# Digital Language Archiving: Reuse and Adaptation of Non-LIS Learning Materials in LIS Education

Sergio I. Coronado[a] and Oksana L. Zavalina[b]

[a b] University of North Texas, US

sergiocoronado@my.unt.edu, Oksana.Zavalina@unt.edu

## ABSTRACT

As a digital repository type of growing prominence, language archives need qualified information professionals to maintain and provide access to their rich collections. Yet, awareness of digital language archives among the US LIS professionals and educators is currently low. Our project's analysis of available relevant training materials for linguists and community members informed the development of a graduate course that prepares LIS professionals for effective and ethical stewardship of digital language archives. In this paper, following a brief literature review of the latest relevant research and publications, results of examination of three major sources of language archiving learning materials are presented and discussed. This discussion's focus is on the ideas for reuse and adaptation of these materials in LIS education and on implementation of some of these ideas in our project-developed course at the University of North Texas.

## ALISE RESEARCH TAXONOMY TOPICS

Information services -- Archives; Information services -- Academic libraries; Education of information professionals -- Curriculum; Education of information professionals -- Online learning

## AUTHOR KEYWORDS

Digital Language Archives; Graduate Education; Endangered Languages Documentation Programme; Archiving for the Future; Collaborative Digital Language Archiving Curriculum.

DOI: https://doi.org/10.21900/j.alise.2024.1736

## INTRODUCTION AND BACKGROUND

With minimal awareness about digital language archives in the US LIS community, there are no LIS academic curricula focusing on them. Nonetheless, there are existing open access online training projects intended for teaching digital language archiving to other audiences. Their learning materials can and should be reused and adapted for helping LIS students develop knowledge and skills required to meet the needs of digital language archives end-users and depositors by stewarding these valuable collections and organizing adequate access to them.

Language collections and archives consisting of multiple language collections are often included in digital repositories of academic libraries. These archives emerged in the 1990s, spearheaded by linguistics scholars, and have since played a significant role in archiving linguistic data for "language revitalization and repatriation" purposes (Burke et al., 2022). Chelliah (2021) states that digital language archives were born out of a necessity for language documentation as well as "long-term preservation and access of documentary linguistics," and function as digital repositories for archiving of linguistic data. Burke and Zavalina (2020) define digital language archives as holding materials "about or in a set of languages, including audio and video recordings, transcriptions, translations, and linguistic annotations." Digital language archives are openly accessible online repositories that help facilitate access to linguistic materials and other cultural heritage resources not only for academic audiences but importantly for communities of language speakers, including endangered languages. However, the needs of each of these target user groups are not fully met, in part due to lack of education on the specifics of digital language archive materials and user needs for LIS professionals who manage these archives (Wasson, Holton, & Ross, 2016; Burke et al., 2022).

Digital language archives have recently become a research topic in information science, with a special issue of *The Electronic Library* journal devoted to them in 2022. They also have been gaining attention and importance in the archival community of practice, particularly in relation to the archival "community paradigm shift" (Cook, 2013). Participatory community language archiving has become an important part of community-based and participatory archives whereby community members contribute to the development and understanding of archival collections, such as through creation, policies, personal stories, and so forth (Thiemer, 2011; Rolan, 2017; Roeschley & Kim, 2019). The development of these archival language collections relates back not only to the cultural preservation and revitalization of endangered languages but also to endangered information, knowledge and culture, and intercultural information ethics. This context guided our comparative examination (presented below) of learning materials available through major online projects for teaching linguists and language community members about digital language archives.

## EXAMINATION OF DIGITAL LANGUAGE ARCHIVE LEARNING MATERIALS

The first online project that we examined is the ***Endangered Languages Documentation Programme (ELDP)*** (https://www.eldp.net), the program continuously funded by the Arcadia Fund since 2002. Its mission and purpose are to help document and preserve endangered languages through funding of the language documentation projects (30-40 grants per year around the world) and making them openly available and freely accessible for everyone. ELDP output is composed of various educational materials, trainings, and resources for teaching and learning about documentary linguistics and language documentation as well as how to create a repository of linguistic resources. These materials are licensed under a Creative Commons Attribution Share-Alike (CC BY-SA) license, allowing for their sharing, adaptation, translation, and/or republishing. *ELDP* output is designed specifically for educating about digital language archives. Particularly, these resources include a Grantee Training (provided only to ELDP grantees) consisting of several language documentation topics, and Training Resources: modules encompassing various topics developed by Endangered Languages Archive (ELAR) and available to the general language documentation community.

