Shifting the Focus of Metadata Training to Community User Needs in Language Archives

Oksana L. Zavalina

University of North Texas, USA

Oksana.Zavalina@unt.edu

ABSTRACT

Repositories of digital collections focusing on languages and cultures are known as digital language archives. In the past 15 years, they have grown exponentially, due to language documentation and revitalization work. Materials resulting from these efforts – mostly housed by GLAM institutions, including in community-centered digital collections – are valuable for education, research, and empowering communities. Information professionals are responsible for organizing and describing them to facilitate access and discovery. A gap exists between the ways these information resources are usually organized and expectations of language communities' members and language preservation and revitalization researchers. GLAM-stewarded community language archive items possess uncommon for other information resources attributes and relationships of importance to target audiences. Their metadata representation – and specific information needs of intended audiences – are not yet in the mainstream GLAM curriculum. The paper describes addressing this training gap as part of the advanced graduate metadata course.

ALISE RESEARCH TAXONOMY TOPICS

Curriculum; Metadata; Cross-language information retrieval; Specific populations; Information needs.

AUTHOR KEYWORDS

Archive users; Metadata education; Language archive materials; Community language archives.

DOI: https://doi.org/10.21900/j.alise.2025.2058

INTRODUCTION

Language documentation and revitalization efforts led by linguists and language communities result in large volume of valuable cultural heritage content deposited to digital repositories of general scope and specialized language archives (Henke & Berez-Kroeker, 2016). Also, recent research positions community-centered lexicography – dictionary creation – projects as archival even without depositing to existing archive (Frederick, Roeschley, & Zavalina, 2025). Language archives steward a variety of cultural heritage information resources, including resources only found in these more specialized archives (such as grammar texts and linguistic analyses), and those found in many other digital repositories (for example, recordings of cultural events, transcripts of oral history interviews conducted with community elders, etc.). The Open Language Archives Community (OLAC) provides centralized access to information resources in language archives collections. In OLAC's searchable database of 300 thousand resources ("in half of the world's living languages") that are held by over 60 archives, resources are represented with metadata in the Dublin-Core-based OLAC standard that utilizes specialized controlled vocabularies for representing attributes of language resources. Two large-scale global bibliographic databases well-known in GLAM community – ArchiveGrid with over 7 million metadata records from 1400 archival institutions and WorldCat with over 586 million records (over 2.3 million for archival resources) – aggregate metadata that represents language archives materials.

LITERATURE REVIEW

As GLAMs most often steward language archive resources (individual items and collections), metadata to represent them is usually created by GLAM professionals, with or without ability to consult with the members of community whose content is deposited or researchers (linguists, anthropologists, historians) who interacted directly with language community members in creating these archival resources (Burke, 2021). Generating high-quality discovery-supporting metadata for legacy materials – collected 30+ years ago in analog form and recently deposited to archives – is the most challenging because information professionals must rely on their own expertise and interpretation of these materials' descriptive context, with those who created (or contributed to creating) archival resources usually no longer around. Tricky metadata creation tasks with legacy materials include but are not limited to the ones related to tracking provenance (Huber, 2023; Weber, 2022).

To support descriptive metadata work, GLAM profession developed conceptual models such as FRBR and LRM (International Federation of Library Associations and Institutions, 1998; 2008; 2017). Academic programs that prepare information professionals have been covering these models and their application for decades: in introductory core information organization courses in programs accredited by American Library Association (ALA) and in advanced specialized courses that focus on library cataloging with MARC and BIBFRAME and digital repository metadata creation with Dublin Core, Encoded Archival Description, etc. FRBR and LRM models define information resources' attributes and relationships to be represented in metadata to enable discovery. Language materials, especially those that are unique to language-focused archival

