Language Archiving Training: A Case Study of a Metadata Course in Library and Information Science Graduate Program, 2020 - 2023

Oksana L. Zavalina
Department of Information Science
University of North Texas
United States of America
Oksana.Zavalina@unt.edu

ABSTRACT

Since the early 21st century, funding agencies have been continuously supporting efforts aimed at language preservation and revitalization. This includes providing online access to unique and valuable collections of language data, which often originates from Indigenous and endangered language communities. Language materials are organized and represented in digital archives mostly by information professionals in the library, museum, and archival fields. However, a gap exists between the way these materials are organized and represented and the understanding of that data – and expectations towards the more functional ways of its organization and representation – by language preservation and revitalization researchers, and by members of language communities. Information resources collected by language archives have unique attributes of importance to their target user groups, and these attributes and their representation are not currently widely addressed by the formal training provided to information professionals. Similarly, specifics of these collections end-users' information needs are not currently examined in this training. In this case study, the project that seeks to address this training gap is presented and its preliminary results are evaluated.

CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Social and professional topics → Professional topics → Computing education → Model curricula • General and reference → Evaluation

KEYWORDS

Language archiving training, metadata training, graduate education, Library and Information Science curriculum.

ACM Reference format:

Oksana L Zavalina. 2023. Language Archiving Training: A Case Study of a Metadata Course in Library and Information Science Graduate Program, 2020-2023. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114299

1 INTRODUCTION

Language archives provide access to various kinds of materials, including those unique for them (e.g., word lists), and those that other kinds of archives also frequently hold (e.g., notebooks, oral histories, recordings of community cultural events, etc.). The Open Language Archives Community (OLAC) specializes in providing access to language archival resources. Its Language Resource Catalog includes metadata records representing over 170 thousand resources held by over 60 archives around the world. For example, 36 of these metadata records represent individual items in archival collections related to Amdo Tibetan language, such as the *Medical Secretary and Doctor in Sokdzong (Sokdzong)* Amdo-Tibetan-language oral history transcript held by *COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)* archive in France. As can be seen in this example record, OLAC metadata records follow the Dublin-Core-based OLAC metadata scheme and make use of OLAC-developed controlled vocabularies.

OLAC is not the only centralized portal through which one may access metadata records representing language archive resources. As of June 2023, two well-known global metadata aggregators include a large number of records representing archival resources: over 2.3 million in the WorldCat database and over 7 million in the ArchiveGrid. It is not clear how many of these metadata records represent language archive resources, and more specifically digital ones: this is not one of the metrics published by developers of these aggregators. However, the estimates can be obtained by searching these databases for archival resources in a specific language.

For example, searching the ArchiveGrid by the phrase "Maori language" retrieves 60 exact matches. Some of these metadata records represent archival items (e.g., *Draft for unpublished second edition of the Grammar of the New Zealand language, 1827-1832 [by Kendall, Thomas]*). Most records though represent archival collections: for example, *United States, Indiana, Bloomington, Polynesian languages, 1949-1957* collection of word lists, dialect texts, speeches, grammatical statements with examples, and other text held by the University of Indiana Libraries.

Similarly, a WorldCat search combining the type of resource (archival material) and language (Cherokee) queries reveals that WorldCat currently includes 22 metadata records categorized as representing Cherokee-language archival resources. This includes some individual items such as *Cherokee Nation's record book* for 1902-1903 years. Most of the records though represent archival collections such as for example William West Long's *Cherokee Medicinal and Magical Texts, 1928-1936* collection of 2 notebooks (recorded in the Sequoyan syllabary) and 91 other items held by the American Philosophical Society Library.

Searches like the ones presented above demonstrate that language archive materials are largely held by libraries of various kinds. Thus, metadata to represent these resources is created by library professionals, sometimes in collaboration with those linguists who collected archival materials or with language community members [3] but often – as is the case with legacy materials – without. In general, providing access to legacy data in digital language archives presents several challenges, including those related to provenance, orphan data, and citation tracking [12].

There is clearly the need for providing the information professionals with training that would allow them to identify those attributes of resources in language archives that are important for the users (linguistics and language speakers, instructors, or learners) and to accurately represent these attributes in metadata. Until recently, such training was not provided. As a result, digital language archive materials are often made available to users in a less functional way (e.g., [1], [11]).

Research also demonstrates what specifically is missing in how digital language archival materials are represented. For example, interviews and observations revealed that, for many users, the Language element of metadata records, as well as and representations of the relationships between items (e.g., an audio recording and textual transcriptions or translations) are most important in their interactions with language archives, yet sometimes not represented [4, 6]. Most of respondents in Burke et al. [4] study noted that multilingual interfaces would enable more users to access digital language archive data. Some users also noted that maps displaying the geographic area where the languages are spoken would allow them to find materials easily.

