Identifying Challenges for Information Organization in Language Archives: Preliminary Findings[1*]

Mary Burke[1[0000-0002-6498-6820]] and Oksana L. Zavalina[1[0000-0002-3354-4923]]

[1] University of North Texas, Denton, TX 76201, USA
mary.burke@unt.edu, oksana.zavalina@unt.edu

Abstract.
Language archives are repositories of linguistic data about a selected set of languages, typically including recordings, transcripts, translations, and linguistic annotations. Digital accessibility of primary language data, particularly that of endangered languages, has long been recognized as necessary for research reproducibility, production of pedagogical materials, and typological discovery, though their potential currently lies dormant because these resources are rarely accessed by linguists or language communities. Reasons for the under-utilization of language archives include the lack of data standardization and decreased focus on metadata quality. The present work seeks to elucidate the issues facing language archive managers and users through two steps: content analysis of information organization in language archives, and semi-structured interviews with language archive managers and users. Primary challenges identified include lacking institutional support and a range of factors which impact authority control in language archives.

Keywords: language archives; information organization; metadata; digital repositories

1 Introduction

Language archives are repositories of linguistic data about a selected set of languages, typically including recordings, transcripts, translations, and linguistic annotations, essential for facilitating language preservation and revitalization and providing access to data on severely endangered languages (Henke & Berez-Kroeker, 2016). Digital accessibility of primary language data, particularly that of endangered languages, has long been recognized as necessary for research reproducibility, production of pedagogical materials, and typological discovery, though their potential currently lies dormant because these resources are rarely accessed by linguists or language communities. Reasons for the under-utilization of language archives include the lack of data standardization and decreased focus on metadata quality. To address this, the Open Language Archives Community (OLAC) was formed in the early 2000's to

[1*] This research has been funded by the US Institute for Museum and Library Services through a planning grant IMLS LG-87-18-0197. We also thank our project team members, Dr. Shobhana Chelliah and Mark Phillips, in addition to the members of our advisory board, Myung-Ja Khan, Drs. Christina Wasson, Gary Holton, Susan Smythe Kung, for help in identifying language archives for analysis and interpreting results.

develop protocols for language archiving and creating interoperable repositories for storing language data. Between OLAC and the Digital Endangered Languages and Musics Archive (DELAMAN), awareness of metadata has significantly increased in the linguistics community. Still, linguists are unsure if the widely used metadata schemes are appropriate for representation of language datasets and if these will ensure the use of their data. They often do not expect non-linguists (who normally make metadata-related decisions and create metadata in data repositories) to adequately represent language data for information retrieval. Need for detailed and accurate metadata to facilitate resource discovery, as well as ease of search in language archives is documented by user studies (Al Smadi et al., 2016; Wasson, Holton & Roth, 2016), informed by a user-centered design approach. Wasson, Holton and Roth (2016) identified five stakeholder groups central to language archives, acknowledging that individuals may belong to multiple groups: language communities, linguists, archivists, user-centered design practitioners, and representatives of funding agencies (p. 659). Representatives of each group gathered at a workshop and shared their perspectives, which were recorded, transcribed and coded using qualitative analysis software Dedoose. While Wasson, Holton, and Roth (2016) report significant insights of these user groups, still, the current knowledge of users' tasks and needs are insufficient to determine whether the present state of information organization is facilitating user tasks, whether additional metadata elements for representing language data, or methods of searching are required to increase the usability of language archives.

To understand more thoroughly the issues facing language archive managers and users, we have undertaken a project funded by the Institute of Museum and Library Services with two phases: Phase 1 included a content analysis of the websites of 20 digital language archives; Phase 2 involves semi-structured interviews with archive users and managers. Findings from Phase 1 are reported in section 2, and the preliminary results from Phase 2 in section 3. Section 4 concludes. The goal of the exploratory study presented in this paper was to identify:

- information organization tools and practices currently employed by language archives across the United States, and
- the needs of depositors and end-users (linguistics researchers, instructors, and students) for information organization functionalities in language archives.

The following information organization tools and practices were considered:

- What item-level and (if applicable) collection-level metadata scheme(s) are used?
- To what extent metadata records are displayed to end users?
- Does the archive allow self-depositing and, if so, are metadata creation guidelines and/or documentation of a metadata application profile used in language archives available?
- How is authority control implemented? What data value standards are used?
- What options for advanced search against indexed metadata fields are available?
- Are metadata records available for harvesting/download?
- What Semantic Web applications are available? For example, is metadata available as Linked Data?