collections, have specific attributes and sets of complex relationships some of which are not captured in GLAM models (Paterson & Coronado, 2025). Description models developed by linguists are unfamiliar to most GLAM practitioners and students and do not fit with the one-to-one principle in metadata creation whereas each information resource must be represented by a separate metadata record (DCMI, n.d.; Urban, 2014), as opposed to representing a "bundle" of information resources together (Max-Planck Institute of Psycholinguistics ISLE Metadata Initiative, 2001). Information professionals need training on identifying and representing in metadata those attributes and relationships of resources important for the language archives users: language speakers, instructors and learners of language and culture on one side, and linguistics and anthropology researchers, educators, and students on another side. Without such training provided until recently, metadata in digital language archive and community archive collections is often lacking in supporting information discovery (e.g., Al Smadi et al., 2016; Wasson, Holton, & Ross, 2016).

To adequately handle descriptive tasks for making an information resource discoverable and reusable, information professionals need to clearly understand and actively consider user needs in both metadata creation and metadata management processes (such as selection of metadata schemes and controlled vocabularies, quality assurance, conversion, etc.). One of the main contributions of FRBR and LRM models is the user-centered approach, where user tasks of *find*, *identify*, *select*, *obtain*, and *explore* are the focus of describing information resources to support user needs. FRBR also maps metadata fields of MARC metadata records to user tasks, providing the way to evaluate the level of metadata records' support of user tasks in metadata quality assurance. Examination of language archives' metadata records with the focus on these user tasks demonstrates that while the *find* user task is overall well supported, support for other four user tasks, especially *obtain* and *explore*, requires strengthening (Zavalin, 2023).

User studies also reveal other areas needing improvement in language archive resources representation. For example, for many users' interactions with language archives, representations of the specific dialect(s) covered by the archival item and relationships between items (e.g., an audio recording and its text transcript) are the most important, yet sometimes lacking in metadata records (Burke et al., 2021, 2022a). Likewise, user studies participants point out the importance of multilingual interfaces and multilingual metadata in enabling broader access to digital language archive resources. Language archive users also emphasize the need for providing map-based geographic browsing option (which only can be enabled with robust geographic representation in metadata) for improved navigation and overall user experience (Burke et al., 2021).

Some publications share the experiences of information professionals working with community members on stewarding digital language archives and best practices emerging from such cooperation. Dale (2022) discussed developing the mediated archiving workflow. Burke and colleagues (2022b) covered challenges and solutions in representing aboutness and South Asian names in metadata. Based on the interviews with digital language archive depositors, Chelliah (2023) developed recommendations for revising the wording of metadata guidelines to make them more accessible to community representatives and the requirements in metadata guidelines for better representation of photographs.

Researchers examining user satisfaction with digital language archive services, including but not limited to metadata, agree that GLAM education needs to cover the digital curation and stewardship of language archival collections. Coverage of user preferences, as well as best practices, in GLAM metadata training will ensure that information professionals responsible for metadata in language archives are prepared to meet community user needs. The preliminary outcomes of the recent project that addresses this curricular need mainly through developing and piloting at University of North Texas (UNT) the broader-scope introductory GLAM graduate course entirely focusing on community language archives – with basic metadata-related knowledge and skills development included but not prioritized in learning objectives – were reported by Coronado and Zavalina (2024), Zavalina and Paterson (2024), Zavalina, O'Neil, and Chelliah (2025). Following Zavalina (2023) early report, this paper presents another project addressing this curricular need as part of the advanced metadata course and covers the evolution through 2025.

PROJECT DESCRIPTION

Our interdisciplinary team began this work with the one-semester experiment developing and teaching digital language archives course to a combined graduate class of UNT GLAM and Linguistics students (Zavalina & Chelliah, 2021). Based on these experimental results, to maximize curricular benefits, we decided to develop the learning module focusing on digital language archives and community archives for the existing advanced elective GLAM graduate course on digital library metadata (INFO5224) that had been part of the initial experiment.

