Abstract

The University of North Texas (UNT) team, including lead applicant (PI) Dr. Oksana L. Zavalina (UNT Department of Information Science), and collaborators (Co-PIs) Dr. Schobhana Chelliah (UNT Department of Linguistics) and Mark E. Phillips (UNT Libraries, Digital Projects), will conduct a planning research project during the period of December 2018-November 2019. The project's goal is to help prepare for a forthcoming collaborative project that will aim to extend the usefulness of existing language data archive collections through a user-centered design of systems incorporating the efficient tools for providing digital access to language data collections at scale.

As part of the planning project, the team will address the library and community needs in background information necessary for improving access to language data archives that are currently underutilized by intended user communities. This planning project will identify the gaps between the information organization methods and techniques currently offered in existing language data archives and the needs of actual and potential language data archive users. As part of the project, we will conduct exploratory content analysis of information organization tools (metadata schemes, controlled vocabularies, Linked Data applications and more) of existing language data archives. We will also interview and observe representatives of intended user communities of language data archives regarding their experience and expectations towards using information organization tools in language data archives. The outcomes and tangible products resulting from the proposed planning project include widely disseminated findings in the form of presentations, published papers and reports.

The intended audience of the planning project includes developers of digital language archives, as well as end-users of language data (researchers and educators in the areas of languages and linguistics, members of language communities). For the language archivists, the planning project is expected to result in increased awareness of information organization barriers that users experience when depositing to language archives or accessing materials deposited to language archives. Measurable changes in behavior of the end-users of language data archives that are expected to result from implementation of the future project informed by results of this planning project include the following:

- increased volume of depositing language data sets to language data archives
- increased volume of interactions with language data archives from the intended user communities (e.g., search, browse, downloads of language materials for use in teaching, research, or other purposes).

Exploring Methods and Techniques for Facilitating Access to Digital Language Archives

The UNT Information Science Department, UNT Linguistics Department, and UNT Libraries are seeking IMLS National Leadership Grant support under *Curating Collections* project category and *Planning Grant* funding category for a collaborative project to identify the gaps between the information organization methods and techniques currently offered in existing language data archives and the needs of actual and potential language data archive users. This planning project seeks to provide necessary background information and preparation for a forthcoming collaborative research project that will aim to extend the usefulness of existing language data archive collections through a user-centered design of systems incorporating the efficient methods and techniques for providing digital access to language data collections at scale.

1. Statement of National Need

Linguists in the United States, often supported by federal funding, continue to create numerous valuable digital datasets. Most of these datasets are unique and many of them represent low-resource or endangered languages. Like most digital preservation activities in various knowledge domains, language data archiving initiatives started in late 1990s and early 2000s. Online language archives are a valuable tool to support language preservation and revitalization, and to providing data on lesser-known languages valuable for linguistic analysis (e.g., Henke & Berez-Kroeker, 2016). Language archives are repositories of linguistic data about a selected set of languages, typically including recordings, transcripts, translations, and linguistic annotations.

At present, language data are archived in a number of specialized language repositories. Some examples of the largest language data repositories include California Language Archive, Archive of the Indigenous Languages of Latin America, Native American Languages Collections at the Sam Noble Museum etc. In the United States, over 20 of such archives were created in at least 9 states. Texas, Pennsylvania, California, and Hawaii are the US states with 3 or more language archives. Approximately half of the language archives located in USA are hosted by universities, others by non-profits and other organizations. To bring this rich language data together to facilitate access to it, an Open Language Archives Community (OLAC) was created, an international collaboration project sponsored by the US National Science Foundation in 2000-2010 and hosted by the University of Pennsylvania Libraries. OLAC put together the combined catalog of all resources from 60 participating language archives located throughout the world: in USA, Australia, Brazil, India, New Zealand, Taiwan, and several European countries. OLAC's combined catalog contains over 300,000 records, covering resources in many languages of the world.

