Information Research - Vol. 30 No. iConf (2025)

Stewardship of digital language archives: training development and testing through collaboration of information scientists, linguists, and communities

Oksana L. Zavalina, Alexandra O'Neil, and Shobhana L. Chelliah

DOI: https://doi.org/10.47989/ir30iConf47311

Abstract

Introduction. Specialized repositories aggregating digital data that focus on languages (of indigenous groups, refugees, etc.) are known as digital language archives (DLA). DLA rapid growth is galvanized by language documentation efforts supported by funding agencies. Recent studies examine DLA user needs and explore support of these needs in digital repositories. Training of LAM professionals in curating and managing DLA to support user needs is emerging, with no research yet into its effectiveness.

Method. Our federally funded interdisciplinary project creates and tests the 1st US LAM graduate course with DLA focus. The curriculum development builds on team's experience with DLA, available DLA user studies, and collaborations of LAM and linguistics researchers with language communities. We report on the current state of curriculum development, discuss analysis results and future steps.

Analysis. Quantitative and qualitative analyses of data collected through course-level and module-level surveys, and content analysis.

Results. Overall, the project-developed training materials are effective in developing learning objectives. Areas for improvement are identified.

Conclusion. The team is refining the LAM curriculum and developing language community DLA training workshops based on student evaluation results, collecting community feedback for future analysis. The project is expected to positively impact LAM education and DLA users' experiences.

Introduction

A digital language archive (DLA) is an archive that includes born-digital and/or digitized language resources in any language(s) or dialect(s), a searchable repository in which resource discovery is powered by metadata. The collections comprising DLAs are often created with the goal of preserving, or documenting, a language, especially an endangered language. Once documented, these materials assist the revitalization efforts of the language(s) in question (Hinton, 2001). Thus, items selected for these collections demonstrate language use in various settings and through different media and file formats (Chelliah, 2021). Some of the common item types in DLAs are also collected by broader-scope digital repositories stewarded by cultural heritage institutions, and include audio-visual recordings of oral histories, performances, interviews, text resources: books, articles, letters, diaries, etc. However, DLA also include a significant proportion of resource types that are not found in other digital repositories: wordlists, field notes, AI-generated and Large Language Models related content, corpus publications essential for computational linguistics research; linguistic annotations, transcriptions, and software (Bird & Simons, 2003).

Literature review

In the past, language communities, despite often being the source of language archive materials, were typically not involved in the creation or maintenance of archives (Henke & Berez-Kroeker, 2016). This practice privileged linguistics and anthropology researchers, missionaries, and colonial administrators as creators of archival records (i.e., items and collections), ascribing provenance to them over the individuals and communities depicted in the archival records (Jones, 2019). However, like other archives, DLAs undergo significant changes and evolve to meet the requirements of the archival paradigm shift toward community-oriented ways of archival thinking and participatory methodologies (Cook, 2013; Genovese, 2016; Rolan, 2017). For DLAs, that means that source language communities of the archival records take an increasingly active part in decision-making regarding donation, appraisal, preservation, metadata, access, dissemination, etc. (Burke, 2021; Dale et al., 2022; Harris et al., 2019; Olson, 2023; Pugh & Christen, 2011; Roeschley & Kim, 2019).

Information science and linguistics researchers explore the archive user needs of language community representatives and scholars. Developing understanding of these needs is important for improving access as they guide user interactions with and expectations towards DLA. Some of the user needs tend to focus on the services offered by the DLA and include stream-able audio and video, user interface accessibility on mobile devices, as well as machine-readable text files and bulk download availability necessary for computational linguistics research (e.g., Burke et al., 2022). Endangered language community members also report needs – related to supporting the revitalization of their languages – that inform collection development decisions as language communities representatives call for inclusion of specific kinds of resources beyond the obvious choice of dictionaries, audio and video files: textbooks and teaching aids covering different disciplines, folktales and storybooks for children, resources on religion and culture, and overall attractive items that capture attention (Burke, 2023). Research demonstrates that information needs of DLA users are often not met, and that one of the key reasons for this is the lack of academic instruction for information professionals on supporting the stewardship and digital curation of these collections (Al Smadi et al., 2016; Wasson, Holton, & Ross, 2016; Zavalina & Chelliah, 2021).

