Descriptive richness of free-text metadata: a comparative analysis of three language archives

Mary Burke, University of North Texas, mary.burke@unt.edu

Oksana L. Zavalina, University of North Texas, oksana.zavalina@unt.edu

Abstract

As archiving became a priority in documentary linguistics only in the 1990's, existing research indicates that language archives are not yet up to date on best practices in information organization. As a result, metadata in language archives varies substantially, depending on the depositor, self-upload procedures, and metadata creation guidelines. Focusing on free-text metadata, known to provide rich information, this study analyzes item-level metadata in three language archives: the Endangered Language Archive, Pacific Regional Archive for Digital Sources in Endangered Cultures, and the Archive of the Indigenous Languages of Latin America. The study identified categories of information included in Description metadata fields and the relative distribution of these categories. The most commonly occurring categories of information observed in this study can serve as a basis for the development of best practice guidelines for item-level metadata in language archives.

keywords: language archives, free-text metadata, comparative analysis, metadata quality

83rd Annual Meeting of the Association for Information Science & Technology | Pittsburgh, Pennsylvania| October 24-28, 2020. Author(s) retain copyright, but ASIS&T receives an exclusive publication license.

1. Introduction

Language archives are defined by their goal of connecting users worldwide to language data (oral histories, examples of speech and writing, etc.). Digital accessibility of primary language data has been recognized by the linguistics community to support research reproducibility and the creation of pedagogical materials (Henke & Berez-Kroeker, 2016; Berez-Kroeker et al., 2018). As language data transitioned to digital formats, the Open Language Archives Community (OLAC) was formed in the early 2000's to develop best practices for language archiving and facilitate interoperable repositories of language data. OLAC and The Digital Endangered Languages and Musics Archives Network (DELAMAN) have educated the linguistics community on the value of archiving

and the creation of basic metadata to accompany language data.

Despite these significant advancements, the field of linguistics is not up to date on best practices in information organization or user-centered web design. As a result, metadata in language archives is highly variable, depending on content management systems, self-upload procedures, metadata creation guidelines, and, crucially, depositors. Collections in language archives come from linguistics researchers worldwide, meaning the metadata of each collection has a different creator: typically, a linguist with minimal (if any) prior experience creating descriptive metadata. Effective metadata management relies on metadata quality evaluations, yet the current understanding of the quality of metadata in language archives is insufficient. The study reported in this paper aims to provide empirical data to address this gap in conjunction with an IMLS-funded project holistically investigating information organization in language archives and how it meets the needs of intended audiences (language communities, educators, researchers) primarily in the United States of America. This analysis adds a metadata quality evaluation component from an international perspective not otherwise captured by the ongoing research.

Metadata describing individual information objects and collections in digital libraries and archives broadly falls in two major categories: metadata taken from controlled vocabularies (e.g., the MODS Name field relies on the Library of Congress Name Authority File, etc.), and free-text metadata. While it is well-known that controlled-vocabulary metadata supports information discovery, the free-text metadata, in particular the Dublin Core metadata field Description and its equivalents in other schemes (e.g., Abstract, Note, and Table of Contents in MODS) have been found to add substantial value to metadata records by providing rich contextual information (Roeschley et al., 2020; Tarver et al., 2016; Zavalina et al., 2008).

This study is the first to systematically analyze free-text metadata in language archives. It examines and compares metadata equivalent to the Dublin Core Description field in three large-scale language archives maintained by institutions in North America, Europe, and Australia respectively: the Archive of the Indigenous Languages of Latin America (AILLA), the Endangered Language Archive (ELAR), and the Pacific Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

2. Related work

Because language archives are an emerging phenomenon, there is a lack of studies on metadata quality in language archives. OLAC attempted an evaluation of Dublin Core records obtained through OAI-PMH based only on completeness

