# Utilizing existing metadata standards and tools for a digital language archive: a balancing act

Mary Burke
*Department of Linguistics, University of North Texas, Denton, Texas, USA*

Hannah Tarver and Mark Edward Phillips
*Department of Libraries, University of North Texas, Denton, Texas, USA, and*

Oksana Zavalina
*Department of Information Science, UNT, Denton, Texas, USA*

Received 5 February 2022
Revised 14 April 2022
Accepted 27 May 2022

## Abstract

**Purpose** – 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.

**Design/methodology/approach** – This paper introduces the reader to the background of this project and discusses some of the areas important for representing language materials where both UNTL metadata and CoRSAL metadata practices were adapted to better fit the needs of intended audiences.

**Findings** – 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 including subject representation for language materials (Subjects and Keywords fields).

**Practical implications** – Although further work is needed to improve language data representation in the CoRSAL digital language archive, the model adopted by the team and lessons learned could benefit others in the language archiving community.

**Originality/value** – This paper is a significantly extended version of the presentation made at the 1st International Workshop on Digital Language Archives in 2021.

**Keywords** Digital libraries, Metadata, Institutional repositories, Controlled vocabularies, Language archives

**Paper type** Case study

## 1. Introduction

The University of North Texas (UNT) Libraries' Digital Collections comprise more than 3.2 million items including text, images, audio/video materials, large-format maps and technical drawings, biological specimens and data sets. These are hosted in a single digital infrastructure including components for archival storage and public access with permanent links. Users access materials via three public interfaces: The Portal to Texas History (cultural heritage materials from institutions in Texas), the Gateway to Oklahoma History (similar materials from the Oklahoma Historical Society and partnering institutions) and the UNT Digital Library (materials owned by the university or created by faculty and staff). UNT Libraries' Digital Collections use UNTL metadata, a locally developed metadata scheme. This article expands on Author (2022) to discuss how both UNTL metadata and

The Electronic Library

0264-0473
DOI 10.1108/EL-02-2022-0028

Computational Resource for South Asian Languages (CoRSAL) metadata practices were adapted to better fit the needs of our intended audiences.

## 2. Literature review

Language archives serve as repositories of data related to languages and cultures, including indigenous language and culture. An important difference between a language archive and a simple repository or corpus is that it allows for the discovery of data via searching and browsing through a database of metadata, that is, by supporting the common user tasks of find, identify, select, obtain and explore, identified in the internationally recognized *Library Reference Model* [International Federation of Library Associations and Institutions (IFLA), 2017]. In a physical archive, after finding metadata records, examining them to identify and select resources of interest and exploring related resources, the user accesses resources in analogue form in the archive building or requests a physical copy. Digital archives (including language archives) not only have surrogate descriptions in the form of metadata but also provide seamless online access to digitized or born-digital materials, regardless of the user's location.

As noted by Henke and Berez-Kroeker (2016) and Berez-Kroeker *et al.* (2017), data accessible in language archives is important for enabling indigenous language revitalization, as it typically includes audio-visual materials and their transcriptions, translations and linguistic annotations for languages facing extinction. These languages are categorized as vulnerable, threatened, endangered, severely endangered and critically endangered on metrics, such as the *Language Endangerment Index*, based on factors, such as transmission from generation to generation, current language speaker numbers and related trends, and domains of language use (Belew and Simpson, 2018; Lee and Van Way, 2016). AQ: 9

Most data collected by linguists was not traditionally shared, other than through secondary resources (e.g. journal articles, conference presentations, etc.). Source data was collected by researchers and shared within a team of researchers or with individual linguists upon request. Development of dedicated digital language archives in the early 2000s and their popularization through awareness programs and training – notably, by the Open Language Archives Community (OLAC) project – changed the situation (Bird and Simons, 2021). Regional language archives (e.g. the California Language Archive, AK Native Language Archive) and larger-scale language archives with collections from multiple continents (e.g. the Endangered Language Archive [ELAR], the Pacific and Regional Archive for Digital Sources in Endangered Cultures [PARADISEC] and Archive of Indigenous Languages of Latin America [AILLA]) emerged during that time. However, the real growth of digital language archives is often attributed to 2011, when the United States National Science Foundation began to require that data generated through its "Documenting Endangered Languages" funding program be made publicly available [National Science Foundation (NSF), 2018].

Creation of descriptive metadata records that make these materials discoverable and reusable is an integral part of the deposit. Prior to language data archiving, the linguistics community used the term *metadata* in a very different sense: to refer to secondary resources, such as linguistic annotations and markup of the primary language resources. According to Good (2002), to avoid confusion among linguists and in communications between linguists and information professionals, the linguistics community started differentiating between metadata designed to facilitate resource discovery in databases vs annotations, transcriptions and markup of primary data. These two kinds of metadata are called *thin* and *thick* metadata, respectively, in Nathan and Austin (2004).

Digital language archives can be hosted by a library or museum, by a university, or by other nonprofit or commercial organizations. The type of repository often determines the overall structure of the archive, metadata schemes and descriptive units (e.g. DSpace or ContentDM digital content management applications that are used by the library's digital collections as a whole, are normally also used by a digital language archive hosted by this library). While traditional archives prioritize hierarchical descriptions (Abraham, 1991), libraries represent each item with its own metadata record and link the related items through data values. Locally developed language archives take a hybrid approach, where related items are grouped together in a bundle: an audio or video recording, textual transcriptions and translations of the content and multiple possible derivatives (e.g. handwritten field notes, grammatical annotations). In many language archives, a single item-level record describes the entire bundle. To accommodate these practices, the OLAC created a specialized metadata application profile and controlled vocabularies that draw focus to unique attributes of language data – subject language, language family, discourse type and so on – used in OLAC aggregation of metadata from over 60 language archives [Bird and Simons, 2003; Open Language Archives Community (OLAC), 2011].