In 2020-2021, we developed new learning module and began revisions of INFO5224 three other modules by integrating examples from language archives in lectures and assignments. The preliminary results were tested in the Spring 2021 course offering. Since then, with continuous revisions, INFO5224 was taught four more times – in Spring semesters of 2022-2025 – with a cumulative enrollment of 65 students. To qualify for INFO5224 enrollment, students must successfully complete an ALA-accredited program's core course: the fundamentals of information organization. Another prerequisite is successful completion of UNT's introductory digital repository metadata course INFO5223, with exceptions granted occasionally based on relevant undergraduate coursework, work experience, or concurrent enrollment. In INFO5223, students develop competencies related to application of major metadata standards. This includes use of data content standards (e.g., DACS, CCO, RDA), data value standards (controlled vocabularies such as LCSH, LCNAF, TGN, AAT, ULAN), data encoding and transmission standards (mainly XML, with introduction to HTML and MARC), and major descriptive metadata element sets for item-level metadata creation (Dublin Core DCTERMS, MODS, VRA Core 4.0) and collection-level metadata creation (Dublin Core Collection Application Profile, Encoded Archival Description, MODS collection application profile, and use of VRA Core 4.0 collection record type). INFO5223 learning materials discuss the role of user needs in developing metadata element sets and controlled vocabularies, as well as in providing information access, with examples.

Prerequisite coursework prepares students to closely examine metadata principles and tools in relation to community language archives in the advanced course INFO5224. With the course

regularly offered in 16-week semesters, the class spends an equal amount of time (4 weeks) on each of INFO 5224 four learning modules (see for example, the Spring 2024 INFO 5224 syllabus and schedule). In the weekly online class meetings, the instructor presents material in an interactive way, with numerous illustrative examples drawn from language archives. Class meetings include brainstorming activities and mini exercises to help digest the content and develop competencies.

During the first week of each learning module, students participate in the class meeting (or review posted instructor presentation slides and recording) and select and read 2 items from the list of 20 or more relevant peer-reviewed professional and/or research publications. Students then critique these readings in the discussion post by addressing the following questions:

1. Which ideas/approaches discussed in the readings do you agree with and why?
2. Which ones do you find unexpected or disagree with and why?
3. Which situations from your experience as user and/or creator of metadata can serve as illustrations to topic(s) raised in the readings?

Students read each other's discussion posts and react to them, with the expectation of providing a reply of substantial length to at least one.

Two out of 5 INFO5224 course-level learning objectives are explicitly user-centered:

- Identify the needs of a user community, types of materials of interest to these users, 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, applying controlled vocabularies.
- 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.

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

1. Identify:
   - information needs and user tasks of digital language archives end-users and depositors (language speaker communities, documentary linguists, language instructors and researchers)
   - types of materials of interest to digital language archive end-users and depositors
   - general and specific metadata standards that can be utilized in representing digital language archive materials.
2. Implement this knowledge in:
   - investigating relations between metadata elements and digital language archive user tasks based on FRBR and LRM models
   - navigating controlled vocabularies and selecting appropriate terms for representing digital language archive materials.

3. Examine and evaluate:
    - Current trends in digital language archive metadata theory and practice
    - Perspectives of developing and applying metadata to provide effective information access for digital language archives users.

During weeks 2-4 of each module, INFO5224 students complete major practical assignments. Module 1 assignment includes two parts. In Part 1 (*Language Materials, their Users, User Tasks, and Metadata*), students answer 5 blocks of questions (with between 2 and 4 interrelated questions in each block) based on 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 conceptual models: FRBR and LRM. In Part 1, students also learn to apply ALA Romanization table for transliterating the title of an information resource in non-Latin script language. Part 2 (*General and Specialized Controlled Vocabularies for Representing Resources in Language Collections to Facilitate Information Access*) currently includes 18 blocks of information-hunt style questions – between 2 and 4 per block. In this part of the assignment, students navigate 15 data value standards (including 6 specialized controlled vocabularies for language resources representation that are not covered by mainstream GLAM metadata coursework) to find entries for terms, names, and codes relevant for representing digital language archive materials, interpret and apply the authorized forms.