Cultural heritage institutions, including libraries, have long curated datasets, with metadata as a key tool for curation. Since 1970s, libraries started providing access to analog for users through online catalogs, with highly expressive MARC (Machine Readable Cataloging) metadata. After approaches for bibliographic control of machine-readable data files in the social sciences were proposed in 1980s (Dodd, 1982; Gray, 2013; Stephenson, 2013), it gradually became common for academic libraries to acquire digital datasets and provide access to them with MARC metadata (Hogenboom & Hayslett, 2017). As early as in 2001, almost 97% of libraries responding to a survey on numeric data products and services reported that they provided MARC metadata records describing datasets in their online library catalogs (Cook, Hernandez, & Nicholson, 2001).

MARC is a rich metadata scheme, that is constantly updated to align with current developments in information organization. The latest version of MARC bibliographic standard includes 231 metadata fields, many of which have more than one subfield. The freely-available online WorldCat database (https://www.worldcat.org/) aggregates hundreds of millions of MARC metadata records that describe various information objects, including archival materials. For example, the search for “language” in the WorldCat database retrieves over 8.8 million of MARC metadata records, including more than 536 thousand of metadata records representing archival materials. MARC metadata records rely on authority control (the use of standard controlled vocabularies to represent agents related to the lifecycle of information objects and various kinds of their subjects: concepts, events, places, works, persons, groups, etc.). Authority control significantly improves discoverability of information objects. Linked Library Data expands on authority control ideas and further facilitates access to digital information through emphasizing relationship representation. Linked Data is a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.

In the United States, access to digital data is now increasingly provided through institutional data repositories that are often part of universities’ open access repositories. Those repositories rely on the metadata that is much simpler and less expressive than MARC: most commonly on Dublin Core metadata scheme (DC Metadata Element set 1.1 with 15 descriptive elements, or extended DC Terms) traditionally used for describing collections of various digitized and born-digital content. Dublin Core is a default metadata scheme that is installed with popular digital content management software such as DSpace, etc. Dublin Core standard is very flexible and does not prescribe the use of controlled vocabularies which often results in lower quality metadata records that affects information retrieval. Often, local metadata application profiles based on Dublin Core are constructed for use in data repository. For example, University of North Texas (UNT) Digital Library, and its component Data Repository, use the UNTL metadata scheme for describing items from all of its collections, including the Lamkang Language Resource collection of materials on an endangered language of Northeast India.

UNTL is the local extended version of the standard Dublin Core metadata scheme, with 20 descriptive metadata elements (to represent the information object itself), and 1 administrative metadata element (to represent data about metadata records: the record creator name and the record creation time and date, record editor name and record modification time and date, record size in bytes, etc.). In the UNTL metadata scheme, metadata elements (e.g., date, description) are accompanied by qualifiers (e.g., date qualifier=“creation”, description qualifier=”physical”). Three major controlled vocabularies are used in UNTL metadata: the Library of Congress Name Authority File (LC NAF) for personal, corporate, and geographic names, the Library of Congress Subject Headings (LCSH) and UNT Libraries Browse Subjects (UNTL-BS) for representing subjects of information objects. Metadata is downloadable from UNT Data Repository in the original UNTL form (in Extensible Markup Language XML format), in 4 different syntaxes of a standard Dublin Core (XML, text, and 2 seralizations of Resource Description Framework: RDF/XML and JSON), and in Metadata Encoding and Transmission Standard (METS, in XML syntax). Additional metadata schemes, beyond Dublin Core and its local versions (i.e., UNTL), are also used in the language archiving community, and the most prominent of them is the OLAC metadata scheme.

In the 2000s, the language archiving community has been building awareness of metadata and working on designing metadata standards through projects such as Electronic Metastructure for Endangered Language Data (E-MELD, http://emeld.org/). This 5-year project conducted best-practice workshops and schools, developed a metadata editor tool, a linguistic ontology of morphosyntactic terms, and an OLAC metadata standard for linguists to describe their data (http://www.language-archives.org/OLAC/metadata.html). OLAC metadata standard is also based on Dublin Core, but unlike UNTL, it has a specific focus on the attributes of linguistic data: language, language family, country, linguistic subfield, data type, format, and more. The available values of these descriptors in OLAC metadata records serve as the basis for powerful faceted search in OLAC database search interface. Digital Endangered Languages and Musics Archive (DELAMAN) is another organization working on standards for language data archives. An example of its most recent activity in the area is the NSF-funded workshop Metadata Editing and Collection Management (MEaCoM) offered in June 2018.