While there is an overlap in the principles and practices of broader digital archiving, digital archiving of linguistics research data, and community digital archiving focusing on language, there are substantial differences in the goals, scale, content, challenges, ethical considerations, and impact. One example is that types of items unique to language archives and specific attributes of these resources, as well as complex relations between these items must be represented differently in bibliographic metadata, archive's navigation, and controlled vocabularies to meet DLA user information needs (e.g., Aljalahmah & Zavalina, 2023; Paterson, 2024; Zavalin, 2023). Greater focus is needed on authorship and intellectual rights (e.g., with AI-generated data), as well as on ownership, archiving spaces, and mutually beneficial relations between communities and partnering researchers and institutions, identifying, and disseminating best practices, and developing collaborations to support community archiving and to engage communities in stewardship of their DLA collections housed by LAMs (Zavalina & Chelliah, 2021; Chelliah, 2023).

University of North Texas (UNT) educators experimented with adding DLA training for LAM students as an individual DLA-focused module in the advanced metadata course. Preliminary results of DLA-focused learning module effectiveness indicated that during 2021-2023 testing, 75% to 90% students attained the target cumulative score of 85% or more on module assignments and activities (Zavalina, 2023). Student survey data on teaching effectiveness however were collected and analysed only at the course-level, not for the DLA-focused module. Realization of the need for a more robust in-depth and open-access DLA training for LAM students led to successful proposal to the US Institute or Museum and Library Services (IMLS) for addressing this need.

Expanding LAM workforce training

To meet this demand, our project's interdisciplinary team develops curriculum with a strong experiential component to educate information professionals in the archiving and curation of resources that provide the means to revitalize community memory and language. The project is producing learning materials intended to be useable in courses offered in LIS and Archival university programs, LAM on-the-job trainings, and archiving workshops for members of indigenous, immigrant & refugee communities who are documenting their community heritage.

Project-developed course, data collection and analysis

Our team developed the online graduate course that consists of 4 modules and introduces LAM students to how the language archives have evolved and function today, to needs of the end-users and depositors of these archives, to material types that DLAs collect, specifics of describing the unique attributes of these complex resources in metadata, and digital curation of these materials. This course also introduces students to working with language communities on collection stewardship, ethical and effective dissemination of language archive resources, and evaluation of DLAs. The course content, developed since 2023, successfully went through university-administered assessment of online course pedagogy, copyright/trademark, and ADA accessibility compliance. It was first offered as a section of an existing information resources and services for special clienteles seminar to UNT LAM graduate students during June 24-July 26, 2024, with 11 IMLS-funded tuition stipends, and 23 students completing the course.

At multiple points of time, we collected and analysed various data on effectiveness of the course and its individual learning modules. Student success in completing the tasks of the 4 module assignments was used in the course evaluation: we analysed range, average, standard deviation of the scores attained by student submissions and compared these across modules. At the end of semester, the team obtained anonymous student survey data to evaluate effectiveness in meeting each of the 8 broader course-level learning objectives. Data collection also included anonymous pre-test and post-test surveys to evaluate each of the 4 learning modules, with questions designed to capture indications of perceived progress made by the student and based on the much more specific module-level learning objectives (a total of 23, ranging 4-7 per module). Finally, our team analysed quantitatively and qualitatively the results of the university-administered online teaching effectiveness survey, after these results were released to the instructor in August 2024.

Course evaluation findings and discussion

Results of assignment submissions evaluation demonstrate that 95.65% of students consistently produced high-quality work, earning 80% or more of possible points. All (100%) students attained the cumulative average score (for all 4 assignments) of 85% or more, higher than in Zavalina (2023). The average score attained ranged from assignment to assignment between 18.26-19.75 on the 20-

point scale (91%-99%), with the highest observed for the module 2 comparative discussion assignment and the lowest for the module 4 practical exercise. The highest standard deviation in the student scores (2.464) was observed in module 4, and the lowest (0.876) in module 2. We believe that one of the reasons for differences between module 2 and module 4 results is the different workload levels, with four module-level objectives addressed by module 3 and seven by module 4.

