Exploration of Information Organization in Language Archives

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

Oksana Zavalina
University of North Texas
Denton, TX, USA
oksana.zavalina@unt.edu

ABSTRACT

This submission reports preliminary results of the first stage of a research project that seeks to identify the information organization methods and techniques currently offered in the language data archives and the gaps between the tools and techniques available and the needs of actual and potential users of language data archives. We conducted an exploratory content analysis of the websites of the language archives (LAs) hosted by the institutions located in the United States and several other countries. The focus of our exploratory content analysis is on the information organization, including documentation on metadata standards, displaying of individual metadata records and availability of harvesting sets of metadata records, provision of advanced discovery and navigation options powered by metadata such as availability of adaptive and personalized search or social tagging functionality. Here, we report our preliminary findings and describe our plan for Stage 2 of the project.

KEYWORDS

language archives; information organization; metadata; Linked Data

ASIS&T THESAURUS

Digital repositories; Archives

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19 – 23 October, 2019

INTRODUCTION

Rich and unique digital language datasets have a potential to make a strong contribution in social science research and education (e.g., Language Science, Geography, History, Sociology) and Computer Science (e.g., natural language processing). However, this potential currently remains largely unrealized as the language data available in language archives are rarely accessed by linguists or indigenous language communities. One main reason for this lost opportunity is the confusing and cumbersome design of language archives [Wasson, Holton & Ross, 2016; Wasson et al., 2018]. For example, as depositors upload information with various levels of granularity, retrieval for educational or research purposes becomes untenable without much additional resorting and organization of the archived data. Also, users of language archives cannot easily compare data across languages. Many theoretical breakthroughs in historical linguistics, syntax, phonology, and other areas would become possible if users could query archival data for cross-linguistic patterns. Currently, language archives place data on each language in a separate collection and do not function as databases [Al Smadi et al., 2016].

Current State of Language Archiving

Online language archives are a valuable tool to support language preservation and revitalization, and to providing data on lesser-known languages valuable for linguistic analysis [Henke & Berez-Kroeker, 2016].

Like most digital preservation activities in various knowledge domains, language data archiving initiatives started in late 1990s and early 2000s. To bring this rich language data together to facilitate access to it, an Open Language Archives Community (OLAC) was created, an international collaboration project sponsored by the US National Science Foundation in 2000-2010 and hosted by the University of Pennsylvania Libraries. OLAC's goal was to develop protocol for archiving language data and to create interoperable repositories for storing and making language data more widely accessible. OLAC put together the combined catalog of all resources from 60 participating language archives located throughout the world: in USA, Australia, Brazil, India, New Zealand, Taiwan, and several European countries. OLAC's combined catalog contains over 300,000 records, covering resources in many languages of the world.

In the 2000s, the language archiving community has been building awareness of metadata and working on designing metadata standards through projects such as Electronic Metastructure for Endangered Language Data (E-MELD, http://emeld.org/) which developed a metadata editor tool, a linguistic ontology, and a linguistic metadata standard (http://www.language-archives.org/OLAC/metadata.html) based on Dublin Core. Digital Endangered Languages and Musics Archive (DELAMAN) is another organization working on standards, including metadata-related, for language data archives.

Language archiving research [e.g., Nathan & Austin, 2004] distinguishes between two kinds of metadata important for

information organization in language archives: so called "thick" metadata that represents text encoding of the linguistic documents itself: transcriptions, commentary, and time-aligned annotations (e.g., TEI), and "thin" metadata that facilitates research discovery (e.g., Dublin Core, OLAC, MARC 21, etc.). In addition to item-level descriptions of individual objects in archives, collection-level metadata that describes the entire collection is important, with major standard examples including Encoded Archival Description (EAD), Dublin Core Collection Description Application Profile, etc.

In the past, language archives were primarily designed for and by linguists, described as a "one-way" model by Nathan [2014, p. 193], with limited communication between archivists and users. The current foci in language archiving are "expanding audiences for archives and breaking traditional boundaries between depositors, users, and archivists" [Henke & Berez-Kroeker, 2016, p. 412]. The emerging adoption of participatory archives and self-depositing practices aims to rectify the previous lack of communication. Self-depositing is a key feature setting apart Endangered Languages Archive (ELAR) and Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADESIC), two of the largest language archives.

