Data Management and Processing for Endangered Language Documentation: A Workflow

R Karthick Narayanan, Meiraba Takhellambam and Pabitra Chettri

The Centre for Endangered Languages, Sikkim University (CEL, SU) by design is an organisation with multiple individuals of varying epistemic perspectives and technical skill sets working towards a goal of documenting the endangered languages of Sikkim and North Bengal region. Hence adopting a streamlined workflow with clearly defined roles and processes to ensure that the goals are met is essential. This document, part of our metadocumentation initiative, presents the workflow adopted by CEL, SU. It covers all the data processing steps followed by us from file renaming to preparing files for archival submission.

Raw file handling

The first step after data collection on the field is the saving and metadataing of the raw files created during fieldwork. This section describes the methods adopted by CEL, SU for the same.

Saving the raw files: Folder structure

The Raw files (audio, video, images, text) recorded in the field need to be sorted and stored systematically on the day of the recording. Following a uniform and systematic organisation of the files will help us locate the files much easier, as our collection of files are expected to grow over the period. The Centre has adopted the following folder structure for saving raw files, to deal with the large number of files that we have produced:

Data recorded each day is saved under a parent directory named after the language (for example: all data recorded from Magar fieldwork will be saved under the folder named 'Magar'), under which there is another sub-directory which would indicate the type or name of fieldwork (For example:'Magar>Pilot'). In this sub-directory, the files are saved under the directories named after the dates of the creation. (For example:'Magar>Pilot>12052017'). Inside the directory named after the date of creation, several sub-directories are made based on the media type (For example:'Magar>Pilot>12052017>Audio'). Typically there are four media types: audio, video, images and text. Inside these directories the raw files are stored after renaming them (see next section for file naming convention). The folder structure is represented below:

Language name> Pilot / Main> Date> Folders as per media types> Files

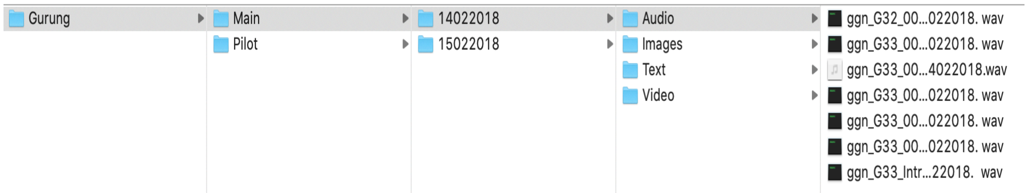

Figure 1: Folder Structure

All media files transferred to the computers must be backed up along with the same folder structure to an external hard drive by the end of the day.

File Naming Convention

All the files created during the documentation activity of the centre is given a unique name. The unique name is structured to represent the ISO code of the language, Informant ID, Session ID, Content code and the Date of creation. ISO code of the language is the unique code assigned to each language by the <body>. Informant ID is the unique code assigned to each language consultant from whom data is recorded. Separate Session ID is assigned to each sitting with the language consultant during the same day. Content code refers to the type of data collected, and may be one of the following:
SLP: sociolinguistic profile
WL: wordlist
SL: sentence list
EL: ethnolinguistic data
VA: verbal art
The pattern of naming convention for the four media types are as shown below:

1. Recording (Audio/Video):
   ISO code_(Informant ID_Session ID)_content code_DDMMYYYY.ext

   e.g.: byh_B1_001_WL_DDMMYYYY.wav

2. Photo:
   ISO code_content.ext

   e.g.: byh_pot_DDMMYYYY.jpg

3. Texts:
   ISO code_text source_DDMMYYYY.ext

   Must be written on top of the text source like: e.g.: byh_SikkimHerald_12052017

Metadata

Each file created in the process of documentation needs to be metadataed at every step of its process. This step is crucial for data management. After the raw files are renamed according to the above mentioned convention, metadata for each file must be created. The metadata scheme is as discussed in appendix. The work-in-charge of the language will be responsible for the compilation of metadata. One copy of compiled metadata and raw data should be handed over to the centralised backup facility.

Data Processing

Audio

Process all Sentence lists and Word lists as audio files. If they have been created as video files, extract the audio as wav file using the software Audacity.

1. Use Praat software to annotate the audio files.

   (a) Begin the annotation by converting stereo sounds to mono sounds by extracting one channel in Praat.

   (b) Run a script mark margins to mark boundaries. Specify the duration of the pause in script values for words: .1 msec or .2 msec & .6 msec for sentences. Or follow the Eri Kashima Elan/Praat Machine Segmenting method (https://yammeringon.wordpress.com/2017/05/01/elanpraat-machine-segmenting/).