1.1 Scoping Definitions

For the purpose of this project, we define language data as any audio, video or textual material representative of authentic language use. A list of words, a speaker explaining how to cook a particular variety of rice, or a traditional song are just a few examples of the various forms language data can take. Language data should be accompanied by metadata describing its relevance and context (minimally, how and where it was collected, name of speaker, and a simple description of the content).

We define a language archive as containing at least one collection (two or more items) of language data. Crucially, language archives do not need to identify as such to be considered in this analysis. For example, many university repositories contain collections of language data (e.g., University of North Texas Digital Library's Lamkang Language Resource, Indiana University at Bloomington's Ethnomusicology Multimedia Materials Collection), but identify more broadly as digital libraries or university repositories. Language archives have materials available for public access with the intent of long-term preservation. For example, Kaipuleohone, housed at the University of Hawai'i at Mānoa (http://ling.hawaii.edu/kaipuleohone-language-archive/), exemplifies a language archive because it contains individual items, makes items available for public access, has structured metadata, and accept deposits of material (or has accepted in the past).

Resources not considered in this analysis, though in no way dismissed, are corpora and resource aggregators. The Language Data Consortium (LDC), a collection of corpora, does not organize corpora into collections, have structured metadata, nor provide public access to its content. Resource aggregators, such as the Karuk Archives (http://karuk.org/) are also excluded from this definition. Such websites provide links to resources available on other websites and archives, but do not house language data itself, nor do they provide structured metadata. So, while these platforms may store language data, they are not considered a language archive for the purpose of this investigation.

2 Phase 1: Content Analysis

In the first phase of our research project, we conducted an exploratory content analysis of websites of language archives to identify the information organization practices currently employed by language archives in the United States.

To answer these questions, we examined a total of 20 language archives, including 16 within the United States and 4 outside the United States for comparison (2 in Europe, 1 in Australia, 1 in Canada). Of the 16 US archives, 6 are collections housed in university libraries, 4 are standalone archives not affiliated with a university, and 6 are archives associated with universities. Burke and Zavalina (2019) provide a full report of Phase 1 findings. Phase 2, discussed below, focused on the decision-making process for these information organization decisions and on the gaps between user needs and the way information is organized in language archives. This preliminary report partially addresses the overall findings of the study.

2.1 **Phase 1 Findings**

The findings of Phase 1 displayed a variety of information organization strategies in language archives.

Metadata Schemes Used. Though language archives make some information available on an item landing page, it is often unclear which metadata schemes the elements belong to. It seems that many language archives are using locally developed schemes to suit their needs. Often, local metadata application profiles based on Dublin Core are constructed for use in data repository. One such example, Kaipuleohone, makes two versions of each record available, a simple and full record. 'Simple item records' can be expanded by selecting 'Show full item record.' The full record displays the namespace 'dc' along with qualifiers that clearly indicate the use of the Qualified Dublin Core metadata scheme. Records from other archives typically contain many of the same elements (e.g., Identifier, Title, Contributor/ Depositor/ Creator, Language, Date, Description, Format, Notes) but do not include any explicit indications of the metadata scheme being used. Only 2 of the 20 archives, both collections within a university digital libraries, have metadata records available for download in RDF (i.e., University of North Texas Digital Library and University of British Columbia Open Collections).

Controlled Vocabularies. Use of controlled vocabularies is not consistent across language archives. ISO 639-2 Language Codes, DCMI Type Vocabulary, and Traditional Knowledge Labels are frequently used. Collections within digital libraries include Library of Congress Subject Headings, while other archives designed for language data do not include a Subject element. Dates are typically encoded using W3CDTF. Because most language archives make records available only on the object's landing page, it is unknown whether additional metadata, including more extensive use of controlled vocabularies, exists on the back end.

Advanced Search Capabilities. 18 of the 20 archives have advanced search capabilities, though not all fields are indexed for advanced search. 10 archives have only a subset of fields available for advanced search; minimally, Title, Author, and Language fields.