Some studies have evaluated how descriptive metadata supports information discovery and access and the quality of language archive metadata. One early example is the OLAC repository's metadata ratings based on use of required elements – Title, Description, Subject, Date and Identifier – and utilization of controlled vocabularies for representing language attributes (e.g. discourse type and specific roles of creators and contributors of language resources) (Hughes, 2005). Recent metadata quality research in language archives explores the Description field. For example, Harris *et al.* (2019), in a user study with language community representatives, identified gaps and errors in metadata for the Papua New Guinea language resources in PARADISEC. These findings informed enrichments to metadata records by adding missing contextual information. Author and Author (2022) comparatively analyzed Description field values in PARADISEC, AILLA and ELAR and the archives' metadata creation guidelines for depositors.

In language archives, metadata creation is often the responsibility of a linguist who collected the language data or a language community member who deposits the data on behalf of their community in a self-deposit or self-archiving process (Hanard, 2001). Depositors rely on guidelines made available by the archive. Subject to resource availability, some archives offer a mediated deposit process, under which the depositor and archival staff create metadata collaboratively (Tillman, 2017). Research shows that a mediated approach increases deposits and decreases barriers to researchers contributing to the repository: the amount of time required and lack of confidence and familiarity with the interface and metadata creation (Daoutis and de Montserrat Rodriguez-Marquez, 2018). A comparative analysis found that metadata in mediated deposits was more complete, consistent and accurate across the board than depositor-created metadata (Kurtz, 2010).

In recent years, language archives are seeking to be accessible to language communities as well as academic audiences (Czaykowska-Higgins, 2009; Henke and Berez-Kroeker, 2016; Woodbury, 2014). As the most common depositors to language archives are linguists, who are not necessarily members of the language community or culture in question, the metadata provided may lack context that is relevant to potential language community users. One manifestation of the effort to increase accessibility is the development of the Tromsø recommendations for citation of research data in linguistics (Andreassen *et al.*, 2019). In addition to establishing citation practices for published materials, Andreassen *et al.* (2019) provided templates for citing archival materials which include all individuals who contribute to generating language material, such as the following example:

> Krauss, Michael E. (Interviewer), Jeff Leer (Interviewer) and Anna Nelson Harry (Speaker). 1975. Interview with Anna Nelson Harry. In Krauss Eyak Recordings, item ANLC0082. Alaska Native Language Archive, available at: www.uaf.edu/anla/. (Andreassen *et al.*, 2019, p. 7)

Gawne *et al.* (2021) noted that researchers are often listed as the creators of items, but language community members (speakers of the language) are not included in the metadata. As these citation practices become more common, metadata creators may provide more complete information about all individuals involved in generating language materials (depending on privacy concerns).

A focus group with language archive stakeholders conducted by Wasson *et al.* (2016) revealed the following perceived barriers for digital language archive use:

- lacking contextual information and annotations;
- lacking opportunities for users to engage with the material;
- lacking culturally relevant categories;
- inability to effectively find data with existing search, browse and display options; and
- issues with accessibility (in terms of interface language, terminology and technology).

In addition to these issues, Author *et al.* (2022) found that language archive managers and depositors were dissatisfied with the current range of controlled vocabulary options. Virtually all of these issues relate to how materials are represented in metadata and presented to users. In this paper, the authors discuss their experiences to address some of these issues by adapting the available resources and tools – such as the UNTL metadata scheme, controlled vocabularies, and metadata practices – and by attempting to balance them with the needs and expectations of intended audiences of the CoRSAL digital language archive. AQ: 10

## 3. UNTL metadata and Computational Resource for South Asian Languages

All records in the UNT Libraries' Digital Collections use a uniform metadata scheme (UNTL) based on the Dublin Core standard with added local fields and qualifiers for more specificity and greater flexibility. UNTL has 21 fields including eight that are required: main title, language, content description, subject (2), resource type, format, collection and institution. See Author (2022, p. 3) for a comprehensive listing of fields. Although these required fields are considered the absolute minimum expectation for a usable record (https://library.unt.edu/digital-projects-unit/metadata/minimally-viable-records/), they do not reflect all of the information that is known – or important – for academic research. For example, required values do not include creators, which are critical for certain collections and material types (e.g. theses/dissertations) (Digital Library Assessment Interest Group's Metadata Assessment Working Group, 2021). Language materials generally include the names of a native speaker and/or researcher, which distinguish among different participants or research teams.

Local qualifiers provide options for representing different values, such as multiple dates (e.g. when a text was published vs submitted or accepted) or reflecting different aspects of the coverage (e.g. places vs dates associated with the content) with an appropriate label to designate the *type* of value. This allows similar types of information (e.g. dates) to be in a single field rather than providing separate fields for every combination; it also provides an opportunity for expansion by adding qualifiers without altering the underlying schema.

Materials in the Digital Collections come from many organizations and departments at UNT and metadata creation is distributed among a number of editors – more than 1,200 unique editors of varying experience levels since 2009. Over time, the authors have developed extensive guidelines providing usage information and example values for each

field. This includes general, system-wide guidelines (available at: https://library.unt.edu/digital-projects-unit/metadata/input-guidelines-descriptive/) and collection-specific instructions for larger, ongoing projects (available at: https://library.unt.edu/digital-projects-unit/metadata/project-specific-guidelines-documents/) to provide support for editors of different skill levels.

The CoRSAL developed over 2016–2019 through a partnership between linguistics faculty, Dr Shobhana Chelliah, and the UNT Digital Library. CoRSAL (available at: https://digital.library.unt.edu/explore/collections/CORSAL/) began with two collections from UNT linguistics faculty (Lamkang Language Resource and Burushaski Language Resource) and now accepts deposits from researchers and language community members. CoRSAL staff developed specialized metadata instructions documenting decisions and best practices based on the first two collections and on input from those who have archived language data in the past.

