Exploring Language Archiving Education for Information Professionals: Interdisciplinary Collaboration to Support Information Access

Oksana L. Zavalina[a] and Shobhana L. Chelliah[b]

[a]Department of Information Science, University of North Texas, USA

[b]Department of Linguistics, University of North Texas, USA

Oksana.Zavalina@unt.edu, Shobhana.Chelliah@unt.edu

ABSTRACT

Preservation and revitalization of Indigenous and endangered languages supports a resilient future. Funding agencies have extensively supported efforts aimed at preserving and providing online access to unique and valuable collections of language data. However, a gap exists between the way language data is organized and represented in digital archives (mostly by the LIS professionals) and understanding of that data – and how it should be organized and represented – by language preservation and revitalization researchers, members of language communities. The specifics of information objects collected by language archives and information needs of these collections' end-users are not currently examined in the LIS education. This paper presents the work of the interdisciplinary team of educators, researchers, and practitioners to address this curricular gap, and discusses lessons learned and future directions.

ALISE RESEARCH TAXONOMY TOPICS

digital humanities; community engagement; metadata; data curation.

AUTHOR KEYWORDS

curation language archives; community language preservation and access; graduate education; interdisciplinary experiential learning.

INTRODUCTION AND BACKGROUND

People increasingly use digital technologies to record language practices with the goal of supporting language vitality. Resulting language collections bring individuals and communities (e.g., Native Americans, refugees) a sense of belonging and positive identity and allow to tell their own story and to have control over how their materials are framed or shared. However, research demonstrates that these materials are often not curated and made available to users in a

functional way (e.g., Al Smadi et al., 2016; Wasson, Holton & Ross, 2016). To date, there is little academic instruction to library professionals on how to support curation and archiving of community language revitalization collections, especially factoring in the lack of digital access and digital literacy for many Indigenous communities. Communities as well do not have access to training on how to use existing library and archival protocols to create, archive, curate, and disseminate their materials of intangible linguistic heritage, many of which are in imminent danger of loss (Hinton & Pérez-Báez, 2018). Crafting a resilient future requires addressing this need.

Once a rarity, language archives are proliferating as standalone repositories, language-focused collections housed in museums, academic libraries, or tribal libraries. An Open Language Archives Community (OLAC), an NSF-funded 2000-2010 international collaboration project aimed to facilitate access to this rich language data. OLAC put together the combined catalog of over 60 language archives, developed best practices for language archiving, facilitated interoperability of language data repositories, and provided training to the linguists. Language archives deposits rose exponentially after the 2011 introduction of the NSF (2018) data management plan guidelines which require federally funded projects documenting endangered languages to make their outcomes accessible to other researchers and language community members by depositing data into language archives.

Studies report that Indigenous communities in several countries, including Australia, Canada, New Zealand, Thailand, are actively involved in developing digital archives that provide access to language data and other materials and contribute to revising archival descriptions in existing archives as part of archival decolonization process (e.g., Frederick, 2019; McKemmish, Chandler, & Faulkhead, 2019; Tasker & Liew, 2020; Ungsitipoonporn et al., 2021). In participatory archives, "community members contribute knowledge or resources, resulting in increased understanding about archival materials" (Thiemer, 2011). Examinations of metadata in some of these repositories find that community-created descriptions provide extremely rich context for archived materials and facilitate discovery (Burke & Zavalina, 2020; Roeschley, Kim, & Zavalina, 2020). Community-language-focused digital archive collections typically include common materials – photographs, texts, and recordings (e.g., personal narratives, traditional stories, etc.) – and domain-specific item types such as wordlists, language transcriptions, translations, word-by-word language analysis, etc.

Some recently-started non-LIS education initiatives aim to support resilient future through serving language preservation. Content on language archiving, data management techniques, and metadata is covered in some Field Methods and Tools linguistics courses (e.g., Berez, 2015). At Tribhuvan University of Nepal and Southern Illinois University Edwardsville, linguistics and some other students receive training in creating digital humanities exhibits, get hands-on experience creating metadata and annotations for language data (Hildebrandt et al., 2019). Also, Open Access training has recently been developed by some language archives. One example is the Archiving for the Future online course for depositors-linguists that provides a background in digital curation (Kung et al., 2020). Similarly, the Collaborative Language Archiving Curriculum (2020) aims to guide language communities through the archiving process. The Training and Resources for Indigenous Community Linguists program aims to

connect linguists and language communities to support each other in language documentation activities and archiving (Centre for Cultural-Linguistic Diversity Eastern Himalaya, n.d.).

