Emerging Role of Libraries in Language Archiving in India A Case Study of SiDHELA

1st Karthick Narayanan R
Independent Researcher

Madurai, India
0000-0001-5234-310X

2nd Meriaba Takhellambam
Department of Linguistics
Manipur University
Canchipur, Manipur, India
0000-0002-2616-0921

***Abstract*—SiDHELA is a language archive developed by the Centre for Endangered Languages, Sikkim University in collaboration with the Central Library, Sikkim University. It is the first language archive developed in India. SiDHELA is a model attempt at digital archiving in collaboration with communities of Sikkim and North Bengal region of India. The main highlight of the paper is the possibilities which emerges out of a collaboration between under resourced indigenous communities and an institutional library backed by a language documentation project to curate digital contents for endangered and lesser known languages from under resourced regions like the Northeast of India.**

***Index Terms*—Digital Libraries, Endangered Languages, Archives, Community Centric Approach, Collaboration**

I. Introduction

Language documentation began to emerge as an independent sub-discipline of Linguistics in the mid-1990s due to the global crisis of language endangerment and loss. Language documentation is 'concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties' [1]. Language documentation, unlike language description, is concerned with linguistic data instead of linguistic descriptions like grammar and dictionaries. One of the essential features of this new sub-discipline is a keen "Concern for long-term storage and preservation of primary data – language documentation includes a focus on archiving in order to ensure that documentary materials are made available to potential users into the distant future" [2]. The concern for archiving is an important concern for any team working on language documentation. The work presented here is on the Centre for Endangered Languages, Sikkim University's effort to address this concern.

II. PROJECT CONCEPTION

A. Background

The Centre for Endangered Languages, Sikkim University (CELSU here on-wards), in many senses, is a unique endeavour in the Indian context; the Centre is a collaboration between native speakers and academic experts to document the languages of Sikkim and North Bengal region in India. The University recruited an eclectic mix of native speakers, linguists, computer application experts etc. as staff of the Centre to be able to have a multidisciplinary approach to language documentation. Thus, efforts to address the concern mentioned above took shape at the Centre in the exchanges between the community members and subject experts. The archive development at the Centre was guided by four expectations shared with us from the community we worked with.

This work is supported by the University Grants Commission Financial Assistance for setting up Centre for preservation and promotion of endangered languages to Sikkim University during XII plan period (2012-2017).

B. Community Expectation

The first of the expectation for the archive comes from the Bhujel community of Sikkim. Bhujel community speaks the critically endangered language Bhujel. It has currently only one fluent speaker [3]. The community collaborated with the Centre to document the fading language use among their people. They expect the record of the linguistic knowledge documented by the Centre to be the basis for the community-driven language revitalization efforts. The Magar community expressed the second exception of the archive. They speak the Magar language, a definitely endangered language [4]. Their interest in collaboration with the Centre lies in the opportunity to document the Magar community's cultural practices and transmit them to the next generation. The community is also actively involved in developing pedagogic material for school children. CELSU also organised a Workshop on Script and Font development of Akkha script which is used for writing Magar. The community was eager to preserve the community's unique cultural practices. Hence they keenly demanded that the community's ritual and cultural practices be documented and preserved for future generations to learn from. Records created with the Magar community range from a video record of 'kul' clan pooja rituals practised by one of the sub-clans of Magar, a Magar food festival staged especially for the documentation purpose, a Magar fort celebration, many instances of Magar dance and other culturally relevant practices. The third expectation for the archive comes from the Sherpa community of Sikkim, with whom the Centre documented their language. The Sherpa community speaks the Sherpa language, a vulnerable language with little interruption in intergenerational transmission. Their language is an officially recognised language of the state and is taught in schools up to class 8.

Their involvement with CELSU was guided by their aspiration to have their language recognised by the Central Board of Secondary Education as a second language for Class 9 to 12 and eventually have university courses on Sherpa language. In line with their aspiration, they expected the archive to be a source for their language development efforts. Apart from these three, another fourth common expectation shared with us by the endangered language communities of Sikkim was that the archive must be a platform for self-training materials on language documentation and revitalisation activities. All these four expectations put together gave us an idea of an archive that moves beyond its traditional preservation function. The communities expect a dynamic platform functioning as a language resource centre. These expectations are not unique to CELSU; Endangered Language Archives as a platform has always had the responsibility to preserve and provide access to language data. Informed by the communities' expectation that CELSU worked to create an Endangered language archive was initiated by the university in 2019. These four expectations were transformed into guiding visions for the archive. Sikkim-Darjeeling Himalayas Endangered Language Archive (SiDHELA) conceptualised an archive with three main functions: Archive as a platform for linguistic resources, as a source of Cultural Documentation, as a platform for popular scholarly communication. Further, as the name suggests, the archive strictly focuses on the region Sikkim-Darjeeling Himalayas to maximise the outreach and potential among the communities of the region.

