DESIDOC Journal of Library & Information Technology, Vol. 40, No. 5, Sept 2020, pp. 292-299, DOI : 10.14429/djlit.40.5.16349

Made in India-SiDHELA : India's First Endangered Language Archive

R Karthick Narayanan

Center For Endangered Languages, Sikkim University, Gangtok - 737 102, India
E-mail: rknjnu@gmail.com

ABSTRACT

Sikkim-Darjeeling Himalayan Endangered Languages Archive (SiDHELA) created by the Centre for Endangered Languages, Sikkim University is India's first endangered language archive. This archive is part of the ongoing language documentation initiatives of the Centre funded by the University Grant Commission. The Centre, formally established in December 2016 aims for preservation and promotion of endangered languages in Sikkim and North Bengal. The Centre carries out documentation and description of the indigenous endangered languages of the region through linguistic and ethnographic fieldwork. SiDHELA conceptualised as a platform for a linguistic resource of the languages spoken in the region, houses the primary data collected through fieldwork. One of the main aims of this archive is to preserve the data for long term usage and dissemination. Central Library, Sikkim University hosts the archive under its digital library. Through this archive the Centre for Endangered Languages, Sikkim University seeks not just to preserve and protect but also to promote the use of endangered languages spoken in the region. This paper presents the journey of this archive from idea to reality. This paper outlines the motivation behind the conceptualisation of SiDHELA as a regional archive and then discusses its development. It includes discussion on the developmental platform, theoretical issues in the conceptualisation of the archive and practical challenges in its design and development and its prospects. This paper thus primarily intends to inform scholars and researchers working with endangered languages of the region about this archive and its development. Finally, it hopes to kindle interest among researchers and librarians for developments of more such regional archives.

Keywords: Endangered languages; Archive; India; Sikkim; DSpace; Library.

1. INTRODUCTION

Language death is one of the 21st centuries profound problems. It is a phenomenon wherein the speakers of a particular language abandon their language and shifts to another language. This threat of language shift is called as language endangerment. Estimates and data modeling[1-2] suggest that by the end of the century, more than 50 per cent of worlds languages will be lost. India tops the list of country with the maximum number of endangered languages in the world[3]. This crisis of loss diversity has invited deliberate effort to counter the threat of endangerment. Language documentation has been one of the essential response to address language endangerment worldwide.

Documentation efforts across the world have been supported by both government and non-governmental organisations for more than two decades now. All the internationally funded documentation projects have allocated separate funds to aid in the creation language archives. In India, there are only two dedicated endangered language documentation support schemes: SPPEL (Scheme for Protection and Preservation of Endangered Languages), administered by Central Institute Indian Languages (CIIL) and Funding Support to Universities for the study and research in indigenous and endangered languages of India, through University Grants Commission (UGC). These two schemes are run by the Ministry of Education and promote language documentation among academicians and universities since 2014. However, neither of these two schemes have any plans to develop language archives in India. Unfortunately, scholarships on building and maintaining archives among social science and humanities research in india are restricted to library science, manuscriptology, archaeology and economics. Linguistic and language documentation efforts could immensely benefit if this skill is put to use in India's attempt to save the endangered languages. In this paper, knowledge gained from building India's first Endangered language archive is presented. It first outlines the relationship between documentation and archiving, second, it discusses the conceptualisation of SiDHELA, followed by it the motivations for creating India's first endangered language is discussed, then a discussion on the implementation of SiDHELA is presented and finally, the challenges and prospects are discussed.

2. DOCUMENTATION AND ARCHIVING

Language documentation as a practice, from its inception in the late 1990s, has distinguished itself from descriptive linguistic with its primary focus on data as opposed to description and theoretical explanation. Language Documentation aimed at recording, preserving and providing access to a representative multipurpose language data. Himmelamann in his seminal

Received : 25 August 2020, Revised : 28 August 2020
Accepted : 31 August 2020, Online published : 04 November 2020

work on language documentation characterised the goal of a language documentation as "to provide a comprehensive record of the linguistic practices characteristic of a given speech community", he empahsised that the record must be "multipurpose and comprehensive record of the linguistic practices characteristic of a speech community" and said in language documentation "the emphasis is on the collection and representation of primary data rather than theory and analysis"[4].