Another standard metadata scheme relevant to describing language archives is Text Encoding Initiative (TEI). Language archiving research distinguishes between two kinds of metadata which are both important for information organization in language archives: so called "thin" metadata that facilitates research discovery and so called "thick" metadata that represents text encoding of the linguistic documents itself: transcriptions, commentary, and time-aligned annotations (Nathan & Austin, 2004). The TEI is an international metadata standard that focuses on this "thick" metadata, and that traditionally uses XML syntax.The latest version of TEI released in July 2017, attempts to integrate TEI with ontologies thus making conversions from TEI (XML) to RDF more useful for corpora databases.

In addition to item-level descriptions of individual objects in archives, collection-level descriptions that describe the entire collection are very important. Archival community traditionally, for centuries, relied mostly on these collection descriptions. The descriptions exist as finding aids: first analog, unpublished or published as books; later in digital form in HTML or PDF formats. Encoded Archival Description (EAD) is the oldest -- initially released in 1996 -- and widely used in archives collection-level metadata scheme. Also, MARC metadata used in libraries allows creating collection-level record to represent archives and has been widely used by library community. These collection-level descriptions in the last decade are harvested and indexed by the ArchiveGrid (https://beta.worldcat.org/archivegrid), an international initiative led by OCLC Research in USA that provides access via the WorldCat database to millions of metadata records representing collections from over a thousand of archival institutions. For example, a search for "linguist" in the ArchiveGrid database retrieves 786 metadata records (e.g., a record describing a collection of Vietnamese-American linguist-lexicographer Dinh Hoa Nguyen's papers from 1949-2000 held by University of Washington Libraries Special Collection). Of the 5.35 million of archival records in ArchiveGrid as of April 2018 (the latest available data at the time of preparing this proposal), almost 5.1 million are MARC records, and most of others (over 171,000) are EAD records. In addition, regional aggregations of metadata that describes archives are available. One example is a Texas Archival Resources Online (TARO, https://legacy.lib.utexas.edu/taro/)

Rich and unique digital language datasets have a potential to make a strong contribution in social science research and education (e.g., Language Science, Geography, History, Sociology) and Computer Science (e.g., natural language processing). However, this potential currently remains largely unrealized as language archives are rarely accessed by linguists or indigenous language communities to use the available language data. One main reason for this lost opportunity is the confusing and cumbersome design of language archives (Wasson, Holton & Ross, 2016; Wasson et al., 2018). For example, as depositors upload information with various levels of granularity, retrieval for educational or research purposes becomes untenable without much additional resorting and organization of the archived data. Also, users of language archives cannot easily compare data across languages. Many theoretical breakthroughs in historical linguistics, syntax, phonology, and other areas would become possible if users could query archival data for cross-linguistic patterns. But current language archives place data on each language in a separate collection and do not function as databases (Al Smadi et al. 2016).

Another issue of concern is that the level of depositing language data remains marginal and much lower than for other types of datasets. One of the important reasons for this is metadata-related. Research (e.g. Wasson, Holton & Ross, 2016, etc.) shows that users often find contextual information (including metadata) misleading and even missing. Also, the linguists are unsure if the widely used metadata schemes are appropriate for representation of language datasets and if these will ensure the use of their data. They often do not expect non-linguists (who normally make metadata-related decisions and create metadata in data repositories) to adequately represent language data for information retrieval. In archives that allow self-depositing, depositors such as linguists and community members involved in language revitalization need help with interpreting and applying metadata themselves. Need for detailed and accurate metadata to facilitate resource discovery, as well as ease of search in language data archives is documented by user studies (e.g., Al Smadi et al., 2016).

The proposed planning project is aimed at preparing and supporting a large-scale project that will identify and test ways to bridge the gaps between the language datasets and data users. This project will build on previous work in the area (e.g., the standards and tools developed by E-MELD, OLAC, DELAMAN for language data community, as well as other metadata standards and tools used for facilitating access to data in repositories). The project includes identifying and implementing the effective information organization techniques – metadata schemes and/or application profiles, controlled vocabularies, linked data applications – that would make sense for both the potential depositors and the potential users of language data and would inform user-centered design of language data archives.