Because CoRSAL prioritizes community language documenters, guidelines are intended to be readily interpretable by first-time metadata creators. Depositors are given a metadata sheet with examples of completed records or template-type values, with a focus on language data attributes: language(s), genre, roles of contributors and creators and item relationships (e.g. audio and transcript; original text and translations). Though subject representation is not typically emphasized in language archive metadata (Author, 2022), the CoRSAL guide encourages depositors to include keywords about the content or topic of the items.

## 4. Metadata elements and controlled vocabularies used

The UNT Digital Library contains research materials from a number of disciplines, which makes it challenging to represent materials for both specialists and laypersons. This has been a concern in the Digital Collections for other materials (Author, 2022); however, it is possible to find a reasonable balance for the majority of cases. This section outlines specific choices that have been made for language materials within the UNTL framework.

### *4.1 Object representation*

In the Digital Collections, discrete items are described separately (e.g. individual photographs or documents), rather than grouping items as bundles, to describe the creation and content of each item accurately (e.g. authors of different translations or analyses). It is difficult to represent multiple items in a single record and details may be crucial for certain users or research. Separate records also facilitate finding or limiting information to specific parameters – for example, only transcriptions in the original language – and allows for interfaces that display items according to material types (e.g. images of text vs audio players).

To address item relationships, UNTL records have a qualified relation field to cross-reference items. For example, an audio recording could have a transcription and a translation of the text. Reciprocal relationships in each of the three records would point to the related items (see Figure 1).

F1

In Figure 1, the Manipuri audio recording relates to a Manipuri transcription and linguistic analysis, all visibly linked for users.

### *4.2 Creator and contributor*

In the UNTL scheme, each creator and contributor has a descriptive role (e.g. author or photographer). Role options come from a controlled vocabulary currently consisting of 74 terms (available at: https://digital2.library.unt.edu/vocabularies/agent-qualifiers/) adapted from the MARC Code List for Relators (available at: https://id.loc.gov/vocabulary/relators.html), with local additions for roles that do not have an equivalence (e.g. manufacturer).

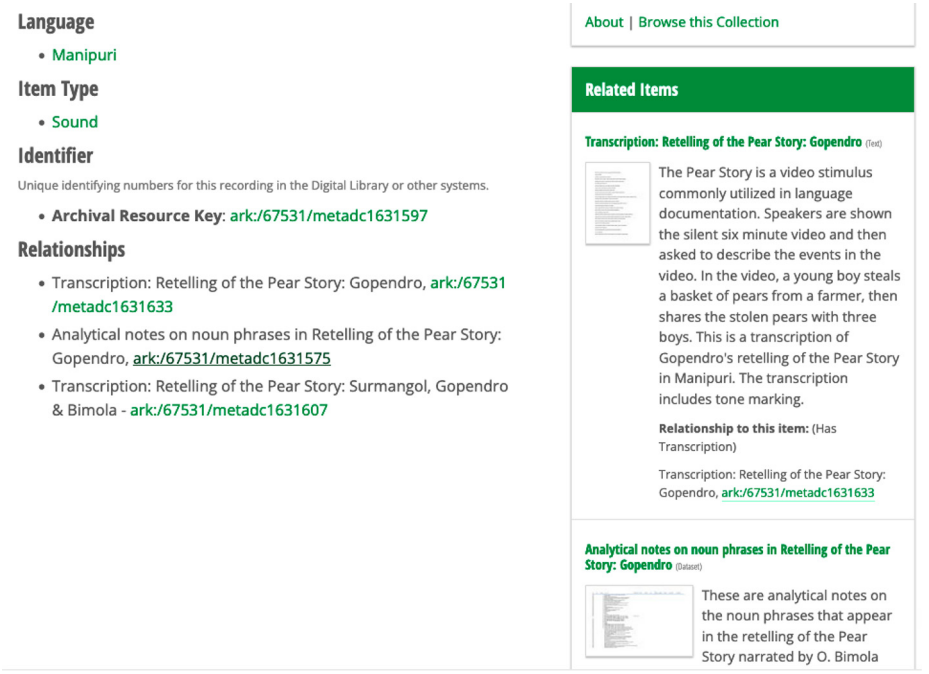


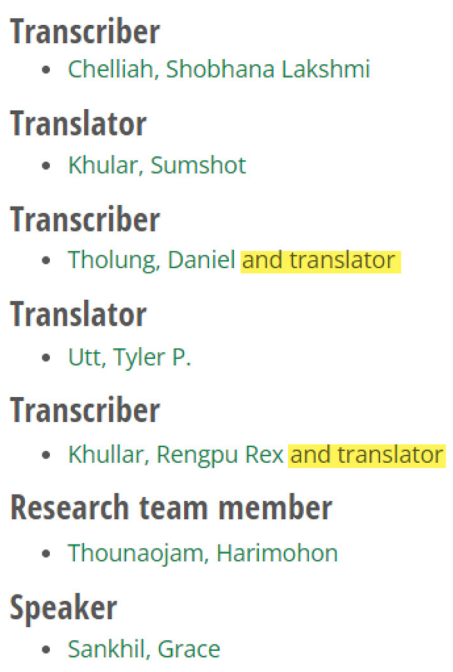

**Figure 1.**
Example of an audio recording record with related items

Despite the range of options, linguistic terminology does not always have a direct correlation. For example, the vocabulary defines *speaker* as "a performer contributing to a resource by speaking words, such as a lecture and speech," and *consultant* as "a person or organization relevant to a resource, who is called upon for professional advice or services in a specialized field of knowledge or training." In the linguistics community, *consultant* would be understood as a *language consultant* or *informant* (i.e. a speaker of the language of interest). The CoRSAL metadata guidelines disambiguate these roles to clarify that *speaker* can be applied to any individual speaking the language, whereas "consultant" more narrowly refers to those providing specialized knowledge (e.g. technical assistance, cultural context) (Computational Resource for South Asian Languages, 2022). A *consultant*, for example, may have identified individuals to participate in interviews and helped design the interview guide, whereas those who participated would be assigned the *speaker* role.