As part of the module-level pre-test and post-test surveys, 70% or more students reported feeling confident or very confident in their objective-related skills and knowledge for 21 (91.3%) of module-level objectives. For 12 (57.1%) of module-level objectives, 81% or more students reported this confidence level. This exceeds our expectations of 70%. However, analysis identified the need for course revisions to ensure improvement in meeting two module-level objectives, where the level of confidence was 52% and 54%. One is the ability to identify digital content management (DCM) tools that could be used for organizing language archives and DCM selection considerations. Another is the ability to determine DLA funding sources and budgeting principles. We attribute this finding to the fact that while these learning objectives are covered in required readings (presentations and examples), they were not reinforced through practical assignment tasks. Yet, even for objectives with lower-than-expected student confidence level demonstrated in the post-test, the difference between pre-test and post-test confidence level was significant.

The end-of-semester project-team-administered survey data indicated that between 91% (1 objective) and 100% students (4 objectives) developed confidence in their knowledge and skills related to course-level learning objectives. This was higher than the 85% expected by the project team. However, unlike with module-level objectives, we did not collect the pre-test survey data for course-level objectives at the beginning of semester.

The university-administered survey results demonstrated high levels of satisfaction with the course content (very good or excellent in 79% of student responses), teaching effectiveness (90%), and the course overall (79%). At the same time, the perceived amount of time to succeed in this course was reported to be higher than average by 63% of students, with 26% considering it much higher. We analysed the answers to open-ended questions, compared to the results of project-team-administered surveys, and used in course revision as the source of additional insight. For example, students suggested adding at least one live online class meeting. Student feedback indicates that a 5-week semester is not ideal for offering a course such as this one, with the variety and complexity of included content. This finding was expected because we developed the course with longer semesters in mind and assigned no required prerequisites, which resulted in a mix of relevant expertise levels among students in the class, with the majority (over 80%) being completely new to DLA topics and needing more time to develop confidence.

Community training and collaboration

This project, implemented by the interdisciplinary team of information scientists and linguists, also has the goal of collaborating directly with language communities to offer training on language archiving and obtain community feedback. The training consists of short workshops in which participants learn about the history, purpose, principles, and practicalities of digital archiving, and take part in hands-on activities designed to familiarize them with digital archives and provide experience with the process of adding an item to a DLA. While community workshops are still in progress, the planning of these collaborations in and of itself has introduced multiple insights that are relevant for future researchers prioritizing community-oriented participatory methods. While the needs of a language community differ in each community, we identify four areas that should be considered in the development of workshop materials for language communities: digital literacy, inter-generational collaboration, technology availability, and language accessibility.

Multiple factors such as socioeconomic status, proximity to urban areas, and community perception of technology, affect the acquisition of digital literacy – skills required to use digital technologies – in communities. Language documentation projects have noted the lack of digital literacy as a barrier for community members to participate in digital resource creation (Frawley, Larkin, & Smith, 2017; Hausknecht et al., 2021; Mwanza, Molepo, & Goduka, 2018). This lack severely impacts the ability of a language community member to contribute to the creation of material in their language through digital archiving, which is detrimental to the language preservation (Domeji et al., 2019). When contributing to a DLA (by submitting an item or a collection), a depositor needs a level of digital literacy that allows performing the associated actions, such as uploading content from a separate device to a computer, file management, creating metadata, and filling out intake information on a web page.

In scenarios where community members lack sufficient digital literacy, DLA community workshops need to cover foundational digital literacy. However, when the digital literacy gap is prominent among older community members, but less pronounced in younger generations, intergenerational collaboration can alleviate the necessity of digital literacy training while simultaneously strengthening relationships between community members. One workshop in our project's series is designed for members of a language community that are university students. While the students have varying levels of proficiency in the language, partnering with older speakers in the community is mutually beneficial as they can support the older speakers navigate the technology used to contribute to archives and learn more about their language and culture from working with the older speaker.

The technology availability affects what happens both during and after the DLA community workshop. The workshops include hands-on activities, such as exploring a digital archive and preparing a file for deposit. To facilitate these activities, the participants must use a computer, that many of them don't have, so it is important to schedule the community workshop in a computer lab or bring laptops for the participants to use. In the non-university settings, access to a computer lab is fairly simple, but in other settings, this can present a substantial coordination challenge. It is also essential to note the technology the participant will be using after the workshop, including but not limited to the general access to computational technology as other factors can cause confusion: for example, differences between the operating systems or text / audio-visual editing software available for participants after the workshop versus what community participants used in training.