This emphasis on the protection and preservation of the primary linguistic data emerged from the discourse on the role of data in Social Science and Humanities among the anglophone countries. EScience (UK), EResearch (Australia) or Cyberinfrastructure (USA) from the early 1990s have created adequate there to promote and preserve data. This effort has led to the creation of various digital tools in the anglophone world that aids researchers in discovering, accessing and using existing data for their research. Further, it has also created enough scholarship on the conversion of analogue data to digital forms, and best practices in producing and archiving born-digital data. Language archiving has emerged from this broader discourse to become a separate subfield of language documentation. From early 2000s linguists working on language documentation have accumulated a vast amount of knowledge on various aspects of language archiving: Discussion on data portability and their long-lasting usability[5,6]; description standards and the development of linguistics specific metadata standards[7,8,9]; workflow and standards for data processing from collection to curation[6,10-13] and archiving models and designs[14-17]. These efforts have led to the development of language archives with varying size and scope. A few significant examples are: The language archive hosted at Max Plank Institute[18] The Endangered Language Archive (ELAR) hosted at SOAS[19] the Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[20] and The archive of the indigenous languages of Latin America (AILLA) hosted at University of Texas[21].

The archives mentioned above host a vast amount of linguistic data from around the world, except for AILLA all the other archives are classified as global archives. Still, they host very little data on Indian languages. Among the three major archives of the world, The language archive, ELAR and PARADISEC we find data only on 30 odd Indian languages, spread over 82 collections. These collections do not represent even a fraction of the linguistic diversity of India. A closer look at the archives reveals that most of the deposits in these archives are created by linguist working in non-Indian universities, this tell us that archiving among Indian university researchers is still not prevalent. One of the reasons for this is that method of language documentation followed by most Indian researchers is different from the standard corpus model of documentation followed in and funded by anglophone organisations. Most language documentation research in India relays on the questionnaire model with a primary focus on producing descriptive grammars and dictionaries. This difference has led to the adoption of different recording principles, workflows, hardware tools, and software tools. This difference in workflow makes it difficult for Indian researchers to meet the requirements of these archives and hence the conspicuous absence of Indian language in these archives. The only solution to archive the already produced language data in India is then to encourage language archiving in Indian universities.

3. SiDHELA- A REGIONAL PROGRESSIVE ARCHIVE

Sikkim-Darjeeling Himalayas Endangered Language Archive (SiDHELA) is India first endangered language archive. It is a special collection archive designed and developed by the Centre for Endangered Languages, Sikkim University to be hosted in the digital repository maintained at the University's central library. It is envisioned as a regional and progressive archive. This archive is part of the ongoing language documentation initiatives of the Centre funded by University Grant Commission. At present, it holds language and cultural data documented by the Centre for Endangered Languages, from the Sikkim-Darjeeling Himalayas region (hence the name). It is developed by the Centre to suit its workflow and research design, and at the same time, it adheres to data and archiving standards acceptable by linguist universally.

As the name suggests, it is a regional endangered language archive embedded into the Sikkim University's Digital Library infrastructure. The nature of this structure has certain advantages and disadvantages. The collection in this archive is strictly regional; thus it primarily contributes to the preservation and documentation of endangered languages of the region and lends support to the revitalisation efforts. This focus on the region bring in one of the most significant advantages for a language archive; it is easily accessible to the community and encourages greater community participation in the documentation. Nevertheless, this high degree of embeddedness, as discussed in by Wasso, et al.[17] brings in challenges to "customising". Andrea Berez-Kroeker, Director of the Kaipuleohone-University of Hawai'i Digital Language Archive, another language archive embedded in digital libraries, said that archives like his face the biggest challenge in user interface, "As for the front end, I have no real control over what kind of information gets displayed, or how things can be searched. It's really geared towards traditional library publications, not media" (as quoted inWasso, et al.[17]). The above-mentioned scenario is true for SiDHELA as well, while we could implement a certain level of customisation in data description standards and submission process in SiDHELA much could not be done in the user interface.