(Hughes, 2004); however, because records were converted into Dublin Core prior to harvesting, any metadata elements with no Dublin Core equivalent were excluded. Studies on general metadata quality note that it is highly context-dependent (e.g., Park, 2009; Bruce & Hillman, 2004). Still, various metrics exist to evaluate metadata quality, such as completeness, accuracy, provenance, conformance to expectation, logical consistency, coherence, timeliness, and accessibility (Bruce & Hillman, 2004). The Information Quality framework (Stvilia et al., 2007), designed for information resources in general, was developed using Dublin Core metadata records and encyclopedia articles, extending to the underlying context of metadata creation and users' interactions with information. More recent work on metadata quality has synthesized these two major frameworks, categorized metadata quality criteria into intrinsic, relational and reputational, and added dimensions of metadata enriching (e.g., Park, 2009; Alemu, 2018). These criteria make increasingly nuanced evaluations of metadata quality possible; however, evaluations based on these criteria require records in one common scheme. Because language archives typically use locally-developed metadata schemes wherein the number and definitions of metadata elements vary, language archive metadata records are not directly comparable (Burke & Zavalina, 2019).

Other evaluations focus on collection-level metadata records. For example, Zavalina (2014) found that, despite redundancies in Description and Coverage metadata fields, free-text Descriptions contained information not represented elsewhere in the record. Metadata creators were able to provide more detailed information in the free-text Description field than was possible through controlled vocabularies alone, demonstrating that the combination of free-text and controlled-vocabulary metadata elements provides richer access (Zavalina, 2014). Another investigation of archival descriptions created by contributors (Roeschley et al., 2020) finds additional contextual information not seen in cataloguer-created metadata, though context may take precedence over the content. Although language archive metadata is also created by contributors (researchers), Harris et. al (2019) note a "surprising lack of detail or information" in the Description field (p. 140). Burke and Zavalina (2019) summarize the results of a content analysis of 20 language archives in the United States, Australia, Canada, and Europe, noting the prevalence of locally-developed metadata schemes and variation in the level of detail in metadata creation guidelines. Interviews with users, depositors, and managers reveal that language archives struggle with funding and implementing authority control (Burke & Zavalina, 2020).

The language archives under consideration in this study are well-established members of the DELAMAN board, with language data and metadata records

freely accessible online. These archives include multiple collections originating from different depositors, using the 'bundle' as a unit, which typically includes an audio or video recording and documents such as a transcription or translation of the recording. The Endangered Language Archive (ELAR) has metadata records containing up to 12 elements, including free-text Description. Jointly hosted by three Australian universities, the Pacific Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) metadata scheme consists of 21 elements. The Archive of the Indigenous Languages of Latin America (AILLA) is a regionally-focused archive, prioritizing deposits from South and Central America, known for its bilingual interface (English and Spanish). AILLA's metadata scheme has 17 metadata elements, including free-text Description and Description [Indigenous] fields. Though they share many metadata fields (e.g., Title, Description, Language) AILLA and PARADISEC contain multiple elements not present in ELAR.

Taking into account the previous findings regarding the information discovery potential of the free-text Description metadata field and the gaps identified by Burke and Zavalina (2019; 2020), this analysis investigates how the Description field is used in language archives.

3. Methodology

This project uses content analysis to evaluate the data values in free-text Description fields of item-level metadata records in the language archives introduced above. The research questions addressed are as follows:

- 1) What are the categories of information included? What are the differences and similarities across the three archives?
- 2) What are the differences and similarities in quantitative characteristics of the data values in free-text Description fields across these archives: the length of data values, the number of categories of information included, etc.?

The data values of Description metadata fields were collected from 130 metadata records from each archive for a total of 390 records. Records were chosen randomly from different collections to avoid records created by the same depositor. Data values were coded for 28 categories of information developed based on the metadata creation guidelines of the archives under consideration and the Computational Resource for South Asian Languages (CoRSAL), another language archive housed at the University of North Texas, as well as on types of information in Description metadata fields observed by Zavalina et al. (2008). A spreadsheet was used for analysis, with new categories added on an as-needed basis in the process. The binary coding approach was used: the code '1' was entered if the category was present; '0' if not. To compensate for researcher bias,

a known limitation of content analysis, an additional coder coded 20% of the dataset, or 78 records. Inter-coder reliability was calculated at 95%. Descriptive statistics were measured for each archive individually and cumulatively.

4. Results

Results of this initial analysis empirically support informal observations of high variation in language archive metadata. The word count and number of descriptive categories are shown in Table 1.