Although only the 1st learning module entirely focuses on digital language archives, including community archives, students are engaged with language archives and their users' needs throughout the semester. In the instructor presentations and class meetings of the remaining 3 learning modules – *Metadata Quality*, *Metadata Interoperability*, and *Linked Data* – examples from digital language archives are used as much as possible. The Module 2 *Metadata Quality* practical assignment also has a significant community language archive component: students collect and analyze a small sample of metadata records representing photographs in community language collections of Computational Resources for South Asian Languages archive based on three major criteria of metadata quality – accuracy, completeness, and consistency – first defined in Bruce and Hillmann (2004) framework widely used in metadata quality assurance and research and refined and operationalized in several studies (most recently, in Zavalin & Zavalina, 2025). Students then compare the results of this evaluation to those for another collection's metadata sample. In this work, students consult the published collection-specific metadata creation guidelines and consider how the metadata quality affects user experiences. The final component of Module 2 practical assignment includes using generative Artificial Intelligence tool of student's choice to automatically generate Dublin Core metadata for a familiar information resource – one of the readings about digital language archives used in discussion posts – and evaluate resulting metadata quality using the same criteria.

While not as significant as in Module 2 practical assignment, digital language archive-related problems and questions are gradually being integrated into two other practical assignments. In Module 4 assignment, students convert Dublin Core and MODS metadata records representing a paper about digital language archives from plain XML into Linked-Data-supporting encodings

using RFDXML and JSON. Also, the small instructor-developed thesaurus that students convert into an OWL-encoded mini-ontology in Module 4 exercise is focused on organizing information in digital language archives. For the next INFO5224 offering, the metadata record that students harvest, convert using MarcEdit metadata management tools, and evaluate resulting metadata effectiveness for supporting user needs in Module 3 assignment, will be the one representing a typical community language archive item.

CONCLUSION

Digital language archives, including community archives, are growing fast and becoming prominent in the information landscape. GLAM education needs to catch up with these developments and equip students with competencies necessary to steward such archives for the benefit of their user communities and ensure discoverability of their resources through functional metadata. This paper presents analysis of professional literature related to metadata challenges and solutions for digital language archives in relation to addressing user needs and supporting user tasks. It follows with presenting the project developing and refining graduate curriculum to enable such training for information professionals. The project aims to help bridge the gap in GLAM profession's understanding of community language archives users and their needs, specifics of attributes and relationships of language archive resources, and metadata needed to adequately represent them. This report will be useful for other GLAM educators working on community-focused metadata curriculum development.

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

Bruce, T. R., & Hillmann, D. I. (2004). *The continuum of metadata quality: defining, expressing, exploiting*. ALA editions. Retrieved from https://hdl.handle.net/1813/7895

Burke, M. (2021). Collaborating with language community members to enrich ethnographic description in a language archive. *Proceedings of the International Workshop on Digital Language Archives: LangArc-2021* (pp. 18-21). https://doi.org/10.12794/langarc1851172

Burke, M., Chelliah, S. Zavalina, O. L., & Phillips, M. E. (2022a). User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. *Language Documentation & Conservation, 16*, 1-24. Retrieved from http://hdl.handle.net/10125/74669

Burke, M., Tarver, H., Phillips, M.E., & Zavalina, O. (2022b). 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

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

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 Archives: LangArc-2023* (pp. 8-10). University of North Texas. https://doi.org/10.12794/langarc2114301

Coronado, S. I., & Zavalina, O.L. (2024). Digital language archiving: Reuse and adaptation of non-LIS learning materials in LIS education. In J. Krammer & M.S. Park (Eds.), *The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024* (pp.1-9). https://doi.org/10.21900/j.alise.2024.1736

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

Dublin Core Metadata Initiative (DCMI). (n.d). *One-to-One Principle*. Retrieved from https://www.dublincore.org/resources/glossary/one-to-one_principle/