2. Project Design

2.1 Goals and Outcomes

The overall goals of the proposed planning project are twofold:

- to identify the following:
  - information organization tools and practices currently employed by language data archives across the United States, and
  - the needs of depositors and end-users (linguistics researchers, instructors, and students) for information organization functionalities in these archives.
- to provide empirical data in support of planning the future large-scale collaborative project focused on development of more efficient and user-friendly tools for access to digital language archives.

The proposed planning project will support an IMLS agency-level goal of Content and Collections. Two performance goals include: (1) broadening access and expanding use of the Nation's content and collections and (2) improving management of the Nation's content and collections in language data archives. The project will result in strengthening libraries as essential partners in addressing the needs of their communities in language archives.

Expected outcomes include publications and presentations based on the study results. In the process of interacting with representatives of stakeholder groups for language data archives, we expect to establish and develop partnerships for the future collaborative research project: identifying both co-PIs from other institutions across the country and a potential advisory board consisting of at least six members to aid the future research project by offering guidance and reviewing documentation and methods.

2.2 Research Questions

The planning project team will carry out an investigation guided by the following research questions:

1. How is information currently organized in the language archives?

2. What are the needs of actual and potential depositors of language data with regards to information organization in language data archives and how they correlate with available information organization functionality?
3. What are the needs of end users of language data (researchers, educators, students) with regards to information organization in language data archives and how they correlate with available information organization functionality?

2.3 Primary Benefits for the Intended Audience

The projects will benefit four groups involved in language data archiving: language archiving service providers, linguists, language community representatives, and library and information science programs. First, language archive practitioners and managers will be able to use the reports developed in this project to organize information in language archives to better serve their user groups. Linguists and communities of language speakers will benefit from more user-friendly language data archives, and from training materials on how to more efficiently use language data archives informed by results of the proposed project. Finally, library and information science educators will be able to use the findings as they develop curriculum to meet the needs of language archivists.

2.4 Input from Community and Consensus Building

The planning project team uses the following methods to ensure that the project design allows for input, consensus building and buy-in from others within the field:

- in the process of working on the proposal, we have obtained feedback from several experts in language archives, who enthusiastically support the proposed project (letters of support are included in the attachment).
- during the project, we will consult with an advisory board. The advisory board will include expert representatives from linguistics and language archiving communities, as well as from broader digital library community. The advisory board will meet online twice during the grant project: in spring 2019, and in early fall 2019. These meetings will allow the project team at UNT to solicit feedback related to research methods, identifying participants of the user study, and to present and receive feedback on preliminary research findings and reports. We will also consult with members of the advisory board individually at the onset of the project and on as needed basis during the project. Please see the Advisory Board section below for details on the structure of advisory board.

2.5 Work Plan

The study will be implemented, under PI supervision, by a research assistant, with contributions by the entire project team.

In the first stage of the proposed planning project, the team will conduct the **exploratory content analysis** of the websites of existing language data archives (institutional and stand-alone), including the UNT language data repository. This stage of analysis will be aimed at answering our research question no. 1. The analysis will be informed by the following:

- user-centered paradigm in information science research (e.g, Dervin & Nilan, 1986)
- theory of user-centered design of digital libraries (e.g., frameworks for digital library evaluation developed by Xie, 2006; Albertson, 2015, etc.).
- theory and practice of metadata and broader information organization in digital libraries in general
- the previous work of language archiving community.

The OLAC list of participating archives will serve as the starting point for creating the list of language archives to analyse; it will be refined and extended through examination of the list of language archives indexed by ArchiveGrid. The focus of the content analysis will be on information organization in language data archives: metadata scheme used, extent to which metadata records are displayed to end users, options for advanced search against indexed metadata fields, availability of adaptive and personalized search, authority control, linked data applications, social tagging functionality etc. We will also collect and examine available auxiliary information (e.g., documentation of a metadata application profile used in a language data archive, metadata creation guidelines, data on the use levels of a language archive, etc.). We expect to complete the exploratory content analysis stage of the project in December 2018-April 2019.