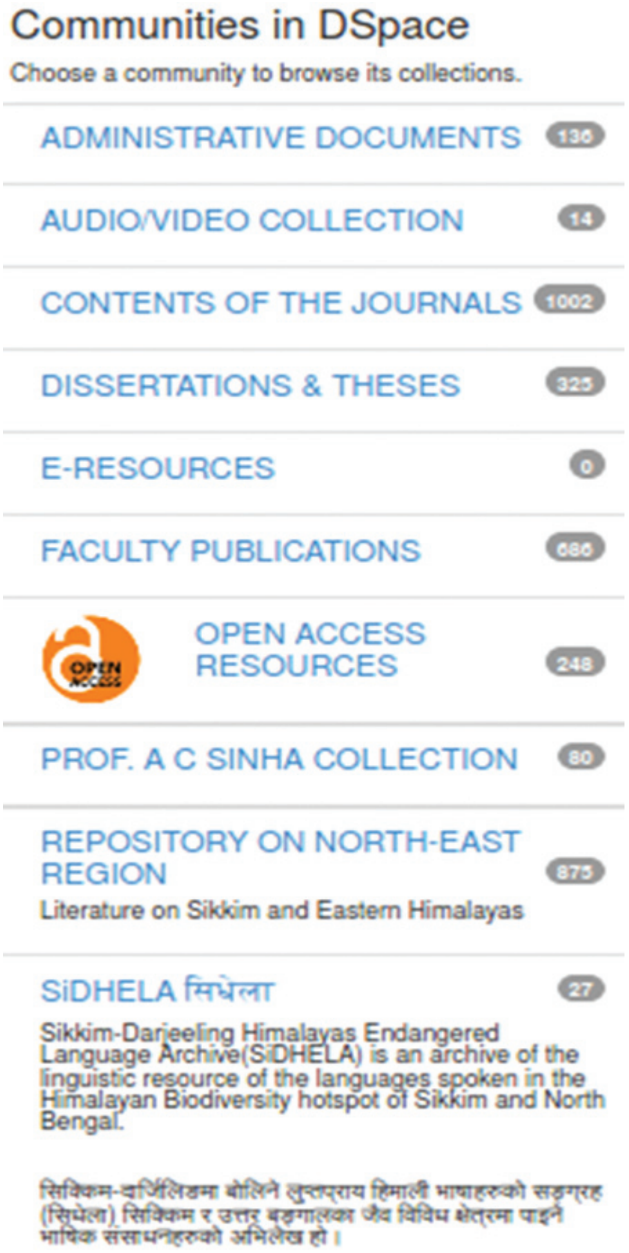
One of the radical changes in language archiving that has been initiated in the last ten years is to move language archiving away from dead scholar archiving model. Dead scholar archiving model are based on conventional archiving models wherein the collection are deposited as final products, and arching of linguistic data was seen as an "end point of documentation"[22]. Under this method, the language data is archived only after the analysis and publication. Progressive archiving, in opposition to this as Nathan[22] point out "encourages and enables incremental archiving, additions to existing deposits and updates and revisions of existing resources". This early deposit method required us to design and implement description of the data earlier in the research.

Centre’s workflow is also geared to meet it (more one this is discussed below).

4. MOTIVATION FOR SIKKIM DARJEELING HIMALAYAS ENDANGERED LANGUAGES ARCHIVE

The centre, formally established in December 2016 aims for preservation and promotion of endangered languages in Sikkim and North Bengal. The centre carries out documentation and description of the indigenous endangered languages of the region through linguistic and ethnographic fieldwork. Since the centre is a UGC special centre, it does not have complete autonomy in the research it carries out. The research of the centre primarily aims to fulfill the requirements dictated by the UGC. As per their condition, the centre was expected “to undertake fieldwork, research, analysis, archiving, and documentation using stat-of-the art speech and language technologies, in formats that are universally acceptable viz. digitised-textual, audio and-video-formats. To produce and publish monographs, grammars, grammatical sketches, dictionaries and lexicon, ethno-linguistic and theoretical descriptions, collection of oral and folk literature and scholarly books on endangered languages.” Under these broad requirements, the centre carries out various tasks on five languages spoken in the region Bhujel, Gurung, Magar, Rai-Rokdung, and Sherpa. Through the last three years, the Centre has accumulated 236.405 hours of audio and 22.59333 hours of video and more than a thousand photos. These recording cover a rage of linguistic data types: lexicon, sentences, folk stories, folk songs and narratives. All these data are born digital and are created following widely accepted standards. Centre for endangered languages sikkim university is probably the only university in India to have acquired this much of data on these languages. Hence we decided to open up this corpus of data for public use making it accessible to both the community members and linguist alike, for the following reasons.

