). Zenodo. https://doi.org/10.5281/zenodo.6829167

Turner, J. (2018). *On writtenness: The cultural politics of academic writing*. Bloomsbury Academic.

W3C Internationalization Working Group. (2022). *Strings on the web: Language and direction metadata* [W3C Group Draft Note]. https://www.w3.org/TR/string-meta/

Woodley, M. S. (2016). Metadata matters: Connecting people and information. In M. Baca (Ed.), *Introduction to metadata* (3rd ed). Getty Publications. http://www.getty.edu/publications/intrometadata/metadata-matters/

Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata*, 11, 51-62. https://doi.org/10.1080/19386389.2011.570654

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality assessment for linked open data: A survey. *Semantic Web – Interoperability, Usability, Applicability*, 1-5. http://www.semantic-web-journal.net/content/quality-assessment-linked-open-data-survey

Zeng, M. L. (2018). Interoperability. In B. Hjørland & C. Gnoli (Eds.), *Encyclopedia of Knowledge Organization*. International Society for Knowledge Organization. https://www.isko.org/cyclo/interoperability

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

Appendix A

Count of item types other than journal article

article	309
proceedings	29
book review	14
chapter	12
technical report	11
protocol	9
digitized backfile	6
journal issue	4
letter to editor	4
retraction	3
editorial	2
encyclopedia entry	2
end matter	2
index	2
news	2
advertisement	1
bibliography	1
book	1
brief	1
communication	1
contributor list	1
editor note	1
issue section	1
journal	1
listicle	1
miscellanea	1
notice	1
notice of meeting	1
technical note	1
table of contents	1
translation	1

Appendix B

Mapping of 32 unique issues to the categories, with examples

Issue	Language	Naming	Status	Geography	Contribution	General
value absent	10.1590/s1516-44462005000100001 (field: language)		10.1590/s1516-44462005000100001 (field: affiliation)	10.1590/s1516-44462005000100001 (field: publisher-location)	10.18535/jmscr/v7i5.40 (field: author)	10.1590/s1516-44462005000100001 (field: license)
translation absent	10.54161/jrs.v2i1.61 (field: abstract and title)					
value in original language absent	10.32598/jmsp.6.4.686 (field: all)	10.3820/jjpe.22.s57 (field: publisher)		10.1080/23802359.2019.1710605 (field: affiliation)		
lang attribute absent	10.1163/1571805042782109 (field: title)					
lang style absent: Romanization only	10.1556/ahista.47.2006.1-4.6 (field: publisher)	10.1016/s1003-6326(20)65424-3 (field: author-family)				
lang style absent: Romanization absent	10.15750/chss..35.201002.013 (field: author)					
vor license terms absent						10.1530/acta.0.0070017ff (field: license)
author/s absent					10.28933/ijsr-2020-12-1605 (field: author)	
not all authors listed					10.31080/asol.2021.03.0355 (field: author)	
ORCIDs absent					10.32598/sija.16.2.1600.1 (field: author)	

Issue	Language	Naming	Status	Geography	Contribution	General
not all persons listed					10.2478/v10008-007-0012-2 (field: N/A – absent person is translator)	
affiliations absent for all authors			10.15556/ijsim.01.02.001 (field: author)			
affiliations absent for all editors			10.1055/b-0036-132151 (field: editor)			
not all publishers listed					10.29252/rmm.5.1.44 (field: publisher – university co-publisher absent)	
related orgs absent					10.1111/j.1945-5100.1996.tb02122.x (field: assertion – rightsholder absent)	
location absent			10.18535/jmscr/v7i5.40 (field: author-1 – affiliation listed as an author)	10.1016/j.forpol.2020.102283 (field: publisher)		
subtitle absent						10.32598/sija.16.2.1600.1 (field: container-title)
outdated						10.1530/acta.0.0070017 (field: container-title)

Issue	Language	Naming	Status	Geography	Contribution	General
registered URL out of date						10.15556/ijiim.02.01.003 (field: doi)
registered URL invalid						10.15556/ijsim.01.02.003 (field: doi)
value in record does not match information on container website					10.29252/archhygsci.8.2.119 (field: publisher)	10.17504/protocols.io.8ihhub6 (field: abstract)
inaccurate	10.3820/jjpe.22.s57 (field: language)					10.15556/ijsim.01.02.003 (field: subject)
affiliations presented as authors	10.15863/tas.2014.02.10.30 (field: author)		10.15863/tas.2014.02.10.30 (field: author)			
multiple languages in single field	10.1163/1571805042782109 (field: container-title)	10.14710/jadu.v2i2.7641 (field: publisher)		10.1590/0074-02760210176 (field: author-affiliation-name)		
multiple values in single field	10.1016/s0005-2760(98)00140-4 (field: container-title)	10.14710/jadu.v2i2.7641 (field: publisher)	10.29252/archhygsci.8.2.119 (field: author-2 and author-3 – affiliations listed as authors)		10.1590/0074-02760210176 (field: publisher)	
original-title used incorrectly: includes value in original language but item is not a translation	chss.72.201905.006 (field: original-title)					