The formal training in digital language archiving offered to information professionals needs to reflect these user preferences, as well as the best practices available. Some best practices for digital language archives were shared by teams of researchers and practitioners from different countries. For example, R and Takhellambam [10] presented the case study of the Sikkim-Darjeeling Himalayas Endangered Language Archive and discussed the collaborative digital language archive development in India. Two studies shared the experiences of the Computational Resource for South Asian Languages (CoRSAL) digital language archive: Dale [7] presented the approaches tested in the development of workflow for mediated archiving while Burke and colleagues [5] discussed the challenges and proposed solutions for name and subject representation in the digital language archive metadata.

Several researchers of digital language archives concluded that education for information professionals needs to cover the specifics of applying archiving techniques and tools in language archives. This paper reports on the project that seeks to address this curricular need, presents intermediate results of this project, and discusses next steps. This paper extends the early results report that was presented at the Association for Library and Information Science Education annual meeting in 2021 [13].

2 COURSE DESIGN

After the initial experiment teaching an interdisciplinary digital language archives metadata course to a combined class of linguistics and library and information science graduate students in the Spring of 2020, the decision was made that to maximize the benefits of this coursework, our team would need to develop the modules focusing on digital language archives and integrate them in relevant courses for information professionals and for linguists. The first candidate for inclusion of such a module was the advanced elective graduate course with the focus on digital library metadata that had participated in the initial experiment. During the Fall of 2020, we developed a digital language archives metadata learning module to integrate in the course and revised the other existing modules to draw examples from language archives in both lectures and assignments.

The new version of the course was tested in the Spring 2021 semester and was taught two more times since then: in Spring semesters of 2022 and 2023, with the cumulative enrollment of 42. To enroll in this course, students must successfully complete the core course in the fundamentals of information organization, followed by the introductory digital metadata course. In the immediate prerequisite (introductory metadata course), students develop knowledge and skills related to the application of major metadata standards, This includes use of data content standards, data value standards (controlled vocabularies), data encoding and transmission standards (XML, HTML, and to some extent MARC 21), and major metadata elements sets for metadata creation to describe items (Dublin Core DCTERMS, MODS, VRA Core 4.0) and collections (Dublin Core Collection Application Profile, Encoded Archival Description, MODS collection application profile, and use of VRA Core 4.0 collection record type). Learning materials of this prerequisite introductory metadata course discuss the user needs and their role in developing metadata element sets, controlled vocabularies, etc., and providing access to information, at the general level, with some examples.

This training prepares students to closely examine the metadata principles and tools in relation to digital language archives, in the advanced digital library metadata course which consists of 4 modules. With the course so far offered only in 16-week spring semesters, the class spends 4 weeks on each learning module. In the weekly class meetings, the teaching team presents material in an interactive way, with numerous illustrative examples, brainstorming and mini-exercise activities for students to help digest the content.

The course opens with Module *1. Metadata for Cultural Works and Specialized User Communities: Language Documentation Case Study* focused entirely on digital language archives. The learning objectives of this module are:

- Identify the needs of a specialized user community, types of materials of interest to these users, general and specific metadata standards that can be utilized in representing these materials for these audiences. Implement this knowledge in metadata work, including investigating relations between metadata elements and user tasks based on conceptual models, navigating controlled vocabularies, and selecting appropriate terms.
- Examine and evaluate current trends in metadata theory and practice, as well as perspectives of developing and applying metadata to provide effective information access for specialized user communities.

During the first week of a learning module, students participate in the class meeting (or review posted slides and recording) and select and read 2 items from the list of 20 or more relevant peer-reviewed professional and/or research publications prepared by the teaching team. These readings are then summarized and critiqued by each student in the discussion post. Students read each other's discussion posts and react to them, with the requirement to provide a substantial reaction to at least one of their classmates' discussions.

Each module has a major practical assignment, that a student completes in weeks 2-4 of the module. For the Module 1 that focuses on digital language archives, the practical assignment includes two parts. In Part 1 (Language Materials, their Users, User Tasks, and Metadata), students answer 4 blocks of questions based on their understanding and critical evaluation of the documentary linguistics workflow and types of materials collected by linguists, as well as user tasks and the specific ways in which metadata fields in a record address them as discussed in two models: the Functional Requirements for Bibliographic Records, and the IFLA Library Reference Model [8, 9]. In Part 2 (General and Specialized Controlled Vocabularies for Representing Resources in Language Collections to Facilitate Information Access), students make use of 16 data value standards (including 4 OLAC controlled vocabularies) to find and examine authority records or other controlled vocabulary entries for terms, names, and codes relevant for representing digital language archive materials.

In the remaining 3 learning modules – Metadata Quality, Metadata Interoperability, and Metadata as Linked Data – examples from digital language archives are used as much as possible, to keep engaging students with the issues related to digital language archives throughout the course. *Module 2. Metadata Quality* also has a significant digital language archiving component. In its practical assignment, students collect and analyze a sample of metadata records from a collection in the CoRSAL digital language archive based on three major metadata quality criteria (accuracy, completeness, and consistency) defined in Bruce and Hillmann [2]. Students compare results of this evaluation to those for a metadata sample in another (non-language-focused) collection that is hosted by the same institution and relies on the same metadata scheme.

