Leveraging Digital Library Infrastructure to Build a Language Archive

Mark Edward Phillips
Libraries
University of North Texas
Denton, Texas
mark.phillips@unt.edu
0000-0002-9679-6730

Mary Burke
College of Information
University of North Texas
Denton, Texas
mary.burke@unt.edu
0000-0002-6498-6820

Hannah Tarver
Libraries
University of North Texas
Denton, Texas
hannah.tarver@unt.edu
0000-0003-2344-9268

Oksana L. Zavalina
College of Information
University of North Texas
Denton, Texas
oksana.zavalina@unt.edu
0000-0002-3354-4923

***Abstract*—Building a digital language archive requires a number of steps to ensure collecting, describing, preserving, and providing access to language data in effective and efficient ways. The Computational Resource for South Asian Languages (CoRSAL) group has partnered with the University of North Texas (UNT) Digital Library to build a series of interconnected digital collections that leverage existing UNT technical and metadata infrastructure to provide access to data from and for various language communities. This article introduces the reader to the background of this project and discusses some of the important for representing language materials areas where UNT metadata has needed flexibility to better fit the needs of intended audiences. These areas include a workflow for standardized language representation (the Language field), defining roles for persons related to the item (Creator and Contributor fields), and representing interconnections between related items (the Relation field). Although further work is needed to improve language data representation in the CoRSAL digital language archive, we believe the model adopted by our team and lessons learned could benefit others in the language archiving community.**

***Index Terms*—metadata, language archives, digital libraries, controlled vocabularies**

I. Introduction

Over the past three years the collaborators at the University of North Texas (UNT) – the College of Information's Department of Linguistics and the Department of Information along with the Digital Libraries Division of the UNT Libraries – have worked to create the Computational Resource for South Asian Languages or CoRSAL. This program seeks to collect, describe, preserve, and provide access to language data and related artifacts from the South Asian region of the world. Initially, two collections from UNT Linguistics faculty (Lamkang Language Resource and Burushaski Language Resource) were uploaded; CoRSAL now accepts deposits from researchers and language community members. A key component of the CoRSAL program is a digital archive that has been built upon existing technical and metadata infrastructure in the UNT Libraries' Digital Collections. While creating the CoRSAL collection in the UNT Digital Library (one component interface of the Digital Collections), the project team has discovered information about metadata modeling and creation that we believe would be beneficial to the wider community. A selection of these lessons learned are presented below.

II. Background

The UNT Libraries' Digital Collections use a uniform locally-developed metadata scheme (UNTL) to describe items regardless of material type, owner, or collection. UNTL is based on the Dublin Core standard, with additional local fields for a total of 21 fields, 14 of which are locally-qualified. Over time, we have developed extensive guidelines (https://library.unt.edu/digital-projects-unit/metadata/input-guidelines-descriptive/) providing usage information and example data values for each of the fields across different material types. For some large, ongoing collections, we also create more specific metadata guidelines that clarify specifics of field usage, state which information applies from the general guidelines, and provide relevant examples.

As we built the CoRSAL collections, it became apparent that specialized metadata instructions would be useful. These were subsequently developed by CoRSAL staff, based on experiences describing the first two CoRSAL collections, with input from those who have archived language data in the past. Because CoRSAL prioritizes deposits from community language documenters, guidelines are intended to be readily understood by first-time metadata creators. Depositors are given a template with examples of completed metadata from other collections. The metadata guidelines development process took into account the relevant attributes of the data typical for language archive deposits: language(s), genre, roles of contributors and creators, and the relationship between items (e.g., between an audio and its transcript; original text and its translations). Though subject representation is not typically emphasized in language archive metadata [1], the CoRSAL metadata creation guide does encourage depositors to include keywords about the content or topic of the items. Finally, templated content descriptions are included to provide examples to depositors.

III. Language-Specific Metadata Usage

Currently there are twelve distinct CoRSAL collections in the UNT Digital Library. This integration process has provided structure to the wide range of language data that is being deposited as well as providing a process for unifying resource description across collections to improve discovery.

For this workshop paper we will focus on three primary areas: language representation, agent roles, and item relations.