Platform form for linguistic resources: Language shift and erosion of linguistic ability is getting common among the communities of the Sikkim and North Bengal. Bhujel one of the communities that Centre work with has only one speaker



Communities in DSpace

Choose a community to browse its collections.

ADMINISTRATIVE DOCUMENTS 136

AUDIO/VIDEO COLLECTION 14

CONTENTS OF THE JOURNALS 1002

DISSERTATIONS & THESES 325

E-RESOURCES 0

FACULTY PUBLICATIONS 686

OPEN ACCESS RESOURCES 248

PROF. A C SINHA COLLECTION 80

REPOSITORY ON NORTH-EAST REGION 875

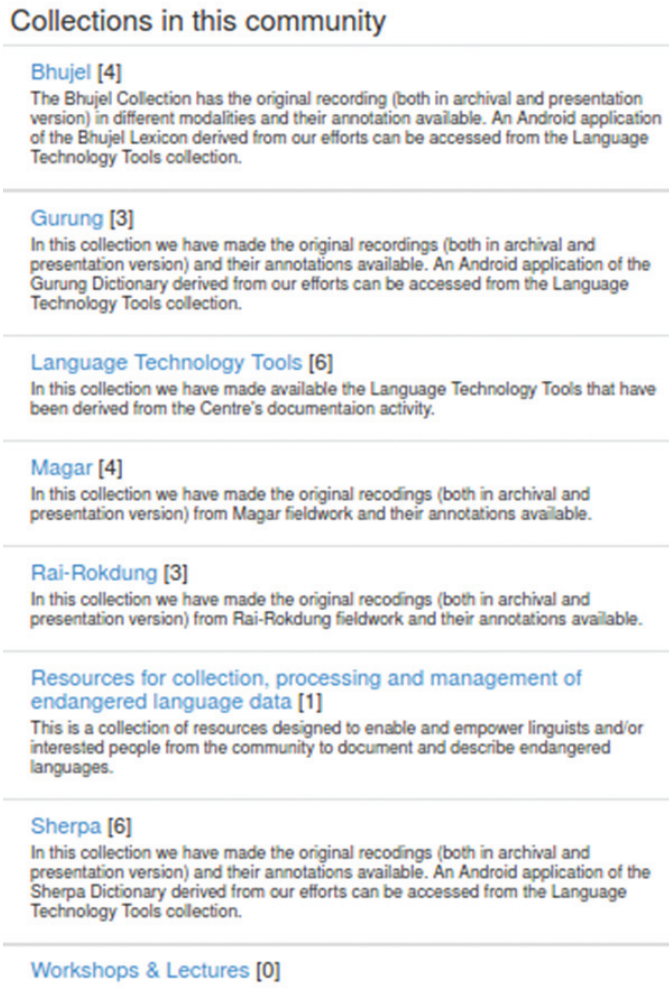
Literature on Sikkim and Eastern Himalayas

SiDHELA सिधेला 27

Sikkim-Darjeeling Himalayas Endangered Language Archive(SiDHELA) is an archive of the linguistic resource of the languages spoken in the Himalayan Biodiversity hotspot of Sikkim and North Bengal.

सिक्किम-दार्जिलिङमा बोलिने लुप्तप्राय हिमाली भाषाहरुको सङ्ग्रह (सिधेला) सिक्किम र उत्तर बङ्गालका जैव विविध क्षेत्रमा पाइने भाषिक संसाधनहरुको अभिलेख हो।

Figure 1. Communities in Sikkim university’s digital repository.



Collections in this community

Bhujel [4]

The Bhujel Collection has the original recording (both in archival and presentation version) in different modalities and their annotation available. An Android application of the Bhujel Lexicon derived from our efforts can be accessed from the Language Technology Tools collection.

Gurung [3]

In this collection we have made the original recordings (both in archival and presentation version) and their annotations available. An Android application of the Gurung Dictionary derived from our efforts can be accessed from the Language Technology Tools collection.

Language Technology Tools [6]

In this collection we have made available the Language Technology Tools that have been derived from the Centre's documentaion activity.

Magar [4]

In this collection we have made the original recodings (both in archival and presentation version) from Magar fieldwork and their annotations available.