   (c) Transcribe the utterances using the control phoneme chart issued by the language in-charge. The language in-charge may use Phonological Assistant software to create the phoneme chart.

   (d) Each audio file should be annotated with the following tiers: word, sentence, meaning in English and meaning in Nepali. The tiers may be named following Tim Gaved and Sophie Salffner (Gaved and Salffner,2014) suggestions (https://www.soas.ac.uk/elar/helpsheets/file122785.pdf).

   (e) The annotation file must have the same name as the audio file. Only the file extension of the audio file and the annotated file should be different.

   (f) Once the annotation is complete, the transcription must be checked and validated by the language in-charge.

   (g) Once the validation is completed the sound files may be chopped as required. To chop the files run a chop sound file script.

   (h) Name the chopped files in the format given below and create a metadata for the chopped files.

   Format

   - Words:

   ISO code_gloss_(01)_UID of file

   eg. byh_sleep_01_byh_B1_001_EL_13032018,

   byh_sleep_02_byh_B1_001_EL_13032018

   - Sentences:

   ISO code_SL code_(A / e_1)_UID of file

   eg. byh_PN_001_byh_B1_001_SL_13032018

   byh_PN_e_1_byh_B1_001_SL_13032018 (different)

   byh_PN_001_e_1_byh_B1_001_SL_13032018 (related)

   - Verbal Art:

   ISO code_content name_UID of file

byh_CreationMyth_byh_B1_001_EL_13032018
Metadata of the newly created TextGrid must be made. It should mention the source files in the 'Relation' field of the metadata.

Video

Process all forms of verbal art as video files.

1. Video files may be annotated using ELAN.
2. Annotate the files in ELAN following Tim Gaved and Sophie Salffner ELAN-FLEx-ELAN workflow (Gaved and Salffner,2014).
3. The annotation file must have the same name as the audio file. Only the file extension of the audio file and the annotated file should be different.
4. Once the annotation is complete the transcription must be checked and validated by the language incharge.
5. Metadata of the newly created TextGrid must be made. It should mention the source files in the 'Relation' field of the metadata.
6. If certain regions requires chopping the annotation should be exported as Praat text grids. The audio from the video file should be extracted using Audacity and saved as 'wav' file.
7. Once the exported files are validated. The sound files may be chopped as required. To chop the files run the chop sound file script.
8. Export annotations if required into Flextext for grammatical analysis. Follow the Gaved and Salffner,2014 workflow to complete the process.
9. For creating subtitles, export the gloss tier (English and Nepali) from ELAN as srt files.
10. Render video and subtitle together as one video using video editing software distribution. Render the video to mp4 (in MPEG4 or H.264 codec) with an aspect ratio of 16:9; frame width x height 960x540, or 1024x576 - web optimized; for HD ready 1280x720; for full HD 1920x1080.

Photographs

1. Create metadata for each Photograph in the collection using the CEL, SU Metadata Scheme.

Born Digital Texts

1. All text should be created in an open access format.
2. All text typed in Resource language writing system must be transliterated to IPA.
3. All transliterated text must be marked for morpheme boundaries.
4. Text data must be glossed as per Leipzig Glossing Rules.
5. Metadata of annotated text files must be created as per CEL, SU Scheme.

Research output

The research output of the CEL, SU documentation project is prescribed by the UGC guidelines. A Trilingual dictionary, Grammatical sketch and Interlinerarised Glossed Text of each language are expected to be produced at end of the project. In this part of the workflow we present the steps in creating the Trilingual dictionary, and Interlinearised Glossed Text.

Dictionary

A trilingual dictionary for each language that the Centre works on is produced using the Fieldworks Language Explorer (FLEx) software. Since multiple researchers are involved in the creation of the dictionary, FLEx's collaboration function is used. The following illustrates the process involved:

1. Establish a FLEx master file for each language.

2. Distribute client files among the collaborators.

3. The following are necessary fields in FLEx Lexicon that should be filled:

   - Lexeme in IPA and subject language's script;
   - Morpheme type; if the unit is a bound morpheme then function of the morpheme;
   - Grammatical category;
   - Gloss in English and Nepali;
   - Example sentence in IPA and subject language's script;
   - Translation of example sentence in English and Nepali;
   - Semantic domain;
   - For each lexeme, include sound files and picture where necessary;
   - Include lexical relationships for lexeme where necessary;
   - Include etymology if available;
   - If details of borrowing are available include it in etymological note;
   - For flora and fauna items include scientific names wherever identified;
   - Include encyclopedic/anthropological information for cultural terms

4. Integrate all client files with master files at regular intervals.

5. Create a BACKUP of the FLEx Project at regular intervals.

6. Print the output of trilingual dictionary. LaTeX export of the lexicon must be generated using PATHWAY and typeset using double-column A4/B4 size, and reversal index must be included in both the gloss languages.

