University of North Texas (UNT)

Exploring Methods and Techniques for Facilitating Access to Digital Language Archives

The UNT Information Science Department, UNT Linguistics department, and UNT Libraries are seeking $49,984 in 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.

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. 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 and digital datasets for users through online catalogs, with highly expressive machine-readable MARC metadata that relies 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. In the US, from 2000s onward, access to digital data is increasingly provided through institutional data repositories that are often part of universities' open access repositories. In addition, language data are also aggregated in several stand-alone specialized language repositories (e.g., California Language Archive, AAILA, DELAMAN, etc.). Data repositories rely on the much simpler and less expressive than MARC metadata: most commonly on Dublin Core metadata scheme – traditionally used for describing collections of various digitized and born-digital content – and local application profiles based on it.

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 for two main reasons. Language data archives are rarely used as depositors upload information with various levels of granularity such that retrieval for educational or research purposes becomes untenable without much additional resorting and organization of the archived data. 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. The linguists are unsure if the widely used metadata schemes are appropriate for representation of language datasets and 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, and in the archives that allow self-depositing, linguists need help with interpreting and applying metadata themselves.

The proposed planning project is aimed at preparing and supporting the large-scale project that will identify and test ways to bridge the gaps between the language datasets and data users. This includes identifying and implementing the effective information organization techniques – metadata schemes/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. Our proposed 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.

Project Design

The planning project team will carry out an investigation guided by the following research questions: How is information currently organized in the language archives? 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? 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? The study designed by PI and co-PIs will be implemented, under PI supervision, by a research assistant, with contributions by the entire project team.

The project team will conduct the exploratory content analysis of existing language data archives (institutional and stand-alone), including the UNT language data repository, with the focus on information organization (e.g., metadata scheme used, extent to which metadata records are displayed to end users, options for advanced search against indexed metadata fields, authority control, linked data applications, etc.). We will also collect end examine auxiliary information (e.g., documentation of a metadata application profile used in a language data archive, metadata creation guidelines, use data, etc.). Next, the team will conduct interviews and observations of a selected sample of actual and potential users of language data archives, representing different groups of stakeholders:

- linguists depositing or planning to deposit their datasets in language data archives
- linguists using or planning to use language data archives for their research and teaching (K12 - higher education)
- students who would benefit from using language data archives in their studies.

In the process of contacting and interviewing 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.

Outcomes and National Impact

Our primary goal is to identify the following: (1) information organization tools and practices currently employed by language data archives across the United States, and (2) the needs of depositors and end-users (linguistics researchers, instructors, and students) for information organization functionalities in these archives. Outcomes include publications and presentations based on the study results that will be shared via UNT Scholarly Works Repository and UNT Data Repository.

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. It will also 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.

Project Team and Partners

The planning project team includes PI Dr. Oksana Zavalina, Associate Professor in the UNT Information Science department, and two Co-PIs: Dr. Shobhana Chelliah, Professor in the UNT Linguistics department and UNT College of Information Associate Dean for Research, and Mark Phillips, Associate Dean for Digital Libraries at the UNT Libraries. It will also include one graduate assistant, a student in Information Science Ph.D. program with a concentration in Linguistics. 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.

Budget

We request $49984 in IMLS funds: The primary expenditure for this project is to fund one graduate research assistant (50% time for 11.5 months) for a total of $17889 plus $6736 in benefits (8.65% for the graduate students), $7550 in tuition reimbursement for the graduate student, $3950 to support team research travel to a language documentation conference to conduct interviews with linguistics researchers an educators, and $13859 in indirect costs at UNT's federally negotiated indirect cost rate (Dept. Health and Human Services, 06/19/2015).