Rai-Rokdung [3]

In this collection we have made the original recodings (both in archival and presentation version) from Rai-Rokdung fieldwork and their annotations available.

Resources for collection, processing and management of endangered language data [1]

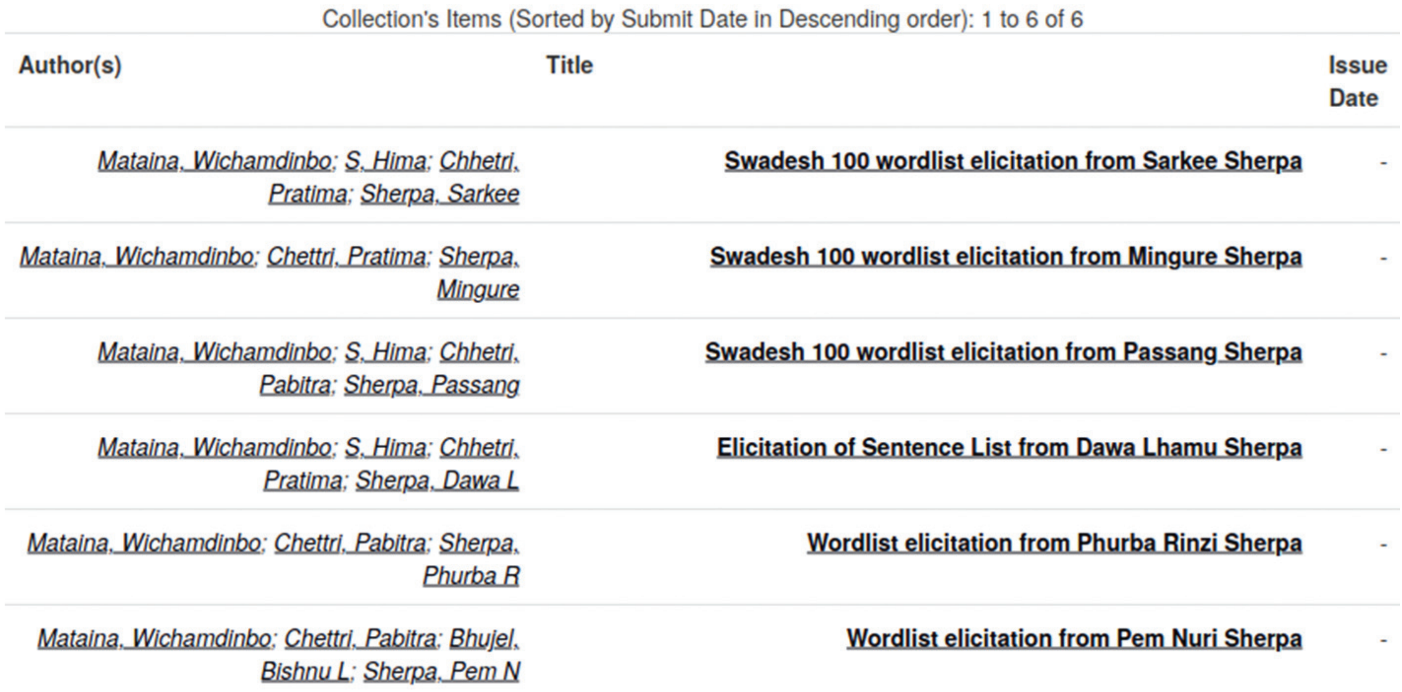
This is a collection of resources designed to enable and empower linguists and/or interested people from the community to document and describe endangered languages.

Sherpa [6]

In this collection we have made the original recodings (both in archival and presentation version) and their annotations available. An Android application of the Sherpa Dictionary derived from our efforts can be accessed from the Language Technology Tools collection.

Workshops & Lectures [0]

Figure 2. Collectons in SiDHELA.