In the second stage of the proposed planning project, the team will conduct **interviews and observations** of a small selected sample of actual and potential users of language data archives. In this stage, we will seek answers to our research questions no.2 and 3. Participants at this stage of our project will be identified by those project team members who possess background in linguistics (please see Project Team section below). In the process of identifying potential participants, project team will also rely on the help of language archives experts in the project advisory board (please see Advisory Board section below). Interview participants will include representatives from the following groups of stakeholders:

- Linguistics researchers depositing or planning to deposit their datasets in language data archives and using or planning to use language data archives in their research
- Language and linguistics educators using or planning to use language data archives for teaching (K12 - higher education)
- Students who would benefit from using language data archives in their studies (linguistics students and information science students)
- Language community members interested in heritage language materials
- Language archiving practitioners and managers.

Participants -- linguists and language community members -- who do not currently use language archives will be interviewed. The semi-structured interviews will be conducted with the purpose of language data archive requirements analysis or user needs assessment, as well as to collect information on how these requirements are met by information organization in language archives based on the previous experience of respondents in using language archives. Participants who already use language archives will also be observed by the project team:

- Depositing participants will be observed as they interact with information organization tools (including metadata) in language archives in the process of depositing.
- Participants who do not deposit themselves but use materials deposited by others in language archives will be observed searching and browsing language archives and interacting with metadata in the process.

Observations will represent the heuristic evaluation of information organization in a selection of language archives. The development of observation protocols will be informed by results of content analysis carried out in the first stage and findings about user needs obtained as part of interviewing in the second stage.

We expect to complete the second stage of data collection and analysis in May-September 2019. This stage will be followed by generation of reports and dissemination of results.

2.6 Project Management

The primary investigator will maintain a work plan identifying the key activities and completion dates for the project. The work plan will be available to all project team members, who will utilize a Web space (e.g., a wiki or blog) to document and share the progress and results of their work. To track the progress toward achieving project results, the graduate research assistant will meet with the PI on a weekly basis, and the entire research team will meet on a monthly basis. Please refer to Schedule of Completion document.

2.7 Project Team

The proposed planning project is a collaboration between the UNT College of Information (UNTCI) and the UNT Libraries (UNTL). As such the responsibilities for leading and managing the grant will be shared between PIs Oksana Zavalina (UNTCI Department of Information Science), Shobhana Chelliah (UNTCI Department of Linguistics), and Mark Phillips (UNTL). The team members, who have successfully collaborated in the past, will bring together the extensive research and practical expertise needed for this project, i.e., information organization, linguistics research and education; building and managing digital repositories (including language data repositories); experience in successful planning and implementing projects funded by IMLS and other national funding agencies.

Dr. Oksana Zavalina will serve as Principal Investigator of the proposed project. She has extensive practical experience in metadata creation (both MARC metadata traditionally used in libraries, and Dublin Core and other metadata used in digital repositories, as well as Linked Data applications). Dr. Zavalina has conducted research on metadata that facilitates resource discovery in digital repositories for over 12 years. This includes research on metadata-related decision-making in designing digital repositories; content analysis studies of metadata, including metadata quality and metadata change; studies of user interactions with digital repositories. Her responsibilities will include: overall project supervision and budget oversight; development of data collection instruments; drafting, editing and submission of reports and grant documentation; and official communication with IMLS.

Dr. Shobhana Chelliah will serve as Co-Principal Investigator of the proposed project. She is a leading expert in Tibeto-Burman linguistics. Her publications include the highly cited grammar of the Tibeto-Burman language Meitei as well as general articles on Tibeto-Burman typology. She has advocated for the use of naturalistic data in linguistic discovery, specifically referencing the need for natural data to supplement current trends in South Asian linguistics to use translation and elicited data. Dr. Chelliah has been extensively involved in language data documentation. She was Program Director of NSF’s Documenting Endangered Languages Program from 2012-2015, where she helped institute a higher standard for Data Management Plans, including mandatory time and funds devoted to secure archiving of DEL-funded data. Her current NSF funded research dealing with Tibeto-Burman lexicography and morphology will allow for evaluation of information organization in language archives from the perspective of a language documenter and typologist. She will contribute to the project her expertise in language data structures, in-depth knowledge of the needs of linguistics researchers and understanding of language data depositing issues. Her responsibilities will include assistance with developing interview and observation instrument and with interpreting results of interview and observation, help in recruiting participants for interviews and observations.

