Data Management Plan

Types of potential data: As this project advances it will produced the following types of data:

- **Text not linked to audio**: This material will comprise native terms for local flora and fauna; scientific identifications to the most precise level possible (i.e., genus or species in the majority of cases); locational data (state, municipality, village, geographical coordinates [either supplied by the individual uploading the data or placed in the database by locating village centers]); descriptive narrative on the referents (e.g., a short summary of a particular plant's use or of an animal's habitat).
- **Photographic and other visual material**: In all instances photographic and other images will be uploaded by project participants (in general, researchers and native speakers) and the overwhelming majority will be in jpeg format with different compression rates. The photos will be stored with the appropriate metadata (depositor, place, associated Indigenous name, Western scientific name).
- **Audio and text linked to audio**: As with all data, the original format of the material is dependent on the contributor. For researchers on endangered languages most material will be in non-compressed .wav format, 48 or 44.1 KHz and 16-bit. Delivery over the internet will be in mp3 at a bitrate of 192/kbits. A basic set of metadata fields will be developed for all recordings though again some of those who contribute materials might not have all the information sought.
- **Transcriptions and translations (English/Spanish)**: Language documentation efforts typically produce time-coded transcriptions, sometimes developed (e.g., in the transcription program ELAN) in three- or four-line interlinear format (transcription, parsed text, glossed text, free translation). In addition to the time-coded and possibly interlinearized transcription of audio data, transcriptions formatted for textual presentation (paragraph style) along with English/Spanish translations will be developed for this project.

Mark-up: All text files will be stored as Unicode text files. Mark-up protocols (XML) will be developed as part of this Level 1 grant though the guiding principle will be open-source format following the best practices data standards for archiving text. Time-coded transcriptions will be txt files encoded in the standard formats used by ELAN and Transcriber and this same protocol will be followed in the permanent archiving of this material. Future project development will explore the use of TEI mark-up for long narrative texts (e.g., formatted transcriptions and translations as those found in printed publications)

PDF: In the present iteration of this project PDF/A will be used for long narratives (transcription and translations) in paragraph style developed from time-coded transcriptions. The formatted transcriptions and translations will be exported to PDF/A for archiving at the Archive of the Indigenous Languages of Latin American (AILLA, University of Texas).

Policies for access, privacy, confidentiality and IPR: Any posted recordings will have to comply with strict adherence to recognized IPR protocols of informed consent. For researchers working on endangered languages this is generally not a problem as both their home institutions and the granting institution (e.g., NSF, ELDP, among others) require IRB approval at home institutions and certification that IPR protocols are being followed. For others, such as native speakers, who upload sound files care will be taken to ensure that informed consent has been obtained. This will be discussed by Amith, Remy, and Ogilvie in designing the user interface for uploading digitally recorded materials on flora and fauna. Strict care will be taken to ensure that all language material may be used for academic, non-commercial purposes as per the provisions to be posted on the open access websites. The protocols established by Mukurtu for cultural material of native communities will be reviewed and adapted as necessary in the present project.

Data management during website development: Website development will be undertaken on the Civic Actions server and the Beta version will be tested on this server. As the project draws to a conclusion the web structure and content will be ported to the Gettysburg College server and opened to the public in search functionality. Uploading permissions will be granted on a case-by-case basis until the site has been fully vetted by the project team.

Permanent archiving: All substantive materials will be deposited at the Archive of Indigenous Languages of Latin America. This will include the text file databases of Indigenous nomenclature of local flora and fauna, associated images, and recordings with their transcriptions and other related materials (e.g., PDF/A for formatted transcriptions and translations). The website will not be operational at AILLA but the substantive content will be safe in permanent storage. It should be noted that many of the recordings that will be used in this initial development of the website have already been archived at AILLA. Thus the material that will be developed in this project, and archived at AILLA, for the most part will comprise a database of ethnobiological terminology (Indigenous term, parsing and glossing of this term, term for biological referent, narrative on habitat or use, metadate on sources and location), photos and other illustrative materials, and formatted narratives.