Collection's Items (Sorted by Submit Date in Descending order): 1 to 6 of 6

Author(s)	Title	Issue Date
Mataina, Wichamdinbo; S, Hima; Chhetri, Pratima; Sherpa, Sarkee	Swadesh 100 wordlist elicitation from Sarkee Sherpa	-
Mataina, Wichamdinbo; Chettri, Pratima; Sherpa, Mingure	Swadesh 100 wordlist elicitation from Mingure Sherpa	-
Mataina, Wichamdinbo; S, Hima; Chhetri, Pabitra; Sherpa, Passang	Swadesh 100 wordlist elicitation from Passang Sherpa	-
Mataina, Wichamdinbo; S, Hima; Chhetri, Pratima; Sherpa, Dawa L	Elicitation of Sentence List from Dawa Lhamu Sherpa	-
Mataina, Wichamdinbo; Chettri, Pabitra; Sherpa, Phurba R	Wordlist elicitation from Phurba Rinzi Sherpa	-
Mataina, Wichamdinbo; Chettri, Pabitra; Bhujel, Bishnu L; Sherpa, Pem N	Wordlist elicitation from Pem Nuri Sherpa	-

Figure 3. Collectons in SiDHELA.

who has a reasonable linguistic ability. The community collaborated with the centre to document the fading language use among their people. They would expect the record of the linguistic knowledge documented by the centre, to be the basis for new language learning materials that can be used to revitalise their language. Hence SiDHELA is conceived as a repository to provide access to resources on the region’s endangered languages. As a part of this effort, SiDHELA is designed to hold not just the language data but also a platform for the dissemination of linguistic resources like android dictionary application.

Linguistic data as a source of Cultural Documentation: As language documenters linguist often assume that data from language documentation is of primary relevance for the linguistic studies. Linguist approach language archives as a warehouse of linguistic data that would help us solve the problem that linguist quibble about. However, records generated from the documentation is primarily multipurpose, a discussion between a fieldworker and language consultant on the origin myth of Rai Rokdung community is not just worth for its narration sample but also for the folk history and belief system it en composes. Similarly, a video record of the rituals of ‘kul’ clan pooja of Magar is not just a record of the ritual use of language but also of the communit’s practices. These records will be an important source of document in the community’s effort to revive and maintain their cultural identity, as these community’s are forced into mainstream culture. Holton observed this[23] and stated that language archives are accessed by the community members primarily as a cultural repository. SiDHELA is thus conceived as a platform for digital preservation of the linguistic and cultural diversity of the region.

A platform for Scholarly Communication: Another motivation that is behind SiDHELA is to create a platform to enhance the scholarly communication on the endangered languages of the region. Through this platform we hope to share resources designed to enable and empower linguists or interested people from the community to elicit, collect, digitise, process and manage linguistic data (words, sentences, narrations, sociolinguistic data, ethnolinguistic information etc) from the endangered languages of the region. As part of this we make available, manuals with detailed instructions on the use of specific software and digital apps for various steps of the process and hopes to cover the entire gamut of documentation activities from the stage of elicitation to archiving for public distribution of the linguistic data.

5. FROM CONCEPT TO CREATION

Long Term Preservation: The primary goal for any endangered language archive is the long term preservation of the data. Long term preservation requires not just state of the art technology but institutional and financial viability too. The present funding mechanism for Indian universities make this a challange. Short term, project based grant schemes are gives money for a period of three to five years. Such short term grants are not enough as the preservation of data is aimed for atleast the next fifty if not for more. To atleast partially address this problem one of the early decision the Centre made is to create SiDHELA on the existing resources of the university. Using the existing infrastructure to create an archive means the archives would share the server space and repository software of the university to ensure the hosting of data. However, using the digital infrastructure of the university makes the archive financially more viable, as the university will take up the responsibility of maintaining the server and

Table 1. CELSU metadata scheme

Label	Definition/Interpretation	Dublin core mapping
Identifier	This should be a unique identifier. It should be same as the file name.	dc.identifier
Title	for the community/collection/resource.	dc.title
Date	date of recording.	dc.date
Place	should be the place where the file was created, esp. for recordings.	dc.coverage.spatial
Source	Source of the data (How is the data sourced? Self or Others? if others please mention their name/or organization name.	dc.source
Publisher	An entity responsible for making the resource available.	dc.publisher
Relation	Reference to related objects in the archive like agreement, associated files(like transcription(TR) and traslation(TL), reviews, photographs, etc.	dc.relation
Researcher	The creator of the data.	dc.contributor.researcher
Creator	A person other than creator responsible for making research contributions to the item.	dc.contributor.author
Consultant	A person responsible for making contributions to the content of the resource language.	dc.contributor.consultant
Language(s) used	language file is in	dc.language
Resource language	language "of interest"	dc.subject.language
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
Genre*	describingribing intellectual content	dc.subject.classification
Discourse_Genre*	specifically about (recorded) discourse	dc.subject
Description	A brief description about the resource	dc.description.abstarct
Elicitation Method	State the Elicitation method	dc.description.elicitation
Type	Audio/Video/Image/Text	dc.type
O.S. Requirement	An operating system required to use a software resource.	dc.format.os
Keywords	Keyword describing the resources.	dc.subject.key
Format	Mention File format like '.jpeg' '.mp4' '.wav'	dc.format.mimetype
Size	Size of the file in MB or GB	dc.format.extent
Length	Length of the audio/video file	dc.format.duration
Pages	(only for documents) No of Pages	dc.format.pages
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