Table 1. Word counts and number of categories of information in Description fields

	Minimum	Maximum	Average	Median	Standard Deviation
Number of Categories	0	10	2.8	2	1.7
Word Count	1	706	52.9	20.5	83.1

The majority of descriptions contain 2 of the 28 categories included in this study. The standard deviation of the number of categories is much lower than in word count. Ten total descriptions contain none of the categories identified; typically these are short (e.g., '*A/R = Rabi*') without adequate context to assign any code. 18.7% of descriptions have only one category, about half which are summary of content or genre. The overall distribution of categories is presented in Figure 1.

Figure 1. Categories of information included in Description fields

A summary of the content, the primary purpose of the Description field, only appears in 66.7% of descriptions, meaning over 30% do not summarize the content of the materials. Some of the most common categories observed in descriptions -- genre, names/roles of speakers, language or dialect, date of creation -- have dedicated elements in each archive's metadata scheme. Personal names of language consultants and sociolinguistic information (age, gender, native language, place of birth) appeared in 30.5% and 4.1% of descriptions respectively. The least common categories are value or uniqueness, audiences or uses, access rights, provenance. One of them -- access rights -- has a metadata element intended for this information in ELAR and PARADISEC's metadata schemes. Figure 2 compares distributions of descriptive categories among the three archives.

Figure 2. Comparative distribution of descriptive categories of information

Descriptions in ELAR most consistently summarize the content of the item, with this category observed in 86.2% of ELAR descriptions compared to less than 60% in PARADISEC and AILLA. ELAR also leads in representing the setting or context and names of those speaking in the recording. Interestingly, ELAR is the only archive with a metadata element designed for recording equipment used; still, this information is found in 8.5% of free-text Description fields.

Compared to AILLA and ELAR, PARADISEC's Description fields include administrative notes and notes from the tape or box of the item significantly more often. The most frequent category in AILLA Description fields is genre, as opposed to summary of content in ELAR and PARADISEC. Though a separate metadata field is dedicated to the description in the subject language (e.g., Curripaco, K'ichee') in AILLA, 3.1% of English-language Description fields include English text followed by a translation into the subject language. Unlike AILLA, translations included in ELAR or PARADISEC Descriptions translate content into a language of wider communication (e.g., Russian, Spanish, Portuguese).

5. Discussion

This exploration of free-text descriptions forms the basis for closer examinations of metadata quality in the future and may benefit managers of language archives in the expansion of metadata creation guidelines. Metadata guidelines provided by AILLA and ELAR leave the Description element open-ended: “Description of the resource” (AILLA, n.d.); “not only the description of the session but also information about its content” (ELAR, n.d.). The mention of the session (recording) in ELAR’s definition may explain the higher proportion of ELAR descriptions providing information about the setting and content of the recording. Examples of descriptive categories suggested in AILLA guidelines include genre, names of speakers, references to related items, and grammatical constructions in the material—categories seen prominently in AILLA descriptions. PARADISEC guidelines do not define the Item Description element, but give the example “Four text stories for interviews” (PARADISEC, n.d.), suggesting depositors should include a genre in their descriptions, and some indication of the extent. More detailed guidelines may improve consistency and usefulness of Description fields by guiding metadata creators to represent both the content and context of the archived material. As needs of language archive users are more thoroughly explored, relevant information to support user tasks may also be incorporated into metadata creation guidelines.

PARADISEC descriptions containing administrative notes likely come from legacy data where no information about the item was available beyond the writing found on the original analog material (e.g., ‘*Side A: 1. Kwento Bibidi-Comía*’), which may be difficult to interpret without context or extensive inference. When legacy data is not accompanied by sufficient information and the original owner cannot be contacted, archival staff have to create metadata to the best of their ability using the resources available. Administrative notes are messages from the metadata creator communicating that no metadata was provided for the item (e.g., ‘*no information available*’; ‘*[to be added]*’).

Burke and Zavalina (2020) report language archive managers’ concern over privacy restrictions as the General Data Protection Regulation (GDPR) develops, requiring suppression of personal information, despite the importance of the names of language consultants for language community users noted by Harris et al. (2019). These findings indicate that language archive managers may have to modify not only the Roles or Participants fields, but also the Description field to redact personal names and suppress personally-identifiable sociolinguistic information if necessary.