III. PROJECT IMPLEMENTATION

Implementing our vision to an actual archive was the challenge, and this is where the Centre sought to collaborate with Sikkim University's Library. Sikkim University is a newly established university funded by the Union government of India. It is presently established in a transit campus spread across the city of Gangtok, the capital of the Sikkim state, and a permanent campus is coming up in Yangang, in rural South Sikkim. Despite being in the early stages of establishment it is one of the few universities in India to have a functional institutional repository. The institutional repository is hosted by the University's Central Library (CL hereafter). CL is emerging as an important knowledge resource centre in the Sikkim-Darjeeling Himalayas region and actively caters to the knowledge demands of the region through various efforts.

A. Challenges in Collaboration

Collaboration between CELSU and CL was not a breeze. The initial attempts to forge the collaboration was met with reluctance. The impulse for such reluctance stems from the items that were to be submitted in the institutional repository. Most Indian institutional repositories are collections of research documents like Thesis, Reports, Articles and Ebooks. Hence, the then Librarian, insisted that we build a prototype to demonstrate the concepts, test out the system's integrity, and even insisted on getting an expert opinion on the prototype. Thus, before SiDHELA was created, a proof of concept was developed using a Dspace repository system. In this prototype, customisation to the Dspace's submission form and metadata registry was first tested. After reviewing the prototype by an external expert, the customisation was implemented in the Central Library's Digital repository and a special collection was created within the institutional repository to host the language archive.

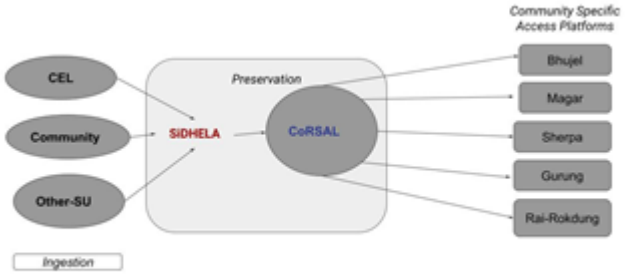
B. Moving the records from the field to the archive

Another challenging aspect of developing the archive was moving the recording from the field to the archives. Typically, an item in an archival collection consists of the recording and its annotation. The recording is stored both in an archival and a presentation format. The Archival version of the record in SiDHELA is a complete, lossless, and unedited version of the recording. It is submitted to the archive along with its metadata as soon as it was created. And the presentation version of the record too is submitted along with archival version. The presentation version is generally web-optimised video and audio formats. On the other hand, the annotation files are uploaded to the archive after further processing. The creation of annotation involves various levels of analytical procedures. Each annotation of a recording must minimally contain transcription in International Phonetic Alphabet (IPA) and translation in at least one of the gloss languages (English or Nepali). Transcription requires analysis of phonetic and phonology of the language. Translation would be possible only after a preliminary understating of the Language's morphosyntactic features. Apart from these, the annotation must be produced in a structured file format. To do this, CELSU used various tools to process them: Time aligned Transcription of audio and video recording are annotated using Praat and ELAN annotation software; Lexical database was developed using Field Language Explorer (FLEx) software while the resulting database is stored in the open-source XML format called Lift Lexicon. And the text corpora too were annotated at the morpheme level using FLEx. The resulting annotated corpora are stored in another open-source XML based file system called FLEXText. Apart from these machine-readable archival versions of the record, a human-readable PDF version is generated and deposited in the archive for ease of use. A detailed workflow [5] is followed to ensure uniformity and quality among the various item types processed by the Centre. Methods of processing the various types of items being generated at the Centre are discussed in the workflow. In addition to creating long-lasting records of languages, the Centre is also actively producing language technology tools to aid the communities in their language maintenance and revitalisation efforts. One of the crucial tools that the Centre actively produces is an Android dictionary application for each of the five languages were worked on. These android applications are derived out of the documented resources and are distributed through the archives. These dictionary applications are also archived in SiDHELA. Before the items could be deposited in the SiDHELA, each documented resource was bundled to combine the archival version of the recordings with their