Issue	Language	Naming	Status	Geography	Contribution	General
original-title used incorrectly: value repeated	10.18535/jmscr/v8i4.83 (field: original-title)					
all authors listed as first			10.12697/akut.2019.25.07 (field: author-sequence)			
first author not identified			10.2118/206525-ms (field: author-sequence)			
input in all caps		10.1016/s1003-6326(20)65424-3 (field: author-family)	tpmj/2009.16.02.2939 (field: author)			
additional persons listed					10.2307/4147866 (field: author-1 and author-2 – authors of reviewed work)	
value incomplete	10.1093/ehr/cel085 (field: title)	10.20523/sapereaude-ano4-vol-12-pg-143-165 (field: author-family)				10.20527/jurnalsocius.v3i2.3259 (field: container-title)
only provides initial/s		10.15863/tas.2019.06.74.35 (field: author)				
acronym only		10.1590/1807-1929/agriambi.v19n4p317-323 (field: publisher and author-affiliation)				

Appendix C

Count of all issues, total and by level and field

Issue	Total	Item DOI	Item Abstract	Item Title	Item License	Person General	Person Given Name	Person Family Name	Person Affiliation	Container Publisher	Container Title	Container Language	Container Subject	Container Rights
registered URL invalid	14	14	0	0	0	0	0	0	0	0	0	0	0	0
registered URL out of date	18	18	0	0	0	0	0	0	0	0	0	0	0	0
value absent	1348	0	221	14	318	0	23	0	1	0	0	253	108	410
translation absent	207	0	51	48	0	0	0	0	0	12	39	0	57	0
value in original language absent	214	0	8	19	0	0	1	2	44	58	25	0	57	0
incorrectly input	246	0	16	37	0	8	24	55	25	63	15	0	3	0
original-title used incorrectly	22	0	0	22	0	0	0	0	0	0	0	0	0	0
language attribute absent	641	0	44	106	0	0	123	134	13	64	112	0	45	0
multiple languages in single field	63	0	3	3	0	0	0	0	9	0	48	0	0	0
value incomplete	21	0	2	0	0	0	0	2	3	10	4	0	0	0
value in record does not match information on container website	71	0	3	2	0	0	0	0	0	49	17	0	0	0
vor license terms absent	63	0	0	0	63	0	0	0	0	0	0	0	0	0
additional persons listed	6	0	0	0	0	6	0	0	0	0	0	0	0	0
author/s absent	41	0	0	0	0	41	0	0	0	0	0	0	0	0
not all authors listed	19	0	0	0	0	19	0	0	0	0	0	0	0	0
all authors listed as first	2	0	0	0	0	2	0	0	0	0	0	0	0	0
first author not identified	13	0	0	0	0	13	0	0	0	0	0	0	0	0
ORCIDs absent	6	0	0	0	0	6	0	0	0	0	0	0	0	0
not all persons listed	1	0	0	0	0	1	0	0	0	0	0	0	0	0

Issue	Total	Item DOI	Item Abstract	Item Title	Item License	Person General	Person Given Name	Person Family Name	Person Affiliation	Container Publisher	Container Title	Container Language	Container Subject	Container Rights
language style absent	290	0	0	0	0	0	120	130	2	23	15	0	0	0
only provides initial/s	72	0	0	0	0	0	66	6	0	0	0	0	0	0
absent for all authors	297	0	0	0	0	0	0	0	297	0	0	0	0	0
absent for all editors	4	0	0	0	0	0	0	0	4	0	0	0	0	0
acronym only	28	0	0	0	0	0	0	0	7	21	0	0	0	0
affiliations presented as authors	42	0	0	0	0	0	0	0	42	0	0	0	0	0
multiple values in single field	56	0	0	0	0	0	0	0	18	33	5	0	0	0
location absent	401	0	0	0	0	0	0	0	2	399	0	0	0	0
inaccurate	158	0	0	0	0	0	0	0	0	0	0	47	111	0
not all publishers listed	6	0	0	0	0	0	0	0	0	6	0	0	0	0
related orgs absent	4	0	0	0	0	0	0	0	0	4	0	0	0	0
outdated	7	0	0	0	0	0	0	0	0	0	7	0	0	0
subtitle absent	6	0	0	2	0	0	0	0	0	0	4	0	0	0
total count	4387	32	348	253	381	96	357	329	467	742	291	300	381	410