repository platform. The design implication of this decision was the choice of Dspace as a platform for archives.

Dspace: DSpace is an open source digital repositories software that is used widely in India for creating Digital repositories. It is an open source easy to deploy repository platform. It allows for easy customisation and can handle any format of data from text documents to digital videos. Its web -based submitting system is easy to use for researchers to create an archival deposit. Each deposited files is stored a bitstreams and along with its technical information and associated description. The bitsream, and their associated technical and descriptive information are together called as an item. The item's exposed metadata is indexed for browsing and searching. Related items are organised in to an collection. Collections are then embedded into a Community the highest level of the DSpace content hierarchy. This hierarchical structure of Dspace was very helpful in creating SiDHELA. SiDHELA is a special thematic community in the Sikkim University's Digital repository (Fig. 1). Under it there are five language collections namely: Bhujel, Magar, Gurung, Rai-

Table 2. New addition required to the DC schema in DSPACE

DC element and qualifier	Scope notes
dc.language.iso639-3	Three-letter ISO 639-3 Codes for Identifying Languages also commonly known as Ethnologue code.
dc.subject.key	Keyword describing the resource.
dc.subject.language	A language which the content of the resource describes or discusses.
dc.description.elicitation	State the Elicitation method used for data elicitation.
dc.format.characterencoding	State the Font Name used in the file. If these are special fonts that are downloadable give us the link to it too.
dc.contributor.consultant	A person responsible for making contributions to the content of the resource language.
dc.contributor.researcher	A person other than creator responsible for making research contributions to the item.
dc.format.os	An operating system required to use a software resource.
dc.format.duration	Length of the audio/video file

Rokdung and Sherpa (Fig. 2). In these collection we have made available, the original recordings from these languages (both in archival and presentation version) and their annotations in structured formats.

Apart from the above-mentioned collections as a part of our effort to aid in language revitalisation, we have made the Android dictionaries derived from our works under Language Technology Tools collection. Finally, to encourage transparency and share our knowledge on language documentation we have made our metadocumentaion available under the collection: 'Resources for collection, processing and management of endangered language data'.

Appraisal and accession: Appraisal is one of the core archival function, at SiDHELA appraisal is done at the item level. The appraisal condition followed right now are not holistic. Presently, it takes into consideration the item's digital preservationability i.e. the standard of files of encoding. Materials collected and documented by the Centre over the past three years are accepted for archiving only as bundles. Each bundle must contain the following files:

- Archival version of the item
- Presentation versions of the item
- Structured Annotation file of the item
- PDF of the annotation file

The primary appraisal condition for each file in the bundle is as follows:

- The Archival versions of the item must be complete, lossless, and unedited to the extent possible; For Data recorded after March 2020 archival version of the item should be deposited into an archive as soon as they are created.
- An annotation file must minimally contain transcription and translation in at least one of the gloss languages(English or Nepali). Annotation files accepted for archival submission are Praat TextGrids and ELAN EAF for audio and video records; Flextext and Lift Lexicon are the only accepted formats for annotated text corpora and lexical corpora. PDFs are accepted only as supplement to the structured annotation format
- Presentation versions of the record for audio files must be MP3 with minimum 128 kbps encoding and video files must be in MP4 format with frame 960 width x 540 height
- Each bundle must be described using the CEL, SU metadata scheme.

Arrangement and description: The principle of provenance and original order is the guiding principle for the organisation of items into collections. Accordingly records created from each linguistic community are kept together and distinguishable from the records of the other linguistic community. That is, item are not grouped together based on their content or themes but based on the subject language being documented. In accordance with this principle, as shown in Fig. 3 all records created on the Sherpa language are archived in the Sherpa collection.

