SFU PKP PUBLIC KNOWLEDGE PROJECT

scholcommlab
https://scholcommlab.ca

Dataset Description: Identifying metadata quality issues across cultures

July 25, 2023

Method

Authors conducted a manual review of 427 JSON-formatted metadata records to identify metadata quality issues related to cultural meanings and identities. Identified issues were then defined and categorized to consider the ways in which issues in metadata quality impact how individuals and communities represent themselves.

Sampling

This sample was drawn from the Crossref API on March 8, 2022. The sample was constructed purposefully on the hypothesis that records with at least one known issue would be more likely to yield issues related to cultural meanings and identity.

Records known or suspected to have at least one quality issue were selected by the authors and Crossref staff. The Crossref API was then used to randomly select additional records from the same prefix. Records in the sample represent 51 DOI prefixes that were chosen without regard for the manuscript management or publishing platform used, as well as 17 prefixes for journals known to use the Open Journal Systems manuscript management and publishing platform. OJS was specifically identified due to the authors' familiarity with the platform, its international and multilingual reach, and previous work on its metadata quality.

Data files

The dataset consists of 5 files:

1. `recordReview.csv` lists the metadata records in the sample and the cultural issues identified in each record, organized by metadata field. Rows represent records and columns correspond to the metadata fields in which issues were found. When multiple issues are noted in a single metadata field for a record, issues are semi-colon (;) separated. Additional information for each record, including the number of issues (cultural only and in total), language policy of the container (e.g., journal, book, etc.), publishing platform, and item type, are also noted.

2. `recordReviewCodebook.csv` explains the organization of recordReview.csv and how column headings map to metadata fields.

3. `issuesCountCultural.csv` provides a breakdown of the number of cultural issues found by element and issue.

4. `issuesGrouping.csv` presents the authors' categorization of the issues, with definitions and examples for each unique issue. An ad-hoc classification of importance is also offered for transparency.

5. `issuesValueList.csv` Where more granular details are provided with umbrella issue types (e.g., "incorrectly input"), definitions and examples of these granular values are provided. The values correspond to those listed in column F "issueValueList" in the issueGrouping sheet, and in the recordReview sheet, they are recorded as "issueDetails: issueValue" (e.g., "incorrectly input: as null").