LIS curriculum has not yet kept in pace with these developments. Professional librarians' input would be beneficial to organizing in language archives (Burke et al., 2021), yet information professionals graduating from LIS programs currently lack knowledge on language archive user needs, attributes of language data, as well as specifics of information organization, metadata quality assurance, and user-centered design for language archives. There is a clear need for diversifying LIS education by incorporating these topics. There is also a need for the LIS field to focus more on rights, ownership, archiving spaces, and mutually beneficial relations between communities and partnering institutions, identify and disseminate best practices, and develop collaborations with linguists to support community archiving (Chelliah, 2021). Our interdisciplinary team that includes LIS and Linguistics educators and researchers, library and digital archive practitioners have started exploring the ways to address these needs.

ACTIVITIES, LESSONS LEARNED, AND NEXT STEPS

In 2019, aiming to bridge the knowledge gap between linguists and information professionals by developing a common ground and shared terminology, a team of LIS and Linguistics faculty and practitioners at the University of North Texas developed an interdisciplinary graduate course on topics of information organization in digital archiving of language data. This work was informed by our research project findings and practical experience related to language archives. The unified syllabus for combined class, the learning modules and assignments, presentations by instructors and guest speakers were designed with the idea to offer the middle ground between skills and interests of two student groups and to provide students with opportunities to learn together and from each other.

In Spring 2020, this combined course was offered to 19 graduate students. Twelve LIS students were enrolled in the synchronous online advanced course INFO5224 Metadata II; they had successfully completed the introductory metadata course in the past. Seven Linguistics students were taking the face-to-face course LING5030 focusing on South Asian languages (SAL). The course aimed to address seven learning outcomes:

1. Develop understanding of language data formats, collections, and archives with reference to South Asian languages (SAL)
2. Be able to discuss relevant major typological features of SAL.
3. For one SAL, gain knowledge of language structure – phonology, morphology, and syntax – based on the datasets curated in digital language archives.
4. Understand important issues and current trends in metadata theory and practice in relation to language data: creation, documentation, and management of metadata; metadata quality and interoperability; metadata as Linked Data; existing technologies and applications; metadata use in information retrieval.
5. Create high-quality item-level metadata for SAL data
6. Create collection-level metadata for collections of SAL data
7. Evaluate metadata quality in a SAL digital collection and develop metadata creation guidelines for SAL data collection(s).

A total of fifteen 150-minute-long class meetings were held, which included topic presentations, examples, demonstrations, and assignment walk-throughs. Both instructors led class meetings together – with content equally split between the two – in the physical classroom with simultaneous Zoom meeting. Slide sets used by instructors and guest speakers, recordings of each Zoom meeting, and automatically-generated transcripts were made available for student review through the combined course website. During the 3-week introductory period, Linguistics students were introduced to the basics of information organization principles. At the same time, LIS students reviewed key content from the previous relevant coursework, narrowed down with language-specific examples and learned the basics of linguistics terminology and general information about SAL that would be necessary for understanding materials in the language collections later in the semester. Linguistics students were also holding separate reading groups to cover more in-depth linguistic content (to address learning outcome 3). The first, fourth, fifth, sixth, and seventh student learning outcomes were assessed through written assignments: one individual exercise, followed by three practical group assignments which contributed to the semester-long real-life project. In that project, interdisciplinary student teams developed common understanding and vocabulary, build skills, and applied these in practice to created and organize the digital archive for Manipur language materials accessible via UNT digital library.[1]

The lessons learned in this experiment were of two kinds:

- Those related to domain differences between LIS and Linguistics terminology. Such differences highlight limitations of some of the traditional library tools when it comes to providing adequate access to language archive materials based on the needs of typical users of these materials
- Those related to instructional design.