The second online project examined is ***Archiving for the Future*** (https://archivingforthefuture.teachable.com/) funded by the US National Science Foundation, and developed by the Archive of the Indigenous Languages of Latin America at the University of Texas at Austin in consultation with Digital Endangered Languages and Musics Archives Network participants and other digital repositories around the world. The project's content is licensed under CC BY-SA 4.0 International, is a self-paced introductory online training course focused on teaching "language documenters, activists, and researchers" about the organization, arrangement and archiving of languages, as well as about "language documentation, revitalization and maintenance" of different "materials and metadata in digital repositories or language archives." *Archiving for the Future* has similar teaching and training materials regarding language documentation to that of *ELDP*. All its training content is also freely and openly available but is designed for broader audiences (language documenters, activists, and researchers), in comparison to *ELDP* whose training focuses on language documentation learners (novice to experts).

The third online training examined is the ***Collaborative Digital Language Archiving Curriculum (CoDA)***, developed and hosted by the Computational Resource for South Asian Languages (CoRSAL) project at the University of North Texas (UNT). According to its website (https://corsal.unt.edu/curriculum), *CoDA* is an "accelerated course for community language documenters, language revivalists, digital archivists, and linguists interested in the documentation and description of language." It is designed for "non-academic documenters," "documentation at a distance," and "linguistic students," along with consideration of varying technological needs. It follows a modular curriculum structure with modules focused on different themes of language documentation and data management. The course is licensed under a Creative Commons Attribution Non-Commercial ShareAlike 4.0 International License, which allows for its materials to be shared, adapted, translated and/or republished. The *CoDA* curriculum audience scope is targeted more towards linguists and non-academic documentalists rather than activists or researchers like *Archiving for the Future* or different levels of language documentation learners (novice to experts) like *ELDP*.

These three projects, despite their differences, appear to provide learning content useful for developing a language archiving and language documentation training for information professionals who will be responsible for maintaining digital language archives. Table 1 provides an overview of the above-described examination of these online projects. A summary of the results of examination of these learning materials' contents—main topics, assignments and activities—and how they can be reused and repurposed as learning materials for teaching about digital language archiving as part of library graduate coursework is provided in the following sections.

**Table 1**
*Online Digital Language Archiving Training Projects—Audiences and Learning Materials*

| | ELDP | Archiving for the Future | CoDA |
|---|---|---|---|
| Intended Audience | Grantees & Language Documentation Learner Community | Language Documenters, Activists & Researchers | Language Documenters, Revivalists, Digital Archivists & Linguists |
| Learning Materials | Grantee Training & Training Resources | Training Course and Phases & Steps | Course Curriculum & Course Modules |

## TOPICS PRESENTED, ASSIGNMENTS, AND ACTIVITIES

***Endangered Languages Documentation Programme's*** (https://www.eldp.net) training consists of the following three main topics, each covered by a module: "Data Preparation," "Metadata Creation" and "Archiving with ELAR." Within each topic, there are sub-topics also. Specifically, Data Preparation is subdivided into "Data Management," "ELAN Fundamentals," "ELAN Intermediate," and "ELAN-FLEX-ELAN Workflow." The Metadata Creation topic is subdivided into "Metadata Creation with Lameta" and "Topics, Keywords and Genres." The Archiving with ELAR topic is subdivided into "Depositing Overview," "Uploading Using PUT," "Uploading using SIP Creator," and "Adding Files to Existing Bundles." Also, although there are no assignments specifically, each module is composed of different reading and viewing activities for training participants, such as video tutorials, PDF guides, documents and teaching sets, all of which provide an explanation of different language archiving concepts, tools and linguistic software for data management (ELAN, FLEx, FFMPEG, SayMore, Keyman, LaMeta, Preservica) and depositing workflows of ELAR, ExLibris SIP, etc.