Another difficulty is that in the UNTL schema, an entity (person or organization) can have only one role and be listed once per record. In the context of the Digital Collections, this 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) makes sense, similar to using qualifiers in other fields rather than listing every option. Language materials, however, are commonly created during fieldwork, wherein one individual has many tasks, and many tasks are collaborative. It is typical, for example, for one individual to be both *translator* and *transcriber* of an audio recording and for multiple native speakers to translate. The metadata creator must determine what the primary role is for each person. Additional clarifications can be added in an optional info subelement that displays to users and is searchable (see Figure 2).

Figure 2 shows the creators and contributors for a transcription and partial translation of a CoRSAL audio recording, with added clarifications using the info field.

F2




```
<creator qualifier="trc">
  <name>Chelliah, Shobhana Lakshmi</name>
  <type>per</type>
</creator>
<creator qualifier="trl">
  <name>Khular, Sumshot</name>
  <type>per</type>
</creator>
<creator qualifier="trc">
  <name>Tholung, Daniel</name>
  <type>per</type>
  <info>and translator</info>
</creator>
<creator qualifier="trl">
  <name>Utt, Tyler P.</name>
  <type>per</type>
</creator>
<creator qualifier="trc">
  <name>Khullar, Rengpu Rex</name>
  <type>per</type>
  <info>and translator</info>
</creator>
<creator qualifier="rtm">
  <name>Thounaojam, Harimohon</name>
  <type>per</type>
</creator>
<contributor qualifier="spk">
  <name>Sankhil, Grace</name>
  <type>per</type>
</contributor>
```

**Figure 2.** Example of "info" subelement of creator metadata element used to indicate multiple roles of an individual

*4.3 Titles*

In library, archival and museum communities, data content standards have long provided instructions on constructing titles for resources that do not have one. For example, *Describing Archives: A Content Standard* section 2.3 instructs metadata creators to provide a *supplied title* when the resource does not have a formal title, or if the formal title is misleading or inadequate; and suggests including two parts: "the name of the creator(s) or collector(s) and the documentary (genre) form of the item." Professional metadata creators supply titles that support information discovery; however, untrained metadata creators (e.g. volunteers) may be less adept, especially when the context is lost in aggregated digital collections. A famous example of this, provided in Foulloneau *et al.* (2005), is the metadata-creator-supplied title "on a horse" for a photograph of President Theodore Roosevelt; once aggregated into a portal for digitized content from 12 university libraries, the title becomes very confusing for users.

To avoid similar situations, title templates are provided based on types of item, such as traditional stories (folk tales), lists of words (e.g. animals, body parts) and elicitations wherein speakers provide example sentences and may be asked clarifying questions. For example, a depositor might provide the title "Hunting and fishing," which could be a traditional hunting story, a conversation about a recent hunting trip or an elicitation of words and phrases to talk about hunting. The CoRSAL guidelines and title templates prompt depositors and provide examples based on familiar language materials (e.g. "Elicitation about hunting and fishing"). For audio/video items which have corresponding transcriptions or translations, "Transcription:" is added to the beginning of the title (e.g. Figure 1).

Oral histories have high variation in titles, as traditional stories may not have official titles. For example, a speaker may provide a title at the time of recording (e.g. "the Squirrel Story"), whereas another speaker may only recognize the story by the names of the main

characters (*Theipaa*, "squirrel" in the target language). Metadata creators generally include multiple titles, using a "parallel title" label when representing translations of the title into other languages (see Figure 3).

UNTL records sometimes use series titles to group items with a shared format, origin or topic. Many Digital Collections materials originate from archives, which create "collections" based on acquisition, usually named for the collector or donor. An archival "collection" may contain only one or two items, which is not useful for browsing a digital archive; however, series titles allow users to filter materials by these archival "subcollections." Similarly, series titles in CoRSAL indicate the source (e.g. researcher); these display as a filter within search results and as a link to "other items in this series" in item records. The Lamkang Language Resource, developed over a long-term research project, includes four series so that users can browse items specifically contributed by Daniel Tholung, Sumshot Khular, Shobhana Chelliah or Rengpu Rex Khullar.

*4.4 Language representation*

A central component of CoRSAL records is language. Within the UNTL edit system, an item's language(s) must be chosen from a drop-down menu with a locally developed