Building on ideas of general accessibility, the language used to present DLA to language community members is influential in facilitating understanding and garnering interest in digital archiving. For this reason, the way the content is presented to information science students or linguistics students must differ from the way it is presented to language community members. While the content remains largely unchanged, field-specific terminology should be modified to be intelligible to people without information science or linguistics background. For example, mentioning that the ability to query the metadata associated with items is one of the major advantages of DLA is not a persuasive point when the person has not been exposed to the concepts of metadata and querying. Similarly, expressing the importance of file naming conventions and organization of files must be substantiated in its practical utility: saving time and not losing things. The language used to discuss DLA must be accessible, both in the terminology that it employs and in the practical grounding of concepts.

Conclusion and next steps

The first offering of the project-developed LAM graduate course demonstrates overall success in curriculum development, as well as substantial demand among current students for coursework that is designed to develop the knowledge and skills necessary for successful stewardship of DLA, including community archives. Now that the course is officially added to UNT graduate catalogue as INFO5385, and students who completed it are sharing their experiences with fellow students, the awareness grows among our LAM student population. With 14 additional IMLS-sponsored tuition stipends for the course, another class of 25 is expected to take INFO 5385 after revisions based on results of student evaluations completed so far and on the upcoming results of language community evaluation of project-developed learning materials.

For the longer (10-week) semester of 2025, we will add to INFO5385 the content covering two course-level objectives that did not reach the 70% student confidence target, break four instructor presentations into 2-3 shorter videos per module, add two live online meetings to the asynchronous course. The changes also include augmenting the data collection by adding the pre-semester survey for course-level objectives. The student feedback will be comparatively analysed and used in further refinements of learning materials.

Our team is also working on integrating relevant content from project-developed course modules into other UNT LAM courses and Indiana University Bloomington linguistics courses. We are disseminating preliminary results to obtain feedback from LAM and Linguistics educators and practitioners through presenting at professional meetings (Coronado & Zavalina, 2024; Zavalina & Paterson, 2024, etc.) and raise awareness of the project's products and opportunities it provides among the library and archival communities of practice, as well as among LAM faculty and students. Once the project is completed, project-developed learning materials will be made available as open access resources and are expected to make a solid contribution to addressing the needs of language communities through the extension of LAM professionals' preparation to stewarding DLAs.

Acknowledgements

The project presented in this paper is funded by the United States Institute of Museum and Library Services (grant RE-254860-OLS-23).

About the authors

Oksana L. Zavalina is Professor in Department of Information Science, University of North Texas, USA. She received her Ph.D. from University of Illinois, and her research interests are in the information organization and user needs in libraries and other repositories, including digital language archives. Dr. Zavalina can be contacted at Oksana.Zavalina@unt.edu

Alexandra O'Neil is Ph.D. student in Department of Linguistics, Indiana University Bloomington, USA. Her research interests are in the enhancing access to internet resources and technology for speakers of under-resourced languages. She can be contacted at aconeil@iu.edu

Shobhana L. Chelliah is Professor in Department of Linguistics, Indiana University Bloomington, USA. She received her Ph.D. from University of Texas in Austin, and her research interests are documentary linguistics and language archiving. Dr. Chelliah can be contacted at schellia@iu.edu

References

Al Smadi, D., Barnes, S., Blair, M., Chong, M., Cole-Jett, R., Davis, A., Hardisty, S., Hooker, J., Jackson, C., Kennedy, T., et al. (2016). Exploratory user research for CORSAL. https://digital.library.unt.edu/ark:/67531/metadc1707416/

Aljalahmah, S., & Zavalina, O.L. (2023). Exploration of metadata practices in digital collections of archives with Arabian language materials. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114295/

Bird, S., & Simons, G. (2003), Seven dimensions of portability for language documentation and description. https://doi.org/10.1353/lan.2003.0149

Burke, M. (2021). Collaborating with language community members to enrich ethnographic descriptions in a language archive. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2021). https://doi.org/10.12794/langarc1851172

Burke, M. (2023). Designing Archival Collections to Support Language Revitalization: Case Study of the Boro Language Resource. Thesis or dissertation, University of North Texas. https://doi.org/10.12794/metadc2137570

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. http://hdl.handle.net/10125/74669

Chelliah, S.L. (2021). Why language documentation matters. Springer

Chelliah, S.L. (2023). Making photographs in language archives maximally useful: Metadata guidelines for community and academic depositors. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114301/