Mr. Mark Phillips will serve as Co-Principal Investigator for the project. He has extensive experience in grant-funded projects for digital libraries and Web archives as well as experience in grant-funded research projects involving ethnographic-based research methods including observation, interviewing and focus groups. His responsibilities will include: assisting with analysis of metadata records and metadata documentation from language archives. He will also provide expertise in system and infrastructure design that will be useful to understand metadata practices present in language archives.

The project will hire a graduate research assistant. It is expected that the research assistant will be a student of the UNT Information Science Department's interdisciplinary Information Science PhD program with a concentration in Linguistics. The research assistant will have major role within the project including data collection, transcription and coding, data analysis, and assistance in writing of research reports and papers. The detailed job description for a graduate research assistant is included in the Resumes attachment.

2.8 Advisory Board

This research project will make use of an external advisory board with members from a wide range of institutions and backgrounds that will assist in guiding the project to successful completion. All members of the advisory board have deep experience in one or more aspects of the research project. The following individuals have already committed to serving on the advisory board for this project: Susan Kung (language archives), Gary Holton (language archives, language data standards), Myung Ja Han (digital libraries, metadata), Christina Wasson (user-centered design). These advisory board members will participate in virtual meetings spaced throughout the project timeline. Please see attached letters of commitment for a more in-depth discussion of advisory board members' interest in the project.

2.9 Dissemination of Results

The results of the proposed project will be disseminated in the form of:

- presentations to the project's advisory board representing language archives experts and experts in other related areas
- presentations and/or panel discussions at national and international conferences in both the linguistics community and information science community. Possible venues include the Association for Tribal Archives, Libraries and Museums (ATALM) conference in October 2019, annual meeting of the Association for Information Science and Technology in November 2019, iConference in March 2020.

3. Diversity Plan

The proposed planning project will inform improvement of information organization in language data archives that will serve the needs of diverse communities: native speakers and learners of various languages, including endangered or under-resourced languages. There is a need to understand the diversity of the participating individuals, institutions and subject matters involved in language data archive creation. Systematic demographic information on language archiving practitioners and depositors of language materials (language community representatives and linguistics researchers) is not yet available so it is impossible to define targeted activities designed to reach underrepresented groups at this time. This proposed project will establish the ethnic and cultural diversity (racial/ ethnic background, age, sex, disability, geographic location, socioeconomic status, etc.) of the language archiving practitioners' community via the data collection mechanism (content analysis) and consultations with project advisory board members. Based on findings from the analysis of this baseline dataset, the project will identify and engage representatives of diverse communities and traditionally underserved groups in interviews and observations. The analysis of interview and observation group results will include looking at variability of possessed understanding of information organization in language data archives and unique service needs among diverse populations. The project team personnel will include both males and females representing different racial and ethnic backgrounds, age groups, and educational backgrounds.

4. National Impact

The proposed planning project seeks to provide necessary background information and preparation for a forthcoming collaborative research project that will aim to extend the usefulness of existing language data archive collections through a user-centered design of systems incorporating the efficient methods and techniques for providing digital access to language data collections at scale.

Our proposed planning project aligns with one of the goals stated in IMLS strategic plan *Creating a Nation of Learners:* “IMLS supports exemplary stewardship of museum and library collections and promotes the use of technology to facilitate discovery of knowledge and cultural heritage”. Additionally, the project fits well into the *Curating Collections* category in that the output of our planning project can assist libraries and other institutions of all sizes across the country in shared service for access, preservation and stewardship of digital collections of linguistic content.