***Archiving for the Future*** (https://archivingforthefuture.teachable.com/) training is subdivided into three major modules—that are referred to as "phases" based on the sequence of tasks completed in digital language archiving—with multiple steps covering various language archiving topics under each phase. The main topics are language archiving; data collection; metadata, file naming and file formats. Examples of subtopics (labeled as "steps") include file naming, file format selection, and planning for metadata (steps of Phase 1: Before Data Collection). There are several activities and assignments that learners are expected to complete in each phase. One activity is textual readings—provided under each step—that give further detailed explanation of the main topic. Another activity is a vocabulary review provided towards the end of each step for each phase to help highlight key terms. There are also practical assignments (referred to in the course as "activities") found within each step. This includes naming files, evaluating file types, examining record metadata, exploring repositories, collecting metadata, and other tasks that help reinforce the skills associated with the topic learned in each of the steps. There is also a Certificate of Completion for the *Archiving for the Future* training (it takes two weeks to process it via email).

***Collaborative Digital Language Archiving Curriculum*** (https://corsal.unt.edu/curriculum) includes nine modules, with three of them intended for researchers collecting the language data, and remaining six dealing with topics of language collection creation, language materials archiving, language materials management and description, language documentation tools, and archival file preparation. In addition to textual readings in every module, this training includes a variety of assignments in each module, including discussion questions and project activities. As part of these projects, training participants are expected to write reports, create video/audio recording, scan a document, explain data corruption or loss, compare a website to an archive, review language collections, create metadata, use ELAN, SayMore and FLEx linguistic software for file creation, develop a collection landing page and interview guide, etc.

Table 2 below provides the summary of topics and assignments presented in each of the 4 training projects we examined in this study.

**Table 2**
*Online Digital Language Archiving Training Projects—Topics, Assignments and Activities*

| | ELDP | Archiving for the Future | CoDA |
|---|---|---|---|
| Topics | Data Preparation, Metadata Creation, Archiving with ELAR | Language Archiving, Metadata Collection, File naming & File Formats | Language Collection Creation, Language Materials Archiving, Description, Management, Language Documentation Tools & Archival File Preparation |
| Assignments and activities | Video Tutorials, PDF Guides/ Documents, Teaching Sets | Textual Readings, Vocabulary Review, Step Activities | Textual Readings, Discussion Questions, Project Activities |

## DISCUSSION OF REUSE & ADAPTATION OF MATERIALS IN LIS EDUCATION

Based on the examination of outputs of these online training projects, the following are some ideas on how some of these materials can be reused and adapted for the purpose of educating LIS graduate students about digital language archiving and curation. Brief information about how they were adapted by our team in the project-developed graduate course are included in this section. More details on the principles of development of our course, its learning objectives, and topics covered can be found in Zavalina and Paterson (2024) and Zavalina, Chelliah and Frederick (2024).

***Endangered Languages Documentation Programme.*** Given that *ELD*P project content contains many training resources on specialized linguistic software tools—such as ELAN, FLEx, FFMPEG, SayMore, Keyman, LaMeta, etc.—these trainings can be integrated into lessons on metadata creation and data preparation tools for student exploration and awareness of these tools. Another way is to incorporate its Video Tutorials and PDF Guides/Documents to corresponding lessons on language archiving, particularly on topics like creating metadata, preparing language data and learning how to archive with the ELAR digital language archive to provide a first-hand experience of these training guides and real-life online tools and facilitate instruction on these subjects.

***Archiving for the Future.*** Since this entire online project is focused primarily on language data collection, the whole training, except for its activities, can be reused in the form of required or recommended reading for providing a general introduction to language archiving and then giving a more detailed explanation about the before, during and after process of collecting language data for addition to a digital language archive. The vocabulary review content can be reused to design a quiz for important terms about language data and other key concepts about data collection. One other way is to apply some of the activities as short practical assignments, particularly on file naming and selecting file formats since those play an important role when archiving language data, which may get overlooked at times by information professionals.