6. Conclusion and future work

This exploration of metadata in three large-scale language archives identified the most prevalent categories of information included in data values of the free-text Description fields using content analysis. Future research could compare the data values in specialized metadata fields to those in free-text Description fields for the most common categories of information observed by this study: genre, names/roles of speakers, language or dialect of the item, title, depositor or creator, and the date of creation—potentially, information from the Description field could be used to enrich the overall record. This study and related future research will inform recommendations to depositors and language archive managers, which may improve metadata quality in language archives and, in turn, the accessibility of archival data.

References

- Alemu, G. (2018). Metadata Enrichment for Digital Heritage: Users as Co-Creators. *International Information & Library Review*, 50(2), 142-156. <https://doi.org/10.1080/10572317.2018.1449426>
- Archive of the Indigenous Languages of Latin America. (n.d.) *Metadata*. <https://www.ailla.utexas.org/site/depositors/metadata>.
- Bruce, T.R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann, (Ed.) *Metadata in Practice*, American Library Association, Chicago. <http://hdl.handle.net/1813/7895>.
- Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1), 1-18. <https://doi.org/10.1515/ling-2017-0032>
- Burke, M., & Zavalina, O. L. (2019). Exploration of information organization in language archives. *Proceedings of the Association for Information Science and Technology* 56, 364-367.
- Burke, M., & Zavalina, O. L. (2020). Identifying challenges for information organization in language archives: Preliminary findings. In Sundqvist, A., Berget, G., Nolin, J., Skjerdingsstad, K. I. (Eds.), *Sustainable Digital Communities: 15th International Conference, iConference 2020, Borås, Sweden, March 23–26, 2020, Proceedings*, (pp. 622-629). Springer International Publishing.
- Endangered Language Archive. (n.d.). *Creating Metadata*. <https://www.soas.ac.uk/elar/helpsheets/file140536.pdf>.
- Harris, A., Gagau, S., Kell, J., Thieberger, N., & Ward, N. (2019). Making Meaning of Historical Papua New Guinea Recordings: Collaborations of Speaker Communities and the Archive. *International Journal of Digital Curation*, 14(1), 136-149).
- Henke, R., & Berez-Kroeker, A. (2016). A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation* 10, 411-457.
- Hillmann, D. I. (2008). Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, 46(1), 65-80. <https://doi.org/10.1080/01639370802183008>
- Hughes, B. (2004). Metadata quality evaluation: Experience from the Open Language Archives Community, In Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, & E. Lim (Eds.), *Digital Libraries: International Collaboration and Cross-Fertilization: Proceedings of 7th International Conference on Asian Digital Libraries, (Shanghai, China, 2004)*, (pp. 320–329). Springer-Verlag Berlin Heidelberg.
- Pacific Regional Archive for Digital Sources in Endangered Cultures. (n.d.) *Depositing material with PARADISEC*. <https://www.paradisec.org.au/deposit/>.
- Park, J. (2009). Metadata quality in digital repositories: A survey of the current

- state of the art. *Cataloging & Classification Quarterly*, 47(3-4), 213-228.
<https://doi.org/10.1080/01639370902737240>
- Roeschley A., Kim J., & Zavalina O.L. (2020). An exploration of contributor-created Description field in participatory archives. In Sundqvist, A., Berget, G., Nolin, J., Skjerdingsstad, K. I. (Eds.), *Sustainable Digital Communities: 15th International Conference, iConference 2020, Boras, Sweden, March 23–26, 2020, Proceedings*, (pp. 638-648). Springer International Publishing.
- Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society of Information Science*, 58, 1720-1733. <https://doi.org/10.1002/asi.20652>
- Tarver, H., Zavalina, O. L., & Phillips, M. (2016). An exploratory study of a Description field in the Digital Public Library of America. *Proceedings of the International Conference and Workshop on Dublin Core and Metadata Applications*. Copenhagen, Denmark, October 13-16, 2016.
- Zavalina, O. L. (2014). Complementarity in subject metadata in large scale digital libraries: A comparative analysis. *Cataloging & Classification Quarterly*, 52(1), 77-89.
- Zavalina, O. L. Palmer, C. L., Jackson, A. S., & Han, M. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata* 8(4), 263–292.