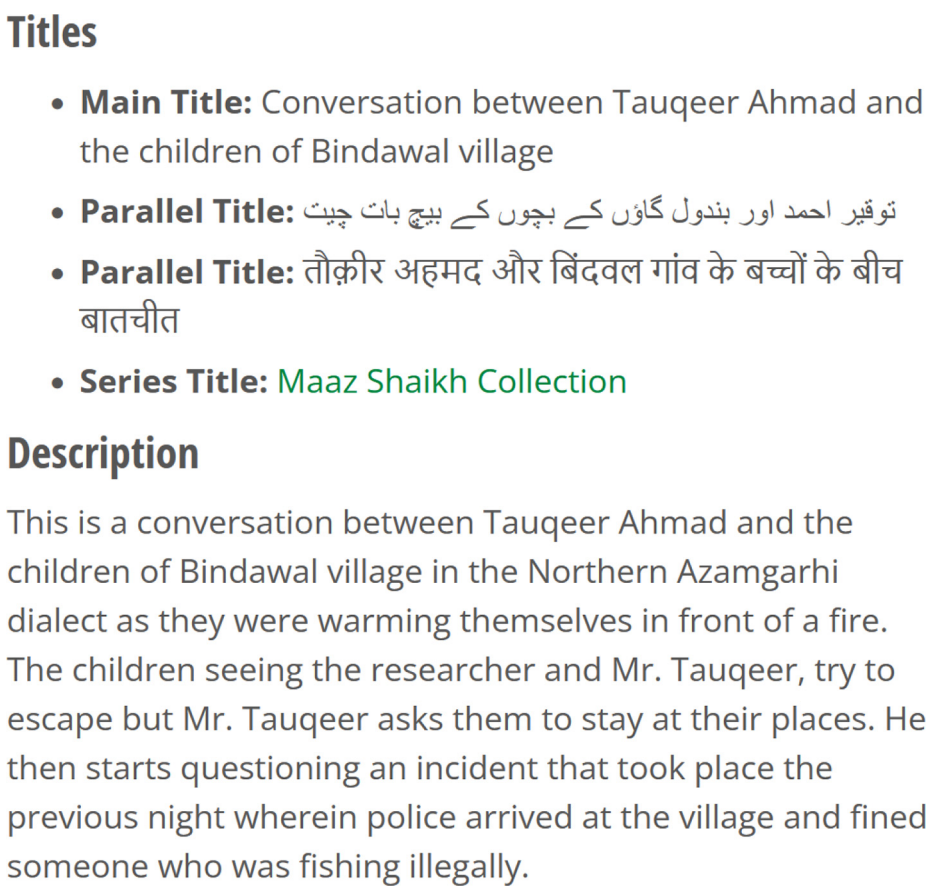



**Figure 3.**
Example of an English supplied title translated into Hindi and Urdu to facilitate access for the relevant audiences

controlled vocabulary (available at: https://digital2.library.unt.edu/vocabularies/languages/). Any time an item in the Digital Collections contains a previously unused language – including most CoRSAL collections – a new term is added to the vocabulary. Prior to the inclusion of linguistics data, these languages were based on the three-letter codes and names in ISO 639-3; however, endangered languages are not always recognized.

CoRSAL and Digital Libraries staff developed procedures to account for additional scenarios. When a researcher submits materials containing a language not yet in the vocabulary, staff check the ISO 639-3 list; if there is no ISO 639-3 code, they consult the Glottolog (available at: https://glottolog.org/) to determine if a standardized "Glottocode" is available. (Glottolog describes itself as the "comprehensive reference information for the world's languages, especially the lesser known languages.") Occasionally, research references a language not appropriately represented in either authority, such as the "Azamgarhi" language. Although a code was available for the larger language group (east2875), it was decided to represent precise languages to avoid possible confusion. In those cases, CoRSAL staff will apply for a new Glottocode to be created, in this case, azam1235.

As with other controlled vocabularies in the Digital Collections, existing vocabularies are used whenever possible; this saves time and makes metadata more sharable. Managing these values within the system allows the support of data entry (e.g. drop-down menus), supplement with "local" values and add details (e.g. descriptions of locally created language codes). This has also been a way to mediate between "authoritative" language values and those preferred by a language community. For example, the Mizo language was historically represented by the ISO 639-3 code "lus," for "Lushai," which is now considered pejorative. The ISO 693-3 code (available at: https://digital2.library.unt.edu/vocabularies/languages/#lus) is used, but the public display has the name "Mizo." The entry can reflect preferred names or alternative terminology while maintaining continuity with other standards.

*4.5 Subject representation*

At least two subject values are required in every UNTL record to facilitate searching and collocation (especially for nontextual materials or to connect items across different collections). Subject terms may come from a number of different standard controlled vocabularies that align with collection or partner needs – for example, Getty Art and Architecture Thesaurus (available at: www.getty.edu/research/tools/vocabularies/aat/), Chenhall's Nomenclature for Museum Cataloging (available at: www.nomenclature.info/) or the Library of Congress Medium of Performance Thesaurus for Music (available at: https://id.loc.gov/authorities/performanceMediums.html). Editors can also assign uncontrolled keywords to describe item content and use the subject qualifier to label the type of subject or vocabulary.

As mentioned in the literature review, there is a very limited set of domain-specific controlled vocabularies for language data, including four small-scale and less known OLAC vocabularies: for linguistic subject (29 terms), role (24 terms), discourse type (10 terms) and linguistic data type (3 terms). Instead, metadata creators rely on more extensive vocabularies used in other practice rather than OLAC. Those vocabularies tend to define terms more broadly or sometimes quite differently, similar to the conflicts regarding creator and contributor roles. Previous studies show that the OLAC vocabularies are not widely used (Author, 2022); currently, CoRSAL does not use any OLAC vocabularies.

The Library of Congress Subject Headings (LCSH, available at: https://id.loc.gov/authorities/subjects.html) is the largest controlled subject vocabulary widely used in library,

archival and museum collections. Two groups of subject terms in LCSH reflect language archive topics:

- "Linguistics," with 85 narrower headings, some of which have subheadings – for example, "Biolinguistics" has "Neurolinguistics" as one of its narrower terms; and
- the "Language and Languages" heading with 60 narrower headings, also with subheadings – e.g., "Colloquial language" has "Conversation" and "Slang" among its narrower terms.