Metadata Creation Guidelines. Self-depositing is not available in most cases; only ELAR, AILLA, the Tromsø Repository of Language and Linguistics (TROLLing), and Language Commons allow users to upload data without an intermediary. Few of the university repositories analyzed have information on making deposits. In these cases, it is unclear whether they are not open to deposits, or whether a potential depositor would have to contact the university library directly to discuss depositing. Similar to the use of controlled vocabularies, the availability and level of detail of metadata creation guidelines vary widely. While some language archives (e.g., PARADESIC, AILLA, Kaipuleohone) provide detailed guidelines including examples and tutorials, others require only a title and a description of the deposit.

2.2 Phase 1 Conclusions

Through content analysis alone, many of our research questions remain unanswered, or only partially answered. For example, we were able to identify that there are no records available for download in RDF on most archive's websites; however, this does not preclude the possibility that such RDF records exist, but are inaccessible to users. Similarly, because full metadata records are not always exposed, there may be controlled vocabularies in place that are not shown in the user interface. Further, content analysis did not indicate whether the archive manager(s) plan to implement any such features, and what factors might impact those choices. To gain a more detailed view of the information organization strategies employed in language archives and user needs, we interviewed these groups in Phase 2.

3 Phase 2: Semi-structured Interviews

3.1 Methodology

In Phase 2, we assess the needs of depositors and various stakeholders for information organization functionalities in these archives via semi-structured interviews. Stakeholders include:

- Linguistics researchers depositing their datasets in language archives and using or planning to use language archives in their research
- Language and linguistics educators using or planning to use language archives for teaching (K12-higher education)
- Students who would benefit from using language archives in their studies (linguistics students and information science students)
- Language community members interested in heritage language materials
- Language archiving practitioners and managers.

The semi-structured interviews were conducted with the purpose of language archive analysis and user needs assessment, as well as to collect information on how these requirements are met by information organization in language archives based on the previous experiences of respondents in using language archives. In addition to the users of archives, we interviewed archivists to learn about the use of metadata schemes and controlled vocabularies in these archives. Participants who already use language archives were also observed depositing data and/or searching and browsing language archives and interacting with metadata in the process. Observations represent the evaluation of information organization techniques in a selection of the language archives examined in Phase 1.

Participants were recruited from major language archives and from language documentation scholarly communication networks (e.g., the Linguistics Society of America's Committee on Endangered Languages and their Preservation (CELP) blog, the LinguistList). To date, we have interviewed and observed 7 archivists and 9 users of language archives. Of the 9 users, 4 had deposited data to a language archive. The present work reports our preliminary findings from the data collection and early data analysis stage of the project.

Interviews and observations were conducted over the Zoom video-conferencing tool. Zoom provides an automated transcription which was then manually corrected. Transcripts were coded according to the recurrent themes in the data using NVivo 12 Plus. The code list was developed based on the interview guide, and was later refined according to topics which emerged from the transcripts and consultation with project team members. Because the data analysis is ongoing, the code list may be further modified as necessary.[2]

3.2 **Preliminary Findings**

With data analysis still ongoing, the final results from the content analysis of interview transcripts is not yet prepared; rather, we share the compelling themes emerging from interviews.

Lacking Resources. Many archive managers identified a lack of funding as a primary barrier they face to maintaining their archive or implementing any new functionalities. With budget cuts to universities throughout the United States, institutions must rely on funding from federal grants to stay afloat. Full-time staff members are vital to the maintenance of the archive, but there is insufficient funding to support them, as short-term funding is not guaranteed renewal. One archive manager remarked "an archive is an institution, and it should be funded at the hundred-year level, not at the one-year level" (arch_5, 2019).

In 2011, the National Science Foundation (NSF) mandated the inclusion of a Data Management Plan (DMP) for all projects to increase transparency of research; this affects language archives directly because one of the most common ways language documentation projects are funded is through NSF's Documenting Endangered Languages (DEL) program. Specifications for the DMP in DEL projects require a Letter of Collaboration from the institution where the data will be archived, confirming their agreement and ability to ingest the material resulting from the project (National Science Foundation, 2018). Another major source of funding for language documentation projects, the Endangered Languages Documentation Program (ELDP) requires that the resulting data be archives in the Endangered Languages Archive (ELAR). This has increased substantially the amount of materials language archives receive each year, creating additional strain on these under-funded institutions.