One example of the first kind of lessons learned is related to vocabulary control. There is a very limited set of domain-specific controlled vocabularies for representing language data. In addition to Glottolog and AUSTLANG language code lists that complement ISO 639-3 standard, it includes four small-scale and less known OLAC vocabularies: for linguistic subject (29 terms), role (24 terms), discourse type (10 terms), and linguistic data type (3 terms). For that reason, metadata creators rely on often much more extensive controlled vocabularies widely used in the libraries, archives, and museums. Those vocabularies often use the same terms but in the much broader or sometimes quite different meaning. For instance, the term Analyst that in the MARC Code List for Relators – used in language archives hosted by academic libraries – is defined as "a person or organization that reviews, examines, and interprets data or information in a specific area" is not represented in OLAC role vocabulary. However, this term is commonly understood by documentary linguists who collect and deposit materials into language archives as referring to a person or a group that specifically provided linguistic analysis of language data. As a result, the information in metadata records is interpreted differently by information professionals and users of language collections, which highlights the need for collaboration and development of common understanding of terminology, and extensions to existing OLAC controlled vocabularies.

[1] https://digital.library.unt.edu/explore/collections/MDR/

Student feedback offered lessons learned in relation to instructional design. Some examples of comments on the strengths of the combined course included:

- "*Interacting with faculty and students from a different program is good preparation for real-life work situations in which we will need to work with people from various backgrounds, and with different interests and goals.*"
- "*Learning the detail required and level of organization needed to accurately archive data. I will be a better linguist as a result of this course*"
- "*This course helped me a lot in expanding my understanding on the subject*"
- "*It was good to understand that we may encounter subjects that we need to understand to provide good quality metadata in the future.*"

Some of the weaknesses that students identified were technological in nature and can be relatively easily overcome in offerings of a combined class like this. For example, although online students had positive experience with instructors' and guest speakers' live presentations, they often could not hear what their fellow students in the physical classroom were saying during live discussion. That problem was resolved when all class meetings were moved online due to COVID-19 pandemic. Other weaknesses pointed out by students were more substantial and less easily addressable as they related to perceived balance of the course content. Interestingly, more than one student on each side (LIS and Linguistics) felt that their discipline's content was overshadowed by the other's:

- "*Understanding the linguistics side seemed to detract from the metadata creation*."
- "*I spent so much time trying to understand the linguistics portion of the class that I feel I neglected the metadata side*."
- "*Hyperfocusing on information science over linguistics information*"
- "*Courses on different themes [are] better taught separately. If students are not equally interested in both courses, it is a hindrance in learning.*"

Student feedback on this experimental course offering has been used to shape curriculum development. We are working on an improved version of a combined course for LIS and Linguistics students with the following changes: ensuring that course assignments cover each of the student learning outcomes, assessing the best-fit course pairs (e.g., an advanced metadata course paired with the field methods or tools course in the language documentation track), selecting methods, procedures, and approaches that better overlap with interests of both audiences, etc. At the same time, our team is experimenting with the modular curriculum approach. We developed a language archives learning module implemented in the Spring 2021 INFO5224 advanced metadata course. The first learning module was designed as a case study of a user community served by information professionals, with the focus on the user needs of language speakers (including Indigenous communities) and linguists, unique attributes of information objects collected in language archives, use of general and domain-specific metadata schemes and controlled vocabularies to represent language archive materials, etc. The other three learning modules were also revised to provide LIS students more opportunities to interact with language archives through evaluation of metadata quality, etc. The student survey revealed substantially higher satisfaction levels than those of LIS students in a Spring 2020 combined INFO5224/LING5030 course: 67% satisfaction score increase for the course content measure

and 56% increase for overall effectiveness.[2] Spring 2021 student survey also included questions about four learning modules. Each module, including the one focused on language archives, received a high satisfaction score of 4.9 out of 5. Student feedback on the course redesign was positive: e.g., *"Presentations about metadata for linguistic user communities were both very informative and interesting, […] opened my eyes to the interdisciplinary nature of the metadata profession."*

In the future, we plan to expand collaboration with LIS and Linguistics education experts and develop the flexible modular curriculum with a strong practical component and a focus on community language archiving. We envision that future curriculum to include the following modules: (1) Language revitalization and language and culture endangerment, (2) Developing and managing a community language archive, (3) Digital content management and metadata for community collections, (4) Preservation and access for community collections, and (5) Dissemination and use of community collections. These individual modules will be integrated in LIS courses in the areas of metadata, digital libraries, data curation, digital humanities, and digital imaging. Individual modules would also be implemented in Linguistics courses such as field methods, tools for language documentation, endangered languages, research methods for Linguistics; and in language archiving training workshops designed for language community members. The modules can also be taught together as part of a specialized LIS course on language archiving. We believe these and similar education initiatives will contribute to crafting resilient future by helping address the need for providing language archiving training to future information professionals and offering training on best practices for linguists and language communities on standards and techniques in preserving valuable language content, in creating and maintaining digital language archives.