7. Digital output (Android App format) must be generated using dictionary app builder software. Android App for each language must be generated in regular intervals and distributed among native speakers for validation.

8. Metadata for Flexfiles, backup files and output files must be created as per CEL, SU scheme.

Interlinearised Glossed Text

1. The FLEx client file used for lexicon building should be used for interlinearising text data.
2. Text data must be entered into FLEx manually or can be imported from ELAN as mentioned above.
3. Integrate all client files with master files in regular intervals.
4. Each word must be marked for morpheme boundaries and glossed using FLEx's 'Text and Words' functions.
5. Once all the items in the text are glossed. Export the Interlinearised Glossed Text Flextext file.
6. BACKUP of the FLEx project must be taken in regular intervals.
7. Create metadata for the exported Interlinearised Glossed Text.
8. Metadata for Flexfile, backup files and output files must be created as per CEL, SU scheme.

Submission for archive

Materials collected and documented by the Centre are to be archived in bundles. Each bundle will contain the following three elements:

1. Archival version of the material
2. Annotation files
3. Presentation versions of the material

Archival version of materials should be deposited into an archive as soon as possible after they are collected. Archival versions of the file must be complete, lossless, and unedited to the extent possible. Every bundle must be described using the CEL, SU metadata scheme. An annotation file must minimally contain transcription and translation in at least one of the gloss languages. Annotation files accepted for archival submission are Praat TextGrids, ELAN EAF, Flextext, Lift Lexicon or other structured data formats. PDFs will be accepted only in exceptional cases.

Appendix-Centre for Endangered Languages, Sikkim University Metadata scheme

The Centre for Endangered languages, Sikkim University metadata scheme is based on the Dublin Core Metadata standards and the Open Language Archives Community's recommendations. This metadata scheme uses all the 15 Dublin core elements and uses the necessary qualifiers to adapt it to describe the attributes of language resources that the Centre produces.

S.No	Label	Definition/Interpretation	Dublin core mapping
1	Identifier	This should be a unique identifier. It should be same as the file name.	dc.identifier
2	Title	For the community/collection/resource.	dc.title
3	Date	Date of recording.	dc.date
4	Place	Should be the place where the file was created, esp. for recordings.	dc.coverage.spatial
5	Source	Source of the data (How is the data sourced? Self or Others? if others please mention their name/or organization name.)	dc.source
6	Publisher	An entity responsible for making the resource available.	dc.publisher
7	Relation	Reference to related objects in the archive like agreement, associated files (like transcription (TR) and translation (TL)), reviews, photographs, etc.	dc.relation
8	Researcher	A person other than creator responsible for making research contributions to the item.	dc.contributor.researcher
9	Creator	The creator of the data.	dc.contributor.author
10	Consultant	A person responsible for making contributions to the content of the resource language.	dc.contributor.consultant
11	Language(s) used	The language the file is in.	dc.language
12	Resource language	Language "of interest".	dc.subject.language
13	Resource language's ISO 639-3	Three-letter ISO 639-3 codes for identifying languages, also commonly known as Ethnologue code.	dc.language.iso639-3
14	Genre*	Describing intellectual content.	dc.subject.classification
15	Discourse_Genre*	Specifically about (recorded) discourse.	dc.subject
16	Description	A brief description about the resource.	dc.description.abstarct

17	Elicitation Method	State the Elicitation method.	dc.description.elicitation
18	Type	Audio/Video/Image/Text.	dc.type
19	O.S. Requirement	An operating system required to use a software resource.	dc.format.os
20	Keywords	Keyword describing the resources.	dc.subject.key
21	Format	Mention File format like '.jpeg' '.mp4' '.wav'.	dc.format.mimetype
22	Size	Size of the file in MB or GB.	dc.format.extent
23	Length	Time duration of the audio/video file.	dc.format.duration
24	Pages	(only for documents) No. of pages.	dc.format.pages
25	Character Encoding	(only for documents/annotation files) State the Font Name used in the file. If these are special fonts that are downloadable give the link to it too.	dc.format.characterencoding