A. Language Field

The UNTL metadata schema has a locally-developed controlled vocabulary for language codes (https://digital2.library.unt.edu/vocabularies/languages/), displayed as a drop-down list for editors. Designating new languages and codes has happened organically as material was added to the collection. Because the scope of content being collected and digitized was primarily focused on English-language resources, the language vocabulary grew slowly. Language codes were historically based on ISO 639-3 three letter codes and designated names. As the UNT Digital Library began adding CoRSAL collections, it became clear that this approach would not always work for language data, especially language documentation data. There were two challenges that came up with the existing approach to language codes and names. First, what happens when the language is not present in the ISO 639-3 language list, and second, what should happen when the "standard" language name assigned by the organization maintaining the standard is not preferred or accepted by the language community?

After discussion, a new process was developed and has been in place for the past year. First, administrators check the ISO 639-3 language code and add that version to the local vocabulary. If the ISO 639-3 language code is not present, the Glottolog (https://glottolog.org/) is used a source of language code. Glottolog describes itself as the "Comprehensive reference information for the world's languages, especially the lesser known languages." Languages in Glottolog have unique identifiers, called Glottocodes, which are added to the local vocabulary with the primary language name [2]. If a language code is not present in either source, the CoRSAL archive team will work with the depositing researcher to submit a request and provide supporting documentation to register the language with Glottolog.

As an example, CoRSAL archive added the Azamgarhi Language Resource collection, however, "Azamgarhi" did not yet have an established identifier in any standard language list. To avoid future confusion, instead of using a near match or the code for the larger group of languages (east2875), CoRSAL staff applied for a Glottocode for this variety (azam1235). This provided a standard code so that the materials could be ingested with a controlled form of the language representation.

In situations where the language community does not recognize the "standard" language name used in the ISO 639-3 documentation [3], the UNTL system has the flexibility to use more acceptable technology and document multiple versions of the name. For instance, the ISO 639-3 code *lus* is based on the language name 'Lushai' which is now called Mizo. While the UNTL language code is *lus* to match the ISO 639-3 code, the language name is represented as 'Mizo.'

B. Creator and Contributor Roles

In representing creators and contributors, the UNTL metadata scheme takes an agent-based approach (i.e., "who made this item") rather than a role-based approach (i.e., "who filled each of these roles in creating an item"). Each agent is assigned a primary role describing their specific contribution to create or steward the item, based on MARC Code List for Relators with some local additions (https://digital2.library.unt.edu/vocabularies/agent-qualifiers/). This makes sense given the wide array of material types and roles, but it means that an entity (person or organization) can only be listed once per record across the creator/contributor fields.

For materials where individuals have multiple roles, it may be challenging to determine which role is the "primary" way that they contributed to the item. For example, the same individual may have transcribed an audio recording and then translated the content into English. In this case, both the Transcriber and Translator roles are applicable. It is possible to represent both, because additional roles and clarifications can be added in an optional Info subfield of the Creator and Contributor fields that displays to users and is searchable, so no information is lost.

Also, there is not always a consensus on role terminology between the information professionals and depositors in the language archiving community. For instance, the term Analyst is defined by MARC Code List for Relators as "a person or organization that reviews, examines, and interprets data or information in a specific area." However, this term is commonly understood by documentary linguists as referring to a person or group that specifically provided linguistic analysis of language data. This difference in interpretation highlights the need for collaboration and development of common understanding of terminology, and possibly extensions to existing OLAC (Open Language Archives Community) controlled vocabularies.

C. Relationships between Items

The CoRSAL collections provide access to a wide range of linguistic data, represented in formats such as audio and video recordings; transcriptions; translations; photographs of cultural events, local flora and fauna; field notes; and collected publications and writings in a given language. Practice in the Digital Collections is to describe each discrete component piece as a separate object – allowing for clear and accurate description of creation information – however, the UNTL system has a robust process for describing relationships between resources, leveraging the Relation metadata field and the available qualifiers (https://digital2.library.unt.edu/vocabularies/relation-qualifiers/). This allows researchers to find specific types of items (e.g., only videos) as well as intellectually-related materials (e.g, a transcript, translation, etc.). Additionally, the UNT Digital Library interface provides features to draw attention to related resources with visual cues (see Figure 1 and Figure 2).