This issue is reflected on the user end during the depositing process—in some cases, researchers utilize student assistants in order to manage the often cumbersome and tedious work of archiving their primary language data. While the DMP has caused a prioritization of archiving, there has not been a corresponding increase in funding to support this additional task not previously emphasized in language documentation projects, and both depositors and archivists alike are striving to meet the shifting demands.

Under-utilization of Authority Control. Interviews reveal that language archives do not implement authority control to the extent typical in traditional library settings. This

[2] Data collection instruments and the current code list can be found at https://docs.google.com/document/d/1EGn_34QyQlpV1EcIj1wuevbvsorbe4UAA9uqw1vLfco/edit?usp=sharing

is especially true for proper names of people. In some cases, archives maintain local name authority files for the depositors in their archive, but do not link this to a higher authority (e.g., the Library of Congress Name Authority File). The use of authority control is connected to the lack of funding; for example, one archive management team explained that they do have a locally developed name authority file, but have not applied it to names in every record because staff must address basic access issues before they can devote time to standardizing data values across thousands of records.

Subject headings are not typically assigned to language data; rather, the language of the data (typically referred to as 'subject language') seems to be of primary importance to users. The long-standing struggle with languages having multiple names has popularized the use of the ISO codes for language names. Still, the ISO 639-3 does not have a value for all existing languages, further complicating the issue of authority control in language archives.

Though this study focuses on language archives based in the United States, there are global factors affecting information organization, namely the General Data Protection Regulation (GDPR) in place throughout the EU. The GDPR is tightening restrictions on personal data use, even if that data is archived outside of the EU, so language archives in the United States must develop solutions to these restrictions. One archive manager noted the effect GDPR is having on their decision to include the names and years of birth of individuals featured in archived recordings, despite the potential value that information, particularly the year of birth, would add for researchers using the archive. It is expected that, in coming years, GDPR and similar legislation will affect information organization in language archives, most notably name authority control.

User Preference for Social Media. Finally, linguists depositing to language archives may feel their archival deposit is not being utilized by other researchers, and especially not by the language communities they work with. Though researchers understand and appreciate the purpose of archiving for long-term preservation of the materials, many express frustration with the inadequate access to materials archives provide. Depositors interviewed compared the depositing process with major language archives to the ease of uploading a video to YouTube or Instagram: [on social media,] "the transaction costs for, let's say, the depositor are extremely low, and yet, the findability and the accessibility is extremely high. So somehow we're at the opposite end in linguistics where, like the burden on the depositor is extremely high, and yet, the findability and accessibility is extremely low" (user_4, 2019). Indigenous language communities, another primary user group of language archives, often rely on mobile devices, making streaming significantly more feasible than downloading large files, which creates a preference for a social media-type user interface over the archival access point.

4 Conclusions

This project is the first step in a series of research and demonstration projects aimed at improving the information organization in language archives throughout the United States. Initial analysis of the interviews reveal several elements of the language archiving process which are problematic for users, depositors, and archivists, including lacking institutional support and a range of factors impacting authority control. These

preliminary results are promising; we are confident that this project will yield fruitful data for developing strategies to improve the usability of language archives, and help bridge the gap between what users need and what archivists are able to provide.

References

1. Henke, R., Berez-Kroeker, A. A brief history of archiving in language documentation, with an annotated bibliography. Language Documentation and Conservation **10**, 411-457 (2016)
2. Al Smadi, D. Barnes, S., Blair,. M., Chong, M., Cole-Jett, R., Davis, A., Hardisty, S., Hooker, J., Jackson, C., Kennedy, T., Klein, J., LeMay, B., Medina, M., Saintonge, K., Vu, A., Wasson, C. Exploratory user research for CoRSAL: report prepared for S. Chelliah, Director of the Computational Resource for South Asian Languages. University of North Texas. Department of Anthropology. (2016)
3. Wasson, C., Holton, G., Roth, H. Bringing user-centered design to the field of language archives. Language Documentation and Conservation **10**, 641-671 (2016)
4. Burke, M., Zavalina, O. L., Exploration of Information Organization in Language Archives. Proceedings of the Association for Information Science and Technology **56**, 364-367 (2019)
5. National Science Foundation Documenting Endangered Languages (DEL) program solicitation, https://www.nsf.gov/pubs/2018/nsf18580/nsf18580.htm. Last accessed 13 September 2019.