We will ensure that results from the proposed planning project are disseminated widely to interested communities by presenting and participating in panel discussions at relevant research and practical conferences, and by publishing results in conference proceedings and journals. Research findings from the proposed planning project will be also shared through the project Web space during the project Published conference and/or journal papers as well as unpublished reports and white papers discussing results will be deposited in the UNT Scholarly Works Repository (https://digital.library.unt.edu/scholarlyworks/) for long-term access upon completion of the project. Tabulated data will be shared using UNT Data Repository (http://digital.library.unt.edu/datarepository).

The proposed planning project is expected to have a far-reaching national impact by addressing the needs of language research and education across the nation and extending the usefulness of existing language data archives. The project will fill the gap in understanding how user needs in information organization of digital language data correlate with functionalities currently offered by various language data archives. As a result of building this understanding and awareness, language archivists will be able to develop and apply more effective and user-centered metadata and information organization tools, and language data archives will grow due to increase in depositing and will become more widely used by language educators, researchers, indigenous language communities, and other interested user groups. The findings of this planning project will also be useful for information science educators, in developing information organization curriculum for language archivists and linguists.

This project is the first step in a series of research and demonstration projects aimed at improving the information organization in language data archives around the country. For example, its findings will inform the future project that will help libraries meet the challenges in ensuring language data curation, availability, and discoverability at scale through identifying and testing solutions to overcome the information organization functionality barriers to active depositing of rich available linguistic data and effective utilization of language data archives.

The project is also expected to serve to build stronger connections between linguists, language archive service providers, and information specialists and researchers developing digital libraries. This interdisciplinary project will help cross-pollinate the library and information science and linguistics communities through a graduate research assistant working in both areas. A goal is that this student will be better able to appreciate the work in each area within the project. Finally, exposing future researchers and practitioners to the technical side of information organization and metadata will benefit the field in general.

The reuse of language data archives in building more discoverable collections is an area of significant future research potential. The ability for libraries, archives, and museums to identify the most meaningful for users ways of information organization in language archives is an important first step in fully realizing the potential of language archives and digital libraries nationwide.

References

Al Smadi, D. et al. (2016). Exploratory user research for CoRSAL: report prepared for S./ Chelliah, Director of the Computational Resource for South Asian Languages. University of North Texas. Department of Anthropology.

Albertson, D. (2015). Synthesizing visual digital library research to formulate a user-centered evaluation framework. New Library World,116(3-4), 122-135.

Cook, Michael N., Hernandez, John J., and Nicholson, Shawn. (2001). Numeric Data Products and

Services: A SPEC Kit Compiled By… Washington, D.C.: Association of Research Libraries.

Dervin, B., & Nilan, M. (1986). Information needs and uses. Annual Review of Information Science and Technology (ARIST), 21, 3-25.

Dodd, Sue A. (1982). Toward Integration of Catalog Records on Social Science Machine-Readable Data Files Into Existing Bibliographic Utilities: A Commentary. Library Trends, 30(3), 335-361.

Gray, Ann S. (2013). Sue A. Dodd’s Lasting Influence: Libraries, Standards, and Professional

Contributions. IASSIST Quarterly, 37(1-4), 15-17.

Henke, R., & Berez-Kroeker, A. (2016). A brief history of archiving in language documentation, with an annotated bibliography/ Language Documentation & Conservation, 10, 411-457.

Hogenboom, Karen, and Hayslett, Michele. (2017). Pioneers in the Wild West: Managing Data

Collections. portal: Libraries and the Academy, 17(2), 295-319.

Nathan, D. & Austin, P.K. (2004). Reconceiving metadata: Language documentation through thick and thin. In Language Documentation and Description, edited by Peter K. Austin, 179-187. London: SOAS.

Stephenson, Libbie. (2013). Social Science Data Files and Bibliographic Control: Contributions of Sue A.

Dodd. IASSIST Quarterly, 37(1-4), 8-14.

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

Wasson, C., Medina, M., Chong, M., Le-May, B., Nalin, E., & Saintonge, K. (2018). Designing for diverse user groups: Case study of a language archive. Journal of Business Anthropology, 7(2).

Xie, I. (2006). Evaluation of digital libraries: criteria and problems from users' perspectives. *Library and Information Science Research*, 28 (3), 433 - 452.