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 examples of the various forms language data can take on. 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, or 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 are platforms where language data is stored, they not considered a language archive for the purpose of this investigation.

Language Archive Analysis

In the first stage of our research project, we conducted exploratory content analysis of websites of language archives, with the goal to identify the information organization tools and practices currently employed by language data archives across the United States. We sought answers to the following specific questions:

- 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 LAs 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?

To investigate these aspects of information organization, a total of 20 language archives were considered, 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.

Findings

Preliminary findings from Stage 1 show a variety of information organization strategies in language archives.

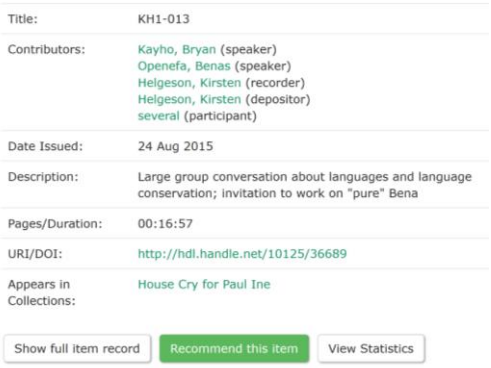
Metadata Schemes Used

Though language archives make some information available on item landing pages, it is unclear which metadata schemes the elements belong to. See Figures 1-3 for examples of the variation in metadata records. It seems that many language archives are using locally developed schemes to suit their needs.

Identifier:	ANLC3234
Title	SLI De Reuse #3
Description:	Grammatical comments on text on singing. The morning of 4/12. Slow repeat of 2nd episode of 4/12
Comments	#3 WdR 3a+b. Tape is located in Annex. Tape 3 of 35.
Contributors	de Reuse, Willem Joseph (interviewer)
Date	1985-04-05
Type	Sound
Subject Language(s)	Siberian Yupik
Collection	SLI De Reuse

Figure 1. Metadata record, Alaska Native Language Archive (ANLA)

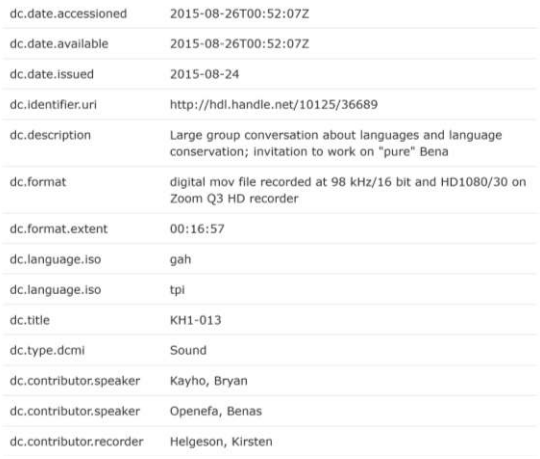
Another language archive, Kaipuleohone, makes two versions of each record available, a simple and full record, shown in Figures 2 and 3. 'Simple item records' can be expanded by selecting 'Show full item record.' The full record seen in Figure 3 includes the namespace 'dc' and qualifiers that clearly indicate the use of the Qualified Dublin Core metadata scheme. The ANLA record (Figure 1) contains many of the same elements, but does not include any other indications of the metadata scheme being used.



Title:	KH1-013
Contributors:	Kayho, Bryan (speaker) Openefa, Benas (speaker) Helgeson, Kirsten (recorder) Helgeson, Kirsten (depositor) several (participant)
Date Issued:	24 Aug 2015
Description:	Large group conversation about languages and language conservation; invitation to work on "pure" Bena
Pages/Duration:	00:16:57
URI/DOI:	http://hdl.handle.net/10125/36689
Appears in Collections:	House Cry for Paul Ine