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

Coronado, S.I., & Zavalina, O.L. (2024). Digital language archiving: Reuse and adaptation of non-LIS learning materials in LIS education. In, The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024. https://doi.org/10.21900/j.alise.2024.1750

Dale, M., Basumatary, P., Iqbal, J., Khullar, R., & Shaikh, M. (2022). Case study of using Facebook groups to connect community users to archived CoRSAL content. Language Documentation & Conservation, 16, 399–416. https://hdl.handle.net/10125/74685

Domeij, R., Karlsson, O., Moshagen, S., & Trosterud, T. (2019). Enhancing Information Accessibility and Digital Literacy for Minorities Using Language Technology– the Example of Sámi and Other National Minority Languages in Sweden. In, Perspectives on Indigenous writing and literacies (pp. 113–137). https://doi.org/10.1163/9789004298507_007

Frawley, J., Larkin, S., & Smith, J.A. (2017). Indigenous Pathways and Transitions into Higher Education: An Introduction, p. 3–11. Springer, Singapore. https://doi.org/10.1007/978-98110-4062-7_1

Genovese, T.R. (2016). Decolonizing archival methodology: Combating hegemony and moving towards a collaborative archival environment. AlterNative: An International Journal of Indigenous Peoples, 12(1), 32–42. https://doi.org/10.20507/AlterNative.2016.12.1.3

Harris, A., Gagau, S., Kell, J., Thieberger, N., & Ward, N. (2019). Making meaning of historical Papua New Guinea recordings: Collaborations of speaker communities and the archive. International Journal of Digital Curation, 14(11), 136–149. https://doi.org/10.2218/ijdc.v14i1.598

Hausknecht, S., Freeman, S., Martin, J., Nash, C., & Skinner, K. (2021). Sharing indigenous knowledge through intergenerational digital storytelling: Design of a workshop engaging elders and youth. Educational Gerontology, 47(7), 285–296. https://doi.org/10.1080/03601277.2021.1927484

Henke, R.E., & Berez-Kroeker, A.L. (2016). A brief history of archiving in language documentation, with an annotated bibliography. Language Documentation & Conservation, 10, 411–457. http://hdl.handle.net/10125/24714

Hinton, L. (2001). The Use of Linguistic Archives in Language Revitalization: The Native California Language Restoration Workshop, p. 419–428. Brill, Leiden, The Netherlands https://doi.org/10.1163/9789004261723_033

Jones, M. (2019). Collections in the expanded field: Relationality and the provenance of artefacts and archives. Heritage, 2(1), 884–897. https://doi.org/10.3390/heritage2010059

Mwanza, A.J., Molepo, J.M., & Goduka, N. (2018). Addressing the digital literacy gap in the use of information and communication technologies in the dissemination of indigenous knowledge: a rural community experience. Indilinga: African Journal of Indigenous Knowledge Systems, 17(22), 154–166. https://www.ajol.info/index.php/indilinga/article/view/186128

Olson, E.A. (2023). Mrs. his name: reparative description as a tool for cultural sensitivity and discoverability. Journal of Western Archives, 14(1), Article 10. https://digitalcommons.usu.edu/westernarchives/vol14/iss1/10/

Paterson, H.J., III (2023). OLAC and serials: An appraisal. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114296/

Pugh, M., & Christen, K. (2011). Opening archives: Respectful repatriation. The American Archivist, 74(1), 185–210. https://doi.org/10.17723/aarc.74.1.4233nv6nv6428521

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(1), 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(3), 195–225. https://doi.org/10.1007/s10502-016-9267-7

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

Zavalin, V. (2023). Ukrainian archival metadata in WorldCat: Exploratory analysis. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114298

Zavalina, O.L. (2023). Language archiving training: A case study of a metadata course in Library and Information Science graduate program, 2020 - 2023. Proceedings of the International Workshop on Digital Language (ACM/IEEE Joint Conference on Digital Libraries 2023). https://digital.library.unt.edu/ark:/67531/metadc2114298

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. https://www.ideals.illinois.edu/items/118795

Zavalina, O.L., & Paterson, H.J., III (2024). Developing graduate curriculum for digital language archive stewardship. In, The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024. https://doi.org/10.21900/j.alise.2024.1657