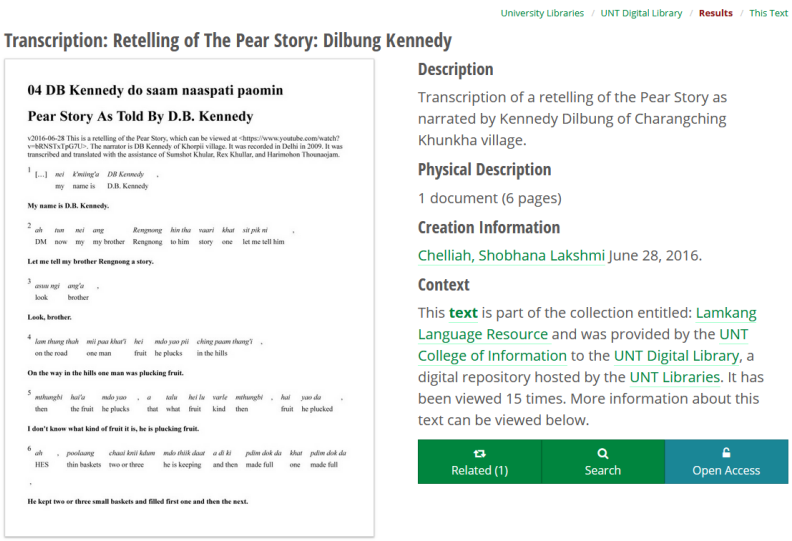

Fig. 1. Example item-level metadata with related items indicated.

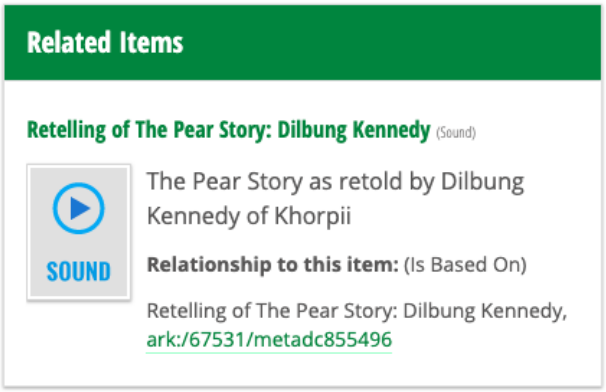

Fig. 2. Example representation of relationship from transcription to original recording.

IV. Conclusion

Although the UNTL metadata scheme is not always a perfect match for the CoRSAL digital language archive collections, since it is not specific to language-based data, it has been easily adapted to these kinds of materials in most cases. We have been able to develop new processes to address specialized concerns (e.g., those related to language names) and are engaging in continuing discussions regarding the best way to handle other issues to ensure robust description that meets the needs of both researchers and the wider, global internet audience.

With any metadata implementation, there is the need for user studies to determine the level of usability for the end-users and the areas of weakness to be addressed. A study focusing on the CoRSAL interface and metadata will help develop a robust understanding of the users' experience when interacting with the digital language archive, and get ideas for potential improvements to future metadata.

Overall, adding CoRSAL collections to the UNT Digital Library has provided a relatively easy way to make materials findable and available to other users while making use of the existing infrastructure and the UNTL metadata schema. While it does require some flexibility and logistical planning, this model and the general success in providing access to these materials show that a similar approach may allow more language researchers to make their materials available online for reuse.

References

[1] M. Burke, O. L. Zavalina, M. E. Phillips, and S. Chelliah, "Organization of knowledge and information in digital archives of language materials," *Journal of Library Metadata*, pp. 1–33, 2021. [Online]. Available: https://doi.org/10.1080/19386389.2020.1908651

[2] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank, "glottolog/glottolog: Glottolog database 4.4," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4761960

[3] S. Morey, M. W. Post, and V. A. Friedman, "The language codes of iso 639: A premature, ultimately unobtainable, and possibly damaging standardization," 2013-01-01. [Online]. Available: http://hdl.handle.net/2123/9838