Schedule of Completion

	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
Identifying language archives for analysis	🟩											
Phase 1: Explorative content analysis: data collection	🟩	🟩	🟩									
Phase 1: Explorative content analysis: data analysis and presentation of preliminary results to advisory board for feedback			🟩	🟩	🟩							
Identifying participants for Phase 2					🟩							
Phase 2: Interview data collection						🟩	🟩					
Phase 2: Interview data analysis							🟩	🟩				
Phase 2: Observation data collection								🟩	🟩			
Phase 2: Observation data analysis and presentation of preliminary results to advisory board for feedback									🟩	🟩		
Preparation of project reports, dissemination of results											🟩	🟩

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The products produced as outcome of our proposed effort are white papers, studies, and possibly datasets.
All white-papers, project reports, blog posts, or other documentation will be made available using a Creative Commons Attribution NonCommercial Share Alike (BY-NC-SA) license. Creative Commons licenses have been used effectively by the UNT Libraries and more broadly as a vehicle to openly disseminate findings from research projects to the general public.
All datasets created during the project will be released with an Open Data Commons Attribution License (ODC-BY) which aligns with the UNT Libraries' Library Data Retention Policy (http://www.library.unt.edu/policies/other/library-data-retention).

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

All output of this research project will be licensed using common open licenses like the Creative Commons BY-NC-SA licenses for white papers, project documentation, and website, and Open Data Commons Attribution License for datasets.

Published articles, conference proceedings, and other scholarly output will be made available in accordance with the UNT Open Access Policy 06.041 (http://policy.unt.edu/policy/06-041) using the UNT Scholarly Works Repository (https://digital.library.unt.edu/scholarlyworks/).

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

As part of this project, we will be conducting interviews and taking notes during ethnographic observations. The data collected via interactions with human subjects will be stored securely and accessed by project investigators only. Such data will be shared only after appropriate anonymization or with explicit consent from participants.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

New Digital Content, Resources, and Assets will not be created as output of this research project.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

Software will not be created as output of this research project.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Data will be collected via phone interviews, in-person interviews, and ethnographic observations, which involve note-taking, recording, and photographs. Interviews and observation will be conducted throughout the project. There will be a content analysis of existing documentation and publications conducted during this study. The project will also include analysis of existing metadata records from language archives.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

Data collection for interviews and observation involves human subjects and requires IRB approval at UNT. IRB application will be prepared and submitted when/if the project is approved for funding.

Data compiled from document analysis or metadata record analysis does not require IRB approval.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

Participants can be identified in interviews, notes, and recordings. Personally identifiable information will be stored securely and only Principal Investigator and co-Principal Investigator will have access to it. Before public release of the dataset all PII will be removed (participants will be assigned coded numbers and any information that may identify them individually will be obscured in the interviews, notes, and transcripts).

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Participants will be provided with informed consent forms, which they will sign. The forms will be stored securely and separately and the relationship to the collected data will be maintained via a study ID that will be recorded in the informed consent forms and in the data files.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

The data will be collected via interviews and observations and will consist of text files, audio and video files, and photographs. Common word processing software and multimedia players may be used to display the data. Processed data may consist of additional spreadsheets and visualizations, which will be stored in non-proprietary formats (e.g., CSV or PNG).

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

Codebooks will be created as part of the analysis of qualitative data (e.g., in the thematic coding procedures codes will be developed in the inductive manner, after close iterative reading of the interviews). Codes, their descriptions and other documentation that describes when and where the interviews and observations took place will be stored in text formats along with the data. The documentation will be associated with the datasets through consistent file naming and through identifiers that refer to each data collection effort separately.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Any datasets created during this project will be managed and archived in the UNT Data Repository, a collection in the UNT Digital Library operated by the UNT Libraries. All content added to the UNT Data Repository will be preserved and made available by the UNT Libraries in perpetuity.

A.8 Identify where you will deposit the dataset(s):

Name of repository: UNT Data Repository

URL: https://digital.library.unt.edu/datarepository/

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

Principal Investigators will monitor the implementation of this data management plan. The plan will be reviewed every 6 months and adjusted according to the amount and types of data generated.