Show full item record | Recommend this item | View Statistics

Figure 2. Simple metadata record, Kaipuleohone



dc.date.accessioned	2015-08-26T00:52:07Z
dc.date.available	2015-08-26T00:52:07Z
dc.date.issued	2015-08-24
dc.identifier.uri	http://hdl.handle.net/10125/36689
dc.description	Large group conversation about languages and language conservation; invitation to work on "pure" Bena
dc.format	digital mov file recorded at 98 kHz/16 bit and HD1080/30 on Zoom Q3 HD recorder
dc.format.extent	00:16:57
dc.language.iso	gah
dc.language.iso	tpi
dc.title	KH1-013
dc.type.dcmi	Sound
dc.contributor.speaker	Kayho, Bryan
dc.contributor.speaker	Openefa, Benas
dc.contributor.recorder	Helgeson, Kirsten

Figure 3. Full metadata record, Kaipuleohone

Most records include the following metadata elements: Identifier of some kind (local, doi), Title, Contributor/ Depositor/ Creator, Language (ISO code, plain text, or both), Date (deposited, created, or both), Description, Format, Notes, Rights, and Related items.

Availability of Records in Various Formats

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). Few of the archives considered here had metadata records exposed using identifiable schemes.

Controlled Vocabularies

The use of controlled vocabularies is not consistent across language archives. Those used most frequently include ISO 639-2 Language Codes, DCMI Type Vocabulary, and Traditional Knowledge Labels. Language archives existing as collections within a digital library include Library of Congress Subject Headings, while other language archives 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 (see Figure 1), it is unclear whether additional metadata, including further use of controlled vocabularies, exists in a format inaccessible to users.

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. None of the archives analyzed allow for personalized or adaptive searches, though PARADESIC and the Historical Books Collection at the University of British Columbia permit users to leave comments on items which are viewable to other users.

Metadata Creation Guidelines

Self-depositing is not allowed in most cases; only ELAR, the Tromsø Repository of Language and Linguistics (TROLLing), and Language Commons allow users to upload data without an intermediary. (The Archive of the Indigenous Languages of Latin America (AILLA) is developing this functionality now, but has not yet released it.) Few of the university repositories analyzed have information on depositing material. 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 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, others require only a title and description.

STAGE 2 PLANS

In stage 2, we plan to assess the needs of depositors and various stakeholders for information organization functionalities in these archives. Stakeholders include:

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

The team will conduct interviews and observations of a small selected sample of members of each user group. The semi-structured interviews will be conducted with the purpose of language data archive requirements analysis or 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 experience of respondents in using language archives. In addition to the users of these archives, we plan to interview archivists to learn about the use of metadata schemes and controlled vocabularies in these archives. Participants who already use language archives will also be observed by the project team:

- Depositing participants will be observed as they interact with information organization tools (including metadata) in language archives in the process of depositing.

- Participants who do not deposit themselves but use materials deposited by others in language archives will be observed searching and browsing language archives and interacting with metadata in the process.

Observations will represent the heuristic evaluation of information organization in a selection of language archives.

CONCLUSION

This project is the first step in a series of research and demonstration projects aimed at improving the information organization in language data archives around the country. The ability for libraries, archives, and museums to identify the most meaningful for users ways of information organization in language archives is an important first step in fully realizing the potential of language archives and digital libraries nationwide. As a result of this project, we will provide empirical data in support of planning the future large-scale collaborative project focused on development of more efficient and user-friendly tools for access to digital language archives.

ACKNOWLEDGMENTS

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 for help in identifying language archives for analysis and interpreting results.

REFERENCES

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., and Wasson, C. (2016). 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.

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.

Nathan, D. & Austin, P.K. (2004). Reconceiving metadata: Language documentation through thick and thin. In Language Documentation and Description, edited by Peter K. Austin, 179-187. London: SOAS.

Nathan, D. (2014). Access and accessibility at ELAR, an archive for endangered languages documentation. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description*, Volume 12: Special Issue on Language Documentation and Archiving, 187–208. London: SOAS.

Wasson, C., Holton, G., & Ross, H. (2016). Bringing user-centered design to the field of language archives. *Language Documentation and Conservation*, 10, 641-671.

Wasson, C., Medina, M., Chong, M., Le-May, B., Nalin, E., & Saintonge, K. (2018). Designing for diverse user groups: Case study of a language archive. *Journal of Business Anthropology*, 7(2).

NOTE: this document has been adapted from the ACM SIGCHI conference template for use with ASIS&T conference.