Available LCSH terms are often insufficient for language archive materials, although relevant general concepts are included (see Figure 4, left). As a result, CoRSAL metadata records rely on free-text keywords; for example, the record in Figure 4 (right) includes the keywords *directionals* and *spatial reference*. F4

Keywords are also used for genre terms (e.g. elicitation, traditional narratives, word lists) because UNTL metadata does not include a separate genre field (see Figure 4). In language archives, genre is an important attribute for many users; researchers, for example, may want to know whether audio was recorded in a natural setting or as part of a formal experiment. Although keywords are not precisely analogous to genres, they are searchable and become clickable links to find all items in the UNT Digital Library containing the same term (see Figure 5). F5

Some digital collections designate genres or similar concepts through types. Although resource type and format are required fields, these values are browseable terms and do not always reflect individual collection needs. For example, there are options for images and text; however, subsets of these tend toward categories such as "technical drawings," which may be architectural or engineering schematics, data-based charts and so on, vs domain-specific terminology. This also applies to "text," which is used for many of the CoRSAL materials and is a catch-all for items not fitting a specific category.

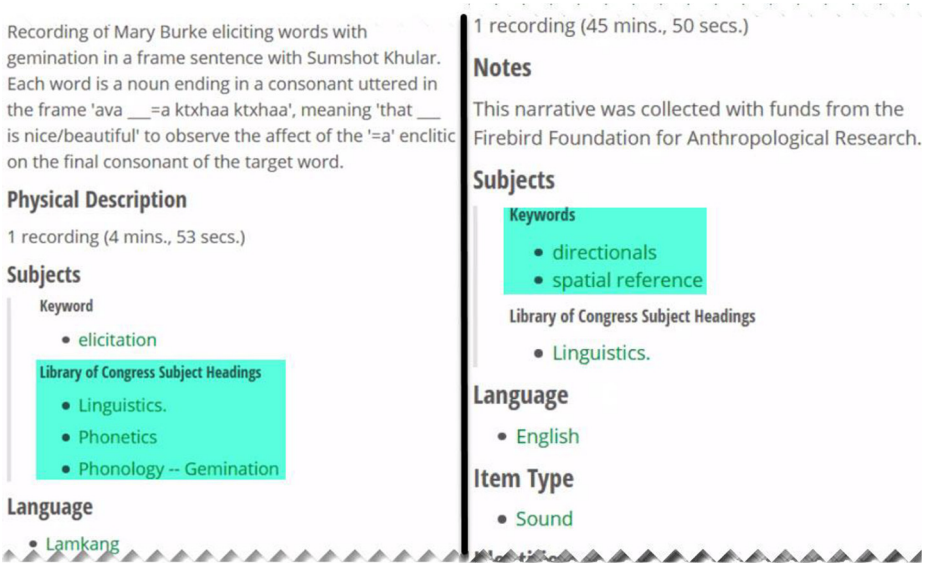



**Figure 4.**
Sample metadata records with LCSH terms (left) and noncontrolled vocabulary keywords (right)

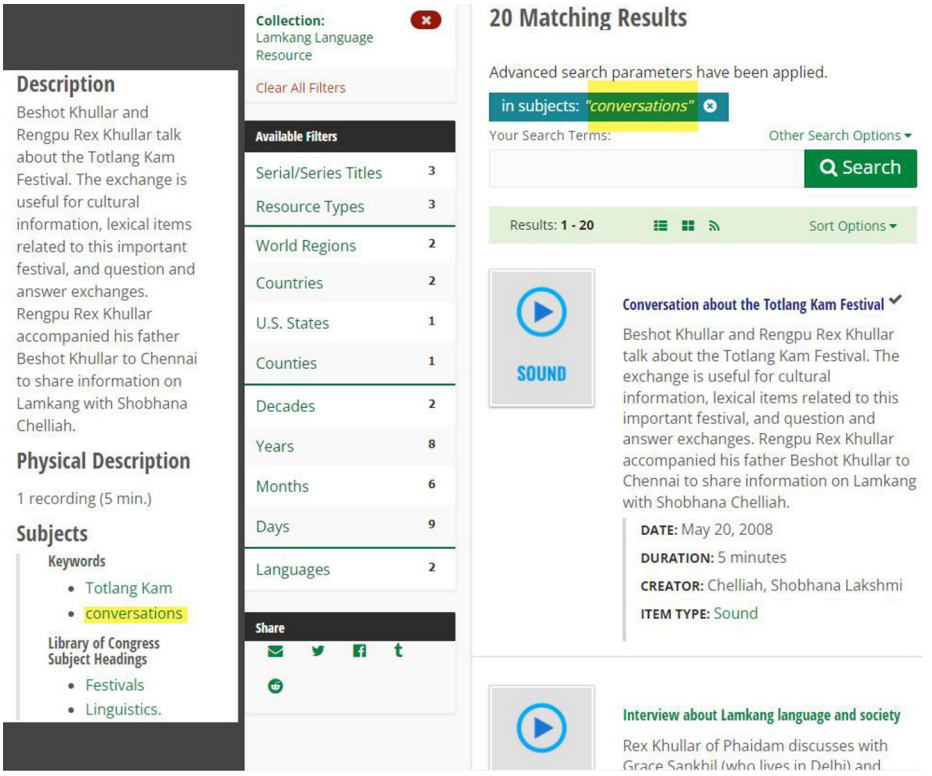



**Figure 5.**
Example of hyperlinked genre term "conversations" represented as keyword in Subject metadata field (left) and search results for term "conversations" in Subject fields (right)

## 5. Discussion

Based on the experiences at UNT, below are highlighted areas for potential challenges.

### *5.1 Domain knowledge differences*

One challenge of digital repositories is integrating materials from many sources – with different standards and needs – and representing them for a variety of users. There are two main concerns about balancing discipline-specific terminology or representation (e.g. through separate, labeled fields vs qualified fields). First, digital archives make information findable (e.g. through Google), so descriptive metadata should be understandable for public users without knowledge of the full collection or jargon. Second, researchers from other disciplines should be able to discover something that might be relevant for cross-disciplinary projects. Just as linguists and librarians use different terminology, metadata decisions impact how users find or understand materials.