Frederick, M., Roeschley, A.K., & Zavalina, O.L. (2025). Beyond language archives: Proposing the archival community informatics framework as an interdisciplinary link to revitalization lexicography. In I. Sserwanga, M.R. Sanfilippo, C. Inskip et. al. (Eds.), Living in an AI-gorithmic World: 20th International Conference, iConference 2025, Proceedings (pp. 1169-1179). https://doi.org/10.47989/ir30iConf46942

Huber, C. (2023). Why it can be difficult to make historic language recordings accessible: A view from a corpus of historic dialect recordings. *Proceedings of the International Workshop on Digital Language Archives: LangArc-2023* (pp. 15–18). University of North Texas. https://doi.org/10.12794/langarc2114302

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

International Federation of Library Associations and Institutions. (1998). *Functional Requirements for Bibliographic Records: Final Report*. K.G. Saur. Retrieved from https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr.pdf

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

International Federation of Library Associations and Institutions. (2017). *Library Reference Model*. Retrieved from https://repository.ifla.org/handle/20.500.14598/40.2

Max-Planck Institute of Psycholinguistics ISLE Metadata Initiative. (2001). *IMDI Part 3: Vocabulary Taxonomy and Structure*. Retrieved from https://archive.mpi.nl/forums/uploads/short-url/c07lIVbfaLy2LYCGFpToQnhvddS.pdf

Paterson III, H.J., & Coronado, S.I. (2025). *Applying Library Science Models to the Arrangement of Language Collections*. Paper presentation at the 9th International Conference on Language Documentation and Conservation (ICLDC). University of Hawai‘i at Mānoa. March 6–9, 2025. Retrieved from https://docs.google.com/presentation/d/11phPLuZXuTi1P0etz5D6R6PDCJhtHxPK

Urban, R.J. (2014). The 1:1 principle in the age of linked data. *Proceedings of the International Conference on Dublin Core and Metadata Applications* (pp. 1-10). Retrieved from https://dcpapers.dublincore.org/article/952136464

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

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

Zavalin, V. I. (2023). Ukrainian archival metadata in WorldCat: Exploratory analysis. In *Proceedings of the International Workshop on Digital Language Archives: LangArc-2023* (pp. 38–41). University of North Texas. https://doi.org/10.12794/langarc2114298

Zavalin, V.I., & Zavalina, O.L. (2025). Assessing comparative effectiveness of metadata creation with specialized and general standards for representing artworks. *Journal of Library Metadata, 25* (3), 153-183. https://doi.org/10.1080/19386389.2025.2499343

Zavalina, O. L. (2023). Language archiving training: A case study, 2020-2023. *Proceedings of the 2nd International Workshop on Digital Language Archives, ACM/IEEE Joint Conference on Digital Libraries.* https://doi.org/10.12794/langarc2114299

Zavalina, O. L., & Chelliah, S. (2021). Exploring language archiving education for information professionals and interdisciplinary collaboration to support information access. *ALISE 2021 Proceedings: Crafting a Resilient Future: Leadership, Education, & Inspiration*, September 20-24, 2021, Virtual. IDEALS. https://hdl.handle.net/2142/110944.

Zavalina, O. L., O'Neil, A.C., & Chelliah, S.L. (2025). Stewardship of digital language archives: Training development and testing through collaboration of information scientists, linguists and communities. In I. Sserwanga, M.R. Sanfilippo, C. Inskip et al. (Eds.), *Living in an AI-gorithmic World: 20th International Conference, iConference 2025, Proceedings* (pp.435-442). https://doi.org/10.47989/ir30iConf47311

Zavalina, O.L., & Paterson III, H.J. (2024). Developing graduate curriculum for digital language archive stewardship. In, J. Krammer & M.S. Park (Eds.), *The Ethics and Evolution of Truth and Information: Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2024* (pp.1-13) https://doi.org/10.21900/j.alise.2024.1657