3 LEARNING EFFECTIVENESS

This advanced graduate metadata course with the content focusing on digital language archives is overall well received by students, with student satisfaction scores in 2021-2023 ranging between 4.1 and 5.0 on a 5-point scale (response rate 50% - 90%). Here we present some preliminary results of basic quantitative evaluation of students' performance in the two modules with significant digital language archives content components.

To measure effectiveness of the digital language archiving learning in this course, we developed the following targets:

1. Individual target: each student receives at least 85% of possible cumulative points for 3 assessments. For that evaluation, we selected both assessments in Module 1 (discussion forum, and practical assignment) and a practical assignment in Module 2.
2. Class target: at least 90% of students meet the individual target.

In Spring 2021 when our grading was lenient because this was the first semester this course was offered in its current form – all 12 enrolled students (100%) met the individual target, with the average score of 95.66% and the median of 96.61% of possible cumulative points. In Spring 2022, 15 out of 20 enrolled students (75%) met the individual target. However, the average and median scores were quite high: 88.07% and 90.48% of cumulative possible points. Also, when only looking at Module 1 that solely focused on digital language archives, the Spring 2022 results were higher: 85% of students met the individual target, with the average of 92.35% and the median of 92.94%. In the most recent semester (Spring 2023), 90% of enrolled students met the individual target. The average percentage of possible cumulative points achieved by the student was 91.41%, and the median was 91.8%.

Our next step would be to conduct a more detailed analysis of digital language archives learning effectiveness using the available data. For example, we would investigate which type(s) of questions on the practical exercise in the digital-language-archives-focused Module 1 students tend to perform better and worse on. This would allow us to assess the implications for further development and improvement of training materials. Also, detailed examination of student feedback on accuracy and completeness of metadata representing items in the CoRSAL archive collections obtained as part of Module 2 exercise will help identify areas of improvement for CoRSAL metadata.

4 CONCLUSION

Our report presents a case study of the graduate course that begins to bridge the gap in information professionals' understanding of digital language archives users and their needs, materials included, and metadata needed. It will be useful for other educators working on addressing this curricular need.

Overall, the results meet our expectations yet the observation that digital language archives learning effectiveness was lower in the semester with the highest so far (yet still reasonable) enrollment of 20 students warrants further monitoring.

The course in question focuses on metadata, so some other important aspects of the digital language archives are outside of its scope, and either were not covered or did not have a practical assignment (or its component) addressing them. As more of the relevant courses for librarians and archivists are starting to integrate content that develops knowledge and skills necessary to successfully manage digital language archives, future studies will need to compare the instructional approaches, course materials, and results, with the goal of improving such training.

REFERENCES

[1] 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. Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1707416/

[2] Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. ALA editions.

[3] Burke, M. (2021). Collaborating with Language Community Members to Enrich Ethnographic Description in a Language Archive. In Proceeding of LangArc-2021 (1st International Workshop on Digital Language Archives), 18-21. https://doi.org/10.12794/langarc1851172

[4] Burke, M., Zavalina, O. L., Chelliah, S.L., & Phillips, M. E. (2022). User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. Language Documentation & Conservation, 16, 1-24. Retrieved from https://scholarspace.manoa.hawaii.edu/handle/10125/74669

[5] Burke, M., Tarver, H., Phillips, M.E., & Zavalina, O. (2022). Using existing metadata standards and tools for a digital language archive: a balancing act. The Electronic Library, 40 (5), 579-593. https://doi.org/10.1108/EL-02-2022-0028

[6] Burke, M., Zavalina, O. L., Phillips, M. E., & Chelliah, S. (2021). Organization of knowledge and information in digital archives of language materials. Journal of Library Metadata, 20(4), 185-217. https://doi.org/10.1080/19386389.2020.1908651

[7] Dale, M. (2022). Creating workflow for mediated archiving in CoRSAL, The Electronic Library, 40 (5), 568-578. https://doi.org/10.1108/EL-02-2022-0027

[8] International Federation of Library Associations and Institutions. (2008). Functional Requirements for Bibliographic Records. Retrieved from https://cdn.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr_2008.pdf

[9] International Federation of Library Associations and Institutions. (2017). Library Reference Model. Retrieved from https://repository.ifla.org/bitstream/123456789/40/1/ifla-lrm-august-2017_rev201712.pdf

[10] R., K.N. & Takhellambam, M. (2022). A collaboratory model for creation of digital language archives in India. The Electronic Library, 40 (5), 594-606. https://doi.org/10.1108/EL-02-2022-0030

[11] Wasson, C., Holton, G., & Ross, H. (2016). Bringing user-centered design to the field of language archives. Language Documentation and Conservation, 10, 641-671. Retrieved from http://hdl.handle.net/10125/24721

[12] Weber, T. (2022). Conceptualising language archives through legacy materials. The Electronic Library, 40 (5), 525-538. https://doi.org/10.1108/EL-02-2022-0029

[13] Zavalina, O.L., & Chelliah, S.L. (2021). Exploring language archiving education for information professionals and interdisciplinary collaboration to support information access. Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2021. Seattle: ALISE. Retrieved from: https://www.ideals.illinois.edu/items/118795.