presentation versions, the structured annotation file, and the annotation file's PDF. Each deposited bundle is then described using the CELSU metadata scheme. The metadata scheme uses all 15 Dublin core elements with the necessary qualifiers to adapt it. It is used to adequately describe the attributes of various language resources the Centre has produced. In total, twenty-five fields of information describe the resources they are: Identifier; title; date; place; source; publisher; relation; researcher; creator; consultant; Language (s) used; resource language: resource language's iso 639-3; genre*; discourse-genre*; description; elicitation; method; type; O.S. requirement; keywords; format; size; length; pages; and character encoding.

IV. Discussion

Despite all the efforts, SiDHELA has met the community's requirements only halfway. The archival software, DSpace, presently used in SiDHELA, is inadequate to meet the challenges of providing wider access. One of the important aspects being the lack of onsite media playback and streaming service. An equally important limitation is the rigidity of the front end in the DSpace system. It offers little customisation and no user-centred design.



SiDHELA-CoRSAL
Collaborative Archiving Model

The limitation of DSpace is overcome at CELSU by adopting a collaborative archive model. CELSU has entered into a collaboration with Computational Resource for South Asian Languages (CoRSAL), University of North Texas. Through this collaboration CELSU has plans to share a copy of the record stored and preserved both at SiDHELA with CoRSAL. Apart from satisfying the LOCKSS principle, this collaboration would further give CELSU the technical advantage available at CoRSAL, one of which is the ability to embed records stored in CoRSAL to other websites. This function could help CELSU meet the communities access expectation by providing community specific access platforms. The model of this planned collaborative archiving is represented through the figure above.

A. Lessons Learned

SiDHELA's experience has a few valuable lessons for language archiving in India. Firstly, for the minoritised language speaker, archives are expected to function outside its traditional domain of preservation. This expectation is common to all endangered language archives; they have the dual function of preservation and providing access. In that sense, it is best for any archiving efforts in India to collaborate with libraries as they specialise in providing access. Secondly, as discussed above, the diversity of items produced as a part of the documentary exercise has been the source of hesitancy among the Indian institutional libraries. These hesitance are not entirely unexpected, as language archiving is a novel exercise for libraries in India. These hesitance can very well be overcome with collaborative efforts between different parties involved in language documentation programs and libraries both institutional and otherwise. Thirdly, another significant lesson is the limitation of the popular archiving platform like DSpace. Overcoming the limitations of Dspace could be addressed either by improving the platform or by adopting a collaborative process across archives. SiDHELA like mini archives, could collaborate with established archives like CORSAL, ELAR or TLA to share the data and know-how amongst them and, in turn, use their capabilities to meet the community expectation. Finally, the most significant of all lessons were the usefulness of developing local archiving capabilities. It is observed in Sikkim that when a local institution acts as a bridge connecting communities and archives it could lead to a significant participation of the endangered language communities in language documentation and archiving.

V. Conclusion

The mirage of a language archive in India has always been a centralised data store which collates all the resources generated across the country. The experience of SiDHELA breaks that spell and points us towards smaller oases spread across the libraries of the country. The aspirations of the smaller lesser known communities of Sikkim and North Bengal and regional institutes provide us the means and resources to create local sanctuaries to protect and conserve indigenous languages.

References

[1] Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel,editors.Essentials of Language Documentation:. De Gruyter Mouton, 2008.

[2] Nikolaus P. Himmelmann.Chapter 1 Language documentation: What is it and what is it good for?:, pages 1–30. De Gruyter Mouton, 2008.

[3] Meiraba Takhellambam and Bishnu Lal Bhujel. Bhujel sociolinguistics sketch. Technical report, Centre For Endangered Language,Sikkim University, Gangtok, Sikkim, nov 2019.

[4] Meiraba Takhellambam and Gangi Maya Mangar. Magar sociolinguistics sketch. Technical report, Centre For Endangered Language,Sikkim University, Gangtok, Sikkim, nov 2019.

[5] Karthick Narayanan Meiraba Takhellambam and Pabitra Chettri. Data management and processing for endangered language documentation: A workflow. Technical report, Centre For Endangered Language, Sikkim University, Gangtok, Sikkim, mar 2018. Available: http://dspace.cus.ac.in/jspui/handle/1/7142.