Additionally, the lack of consensus on existing standards and controlled vocabularies in the linguistics domain makes it difficult to represent terminology consistently. Without established standards, metadata records may be interpreted differently by information professionals and users of language collections from various backgrounds. As specialists from the linguistics domain become more involved in metadata creation and advocate for the continued refinement of these standards within the language archiving community, the development of a common understanding of terminology is anticipated.

*5.2 Opportunities for sharing records with metadata aggregators*

The Digital Collections infrastructure was developed to make metadata shareable and to intentionally expose data for search engines and via Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). For example, records in The Portal to Texas History are harvested into the Digital Public Library of America. Additionally, metadata records in the Digital Collections can be downloaded in multiple formats (e.g. UNTL and simple Dublin Core) in bulk or individually. Documentation for accessing CoRSAL metadata is at: https://digital.library.unt.edu/explore/collections/CORSAL/api/#oai-pmh. Dublin Core is a simpler metadata scheme so some information is omitted, including local fields and qualifiers (e.g. Figure 6 shows the UNTL and Dublin Core versions of a record). F6

Records can be harvested as a collection, which supports a potential future aggregation to collocate linguistic-related resources and data. OLAC, the primary aggregator for linguistic resources, uses OAI protocols to harvest metadata from 62 partners (40 active feeds) including digital language archives, corpora, physical libraries and websites. At the time of writing, CoRSAL has not registered with the OLAC repository.

## 6. Conclusions

This paper describes the collaboration between the UNT Digital Libraries and CoRSAL to make language resources available online. Although the UNTL metadata scheme is not always a perfect match for resources in CoRSAL, it has been easily adapted to language materials in most cases. The authors developed processes to address concerns (e.g. related to language names, and genres) and are continuing discussions regarding other issues to ensure robust description that meets the needs of researchers and the wider Internet audience.

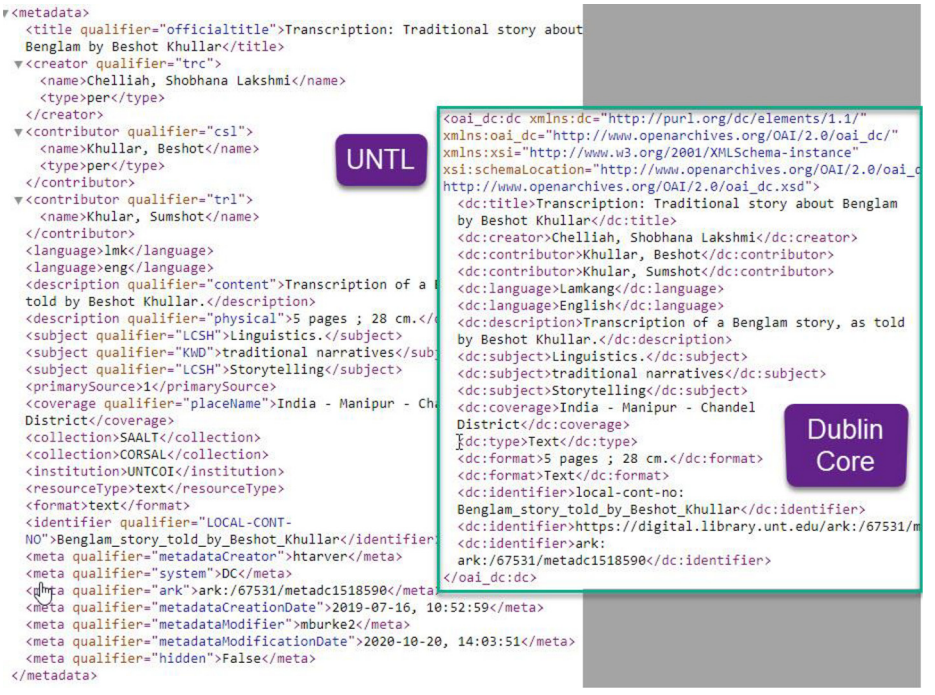



**Figure 6.** Example of a CoRSAL record in native UNTL metadata scheme and converted into OAI DC metadata scheme for harvesting

With any metadata implementation, user studies would determine the level of usability for end users. Currently, little is known about language archive user groups; a study focusing on the CoRSAL interface and metadata may reveal insights for various use cases. Conducting studies across multiple language archives may guide decision-making in the broader community about the use of OLAC vocabularies and subject representation in language materials.

Another future possibility is to expand training for metadata creators in the language archiving community and in the information science and library communities. For example, students in metadata classes at UNT have performed metadata quality evaluations on randomly selected CoRSAL records and discussed collection-specific guidelines. More extensive training could result in higher metadata quality and promote cross-disciplinary communication.

Overall, adding CoRSAL collections to the UNT Digital Library has provided a straightforward way to make materials findable and accessible while making use of the existing infrastructure and the UNTL metadata schema. While it requires some flexibility and logistical planning, this model and the general success in providing access to these materials demonstrates the potential to make more items available, even when a discipline-specific venue may not be available to researchers.

## References


Abraham, T. (1991), "Oliver W. Holmes revisited: levels of arrangement and description in practice", *The American Archivist*, Vol. 54 No. 3, pp. 370-377, available at: https://americanarchivist.org/doi/pdf/10.17723/aarc.54.3.2urn146354t3704r (accessed 1 February 2022).

Andreassen, H.N., Berez-Kroeker, A.L., Collister, L., Conzett, P., Cox, C., De Smedt, K. and McDonnell, B. and Research data alliance linguistic data interest group (2019), "Tromsø recommendations for citation of research data in linguistics (version 1)", doi: 10.15497/rda00040, available at: www.rd-alliance.org/system/files/FINAL%20Version_Troms%C3%B8-Recommendations-Citation-Research-Data-Linguistics.pdf (accessed 1 February 2022).