REFERENCES

Al Smadi, D. et al. (2016). Exploratory user research for CoRSAL [language archive]: report prepared for the Computational Resource for South Asian Languages. University of North Texas.

Berez, A. L. (2015). Reproducible research in descriptive linguistics: Integrating archiving and citation into the postgraduate curriculum at the University of Hawai ʻi at Mānoa. In A. Harris, L. Barwick, & N. Thieberger (Eds.), *Research, Records and Responsibility: Ten Years of PARADISEC* (pp. 39-51). University of Sydney Press.

Burke, M., & Zavalina, O. L. (2020). Descriptive richness of free-text metadata: a comparative analysis of three language archives. *Proceedings of the Association for Information Science and Technology*, 57(1), e429.

Burke, M., Zavalina, O.L., Phillips, M., & Chelliah, S. (2021). Organization of knowledge and information in digital archives of language materials. *Journal of Library Metadata*, 21(2), 1-30.

Centre for Cultural-Linguistic Diversity Eastern Himalaya. (n.d.) *TRICL*. Retrieved from http://ccld-eh.org/tricl/

[2] Even though Spring 2020 student satisfaction scores were likely affected by the onset of the COVID-19 pandemic situation, the increase is sizeable.

Chelliah, S. (2021). *Why Language Documentation Matters*. Dordrecht: Springer.
Computational Resource for South Asian Languages. (2020). *Collaborative Language Archiving Curriculum*. Retrieved from https://corsal.unt.edu/curriculum
Frederick, S. (2019). Decolonization in the archives: at the item level. *iJournal*, 4(2), 14-22.
Hildebrandt, K. (2020). *Archives: Perspectives from Three Scholars of Tibeto-Burman*. (Presented at the Computational Resource for South Asian Languages 4th Annual Meeting, 1 October 2020.) Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1727526/
Hinton, L., & Pérez-Báez, G. (2018). The Breath of Life workshops and institutes. In L. Hinton, L. Huss, & G. Roche (Eds.), *The Routledge Handbook of Language Revitalization* (pp. 188-196). Routledge.
Kung, S. S., Sullivant, R., Pojman, E., & Niwagaba, A. (2020). *Archiving For the Future: Simple Steps for Archiving Language Documentation Collections*. Retrieved from https://archivingforthefuture.teachable.com/
McKemmish, S., Chandler, T., & Faukhead, S. (2019). Imagine: a living archive of people and place "somewhere beyond custody". *Archival Science*, 19, 281-301.
National Science Foundation (2018). *Data Management Plan for SBE Proposals and Awards.* Retrieved from https://www.nsf.gov/news/news_summ.jsp?cntn_id=118038
Roeschley, A., Kim, J., & Zavalina, O.L. (2020). An exploration of contributor-created Description fields in participatory archives. In A. Sundqvist, G. Berget, J. Nolin, & K. Skjerdingstad (Eds.), *Sustainable Digital Communities. iConference 2020. Lecture Notes in Computer Science, 12051* (pp. 638-648). Springer.
Tasker, G. & Liew, C.L. (2020). 'Sharing my stories': genealogists and participatory heritage. *Information, Communication & Society*, 23(3), 389-406.
Theimer, K. (2011). *Exploring the Participatory Archives*. (Presented at the 2011 Society of American Archivists annual meeting.) Retrieved from http://www.slideshare.net/ktheimer/theimer-participatory-archives-saa2011
Ungsitipoonporn, S., Watyam, B., Ferreira, V., & Seyfeddinipur, M. (2021). Community archiving of ethnic groups in Thailand. *Language Documentation and Conservation*, 15, 267-284.
Wasson, C., Holton, G., & Ross, H. (2016). Bringing user-centered design to the field of language archives. *Language Documentation and Conservation*, 10, 641-671.