***Collaborative Digital Language Archiving Curriculum***. Since it is a course composed of several modules on digital language archiving, one way it can be reused is to inform the selection and organization of course content for a graduate course for information professionals on this topic. More specifically, its readings can be reused to supplement course content on topics like creation of language collections, describing and managing language materials, preparing archival files, among others. They could also be used to familiarize LIS students with essential language documentation tools, like ELAN, SayMore and FLEx and archival files resulting from the use of this software. *CoDA* discussion questions can be reused as weekly or bi-weekly prompts to engage students in sharing reflections on course topics. Finally, *CoDA* project activities can be adapted and scaled-up for major assignments of an LIS graduate course, such as examining a language collection as a case study, developing audiovisual recordings documenting the use of a language, creating metadata on language recording, building a language collection, and many more.

Based on the ideas discussed above and informed by our analysis, as part of the IMLS-funded project, we created and tested in the Summer semester of 2024 the UNT graduate course *Community Language Archiving and Curation for Information Professionals*. In our curriculum development, some components of the content from these 3 online training projects have been integrated—with appropriate attribution—into readings and practical assignments. Specifically, some content from *CoDA*—such as language material types, file structure, file naming, etc.—was integrated into one module. Likewise, some content from *Archiving for the Future*—such as history and file formats—was integrated into one module; some content from *ELDP*—such as file structure and website examples from ELAR—was integrated into two modules. The content of the project-developed course was also integrated into other relevant LIS program courses (*Advanced Metadata* and *Cultural Heritage Stewardship*), as well as in linguistics courses (*Corpus Linguistics* and *Field Methods*). Our IMLS-funded project is ongoing, and we are in the process of revising the course content based on feedback received from students (reported in Zavalina, Chelliah &

Frederick, 2024) and digital language archive experts. Some findings from the course evaluation were reported in

## CONCLUSION

Results of this analysis can contribute to the design, development and implementation of curricula for organized graduate and undergraduate courses in LIS academic programs, as well as for professional development of existing librarians and archivists employed in positions that require stewarding language archive collections and ensuring their digital curations. They could also be used to inform continuing professional development for researchers, practitioners, linguists, educators and language community members interested in working with information professionals on the creation of digital language archives. These educational initiatives will help increase visibility and promote digital language archives and their workforce.

## ACKNOWLEDGEMENTS

This project is supported by the United States Institute of Museum and Library Services Laura Bush 21st century Librarian program (grant RE-254860-OLS-23).

## REFERENCES

Burke, M., & Zavalina, O.L. (2020). Identifying challenges for information organization in language archives: Preliminary findings. In: Sundqvist, A., Berget, G., Nolin, J., Skjerdingstad, K. (eds). *Proceedings of the 15th International Conference of Sustainable Digital Communities. iConference 2020. Lecture Notes in Computer Science, 12051*. Springer, Cham. https://doi.org/10.1007/978-3-030-43687-2_52

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 and Conservation, 16*, 1-24. https://scholarspace.manoa.hawaii.edu/handle/10125/74669

Chelliah, S. (2021). Why Language Documentation Matters. Dordrecht: Springer.

Cook, T. (2013). Evidence, memory, identity, and community: Four shifting archival paradigms. *Archival Science*, *13*(2-3), 95-120. https://doi.org/10.1007/s10502-012-9180-7

Roeschley, A., & Kim, J. (2019). Something that feels like a community: the role of personal stories in building community-based participatory archives. *Archival Science, 19*, 27–49. https://doi.org/10.1007/s10502-019-09302-2

Rolan, G. (2017). Agency in the archive: a model for participatory recordkeeping. *Archival Science, 17*, 195–225. https://doi.org/10.1007/s10502-016-9267-7

Thiemer, K. (2011). Exploring the participatory archives [Slides]. https://www.slideshare.net/slideshow/theimer-participatory-archives-saa-2011/9071551

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

Zavalina, O.L., Chelliah, S.L., & Frederick, M. (2024). Supporting the needs of community digital language archive users through training for information professionals: Presentation at the Digital Library Federation (DLF) Forum, July 30, 2024. [Slides]. https://osf.io/hgq75/

Zavalina, O.L., & Paterson, H. J., III. (2024). Developing graduate curriculum for digital language archive stewardship. In, *ALISE 2024 Proceedings* (13 pp.). https://doi.org/10.21900/j.alise.2024.1657