Author (2022).

Belew, A. and Simpson, S. (2018), "The status of the world's endangered languages", Rehgm, K.L. and Campbell, L. (Eds), *The Oxford Handbook of Endangered Languages*, Oxford University Press, New York, NY, pp. 21-47.

Berez-Kroeker, A., Gawne, L., Kung, S., Kelly, B., Heston, T., Holton, G., Pulsidier, P., *et al.* Woodbury, A., (2017), "Reproducible research in linguistics: a position statement on data citation and attribution in our field", *Linguistics*, Vol. 56 No. 1, pp. 1-18, doi: 10.1515/ling2017-0032.

Bird, S. and Simons, G. (2003), "Extending Dublin core metadata to support the description and discovery of language resources", *Computers and the Humanities*, Vol. 37 No. 4, pp. 375-388, doi: 10.1023/A:1025720518994.

Bird, S. and Simons, G. (2021), "Towards an agenda for open language archiving", *ACM/IEEE JCDL LangArc Workshop Proceedings*, pp. 25-28, available at: www.ideals.illinois.edu/bitstream/handle/2142/111675/LangArc2021_Proceedings_7Oct2021.pdf

Computational Resource for South Asian Languages (2022), "Collaborative language archiving curriculum", available at: https://corsal.unt.edu/curriculum (accessed 1 February 2022).

Czaykowska-Higgins, E. (2009), "Research models, community engagement, and linguistic fieldwork: reflections on working within Canadian indigenous communities", *Language Documentation and Conservation*, Vol. 3, pp. 15-50.

Daoutis, C.A. and de Montserrat Rodriguez-Marquez, M. (2018), "Library-mediated deposit: a gift to researchers or a curse on access? Reflections from the case of surrey", *Publications*, Vol. 6 No. 2, p. 20.

Digital Library Assessment Interest Group's Metadata Assessment Working Group (2021), "Case studies: metadata assessment", Digital Library Foundation Blog, available at: www.diglib.org/case-studies-metadata-assessment/ (accessed 1 February 2022).


AQ: 11

Foulloneau, M., Cole, T.W., Habing, T.G. and Shreeves, S. (2005), "Using collection descriptions to enhance an aggregation of harvested item-level metadata", in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL '05)*, Association for Computing Machinery, New York, NY, pp. 32-41, doi: 10.1145/1065385.1065393344.

Gawne, L., Berez-Kroeker, A.L., Andreassen, H.N., Conzett, P., De Smedt, K., Cox, C. and Collister, L.B. (2021), "Using the Tromsø recommendations to cite data in language work", Paper presented at the 7th International Conference on Language Documentation and Conservation, 4-7 March 2021, available at: www.youtube.com/watch?v=GyBCslbn6tc&ab_channel=ICLDCYT (accessed 1 February 2022).

Good, J. (2002), "A gentle introduction to metadata", available at: www.language-archives.org/documents/gentle-intro.html (accessed 1 February 2022).

Hanard, S. (2001), "The self-archiving initiative", *Nature*, Vol. 410, pp. 1024-1025.

Harris, A., Gagau, S., Kell, J., Thieberger, N. and Ward, N. (2019), "Making meaning of historical Papua New Guinea recordings: collaborations of speaker communities and the archive", *International Journal of Digital Curation*, Vol. 14 No. 1, pp. 136-149.

Henke, R. and Berez-Kroeker, A.L. (2016), "A brief history of archiving in language documentation, with an annotated bibliography", *Language Documentation and Conservation*, Vol. 10, pp. 411-457.

Hughes, B. (2005), "Metadata quality evaluation: experience from the Open Language Archives Community", in Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E. and Lim, E. (Eds), *Digital Libraries: International Collaboration and Cross-Fertilization*, Springer, Berlin, pp. 320-329, doi: 10.1007/978-3-540-30544-6_34.

International Federation of Library Associations and Institutions (IFLA) (2017), "IFLA library reference model: a conceptual model for bibliographic information".

Kurtz, M. (2010), "Dublin core, DSpace, and a brief analysis of three university repositories", *Information Technology and Libraries*, Vol. 29 No. 1, pp. 40-46.

Lee, N.H. and Van Way, J.R. (2016), "Assessing levels of endangerment in the catalogue of endangered languages (ELCat) using the language endangerment index (LEI)", *Language in Society*, Vol. 45 No. 2, pp. 271-292.

Nathan, D. and Austin, P.K. (2004), "Reconceiving metadata: language documentation through thick and thin", in Austin, P.K. (Ed.), *Language Documentation and Description*, Vol. 2, pp. 179-188. AQ: 12

National Science Foundation (NSF) (2018), "Data management plan for SBE proposals and awards", available at: www.nsf.gov/news/news_summ.jsp?cntn_id=118038&org=SBE (accessed 1 February 2022).

Open Language Archives Community (OLAC) (2011), "OLAC mission", available at: www.language-archives.org/index.html (accessed 1 February 2022).

Tillman, R.K. (2017), "Where are we now? Survey on rates of faculty self-deposit in institutional repositories", *Journal of Librarianship and Scholarly Communication*, Vol. 5 No. 1, p. eP2203.

Wasson, C., Holton, G. and Roth, H. (2016), "Bringing user centered design to the field of language archives", *Language Documentation and Conservation*, Vol. 10, pp. 641-671.

Woodbury, A.C. (2014), "Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire", *Language Documentation and Description*, Vol. 12, pp. 19-36.

**Corresponding author**
Mary Burke can be contacted at: maryburke@my.unt.edu