Items deposited to each collection are described as per CELSU metadata scheme. As per the CELSU workflow, each file created in the process of documentation needs to be metadataed right after their creation. This step in the data management workflow is crucial to achieving progressive archiving. The CELSU metadata (Table 1) scheme uses all the 15 Dublin core elements with the necessary qualifiers to adapt it. It is used to adequately describe the attributes of resources the center has produced. In total twenty five fields of information are used to describe the data 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.

This metadata scheme is implemented in the Dspace at two levels. First the existing metadata registry of the Dspace Dublin core metadata schema is customised by adding the custom metadata fields in Table 2.

After this a custom input form is created based on the updated metdata scheme, and this custom input form is defined as the submission form for the collections in SiDHELA. Thus the SiDHELA community in the Sikkim University's digital repository is a special community which uses the CELSU metadata scheme and has custom input form assigned for its collection to meet the descriptive standards.

Apart from this each languages collection in the SiDHELA community also presents information on the Name and contact information of the primary creator, description of the scope and duration of the project, date range during which the records were created, ISO-639 codes for the language(s) documented in the records, names of language(s) or dialect s documented in the records and sociolinguistic information relevant to the language.

6. CHALLENGES AND FUTURE

SiDHELA's recent launch during the Sikkim University's foundation day on the July 2 2020 marked the beginning in a new phase of development. The SiDHELA has so far able to achieve customisation at the metadata and submission process and successfully deploy the same. The launch of the archive should be seen as an invitation for further work and not as a culmination. SiDHELA's web based submission and updating process has moved us closer to establishing a progressive archive. However several key design challenge are yet to be worked on in the development of SiDHEAL. Some of the challenges that SiDHELA would take up in the near future are: Version histories - to track and present information on modification to the submission; Access protocol - a sophisticated access protocol that restricts access to sensitive materials necessary to safeguard personal and cultural rights of the linguistic communities; and a onsite feedback method that would allow users to suggest edits and enable them to actively participate in the knowledge creation.

REFERENCES

1. Krauss, M. The world's languages in crisis. Language 68, 1992, **1**, 4-10. https://sustainableunh.unh.edu/sites/sustainableunh.unh.edu/files/images/Krauss(1992).pdf (accessed on 17/07/2020).
2. Campbell, L.; Nala, H.L.; Eve, O.; Sean S. & Kaori, U. New knowledge: Findings from the catalogue of endangered languages, 2013. http://hdl.handle.net/10125/2614 (accessed on 17/07/2020).
3. Moseley, C. Atlas of the World's languages in danger. Unesco, 2010. http://www.unesco.org/culture/en/endangeredlanguages/atlas (accessed on 17/07/2020).
4. Himmelmann, N. Documentary and descriptive linguistics. Linguistics, 1998, **36**, 161-196.
5. Bird, S. & Gary, S. Extending dublin core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 2003b, **37**, 375–388.
6. Austin, P.K. Data and language documentation. In: Essentials of language documentation, Jost Gippert, Nicholas P Himmelmann and Ulriek Mosel, De Gruyter, Berlin, 2006, 87-112.
7. Johnson, H. & Arienne, D. Customizing the IMDI metadata schema for endangered languages. *In* Proceedings of The International Conference on Language Resources and Evaluation. 2002. http://www.mpi.nl/lrec/2002/papers/lrec-pap-05-JohnsonDwyer.pdf (accessed on 17/07/2020)
8. Bird, S. & Gary, S. Seven dimensions of portability for language documentation and description. *Language*, 2003, **79**, 557–582.
9. Nathan, D. & Austin, P. Reconceiving metadata: Language documentation through thick and thin. In Peter K. Austin(ed.) language documentation and description, vol 2. SOAS: London.2004 pp. 179-188
10. Good, J. Data and language documentation. *In* Peter K. Austin & Julia Sallabank (eds.), The Cambridge handbook of endangered languages. Cambridge University Press, Cambridge, 2011, 212–234.
11. Nathan, D. Digital archiving. In Peter K. Austin & Julia Sallabank (eds.), The Cambridge handbook of endangered languages. Cambridge University Press. Cambridge. 2011. pp. 255–273.
12. Thieberger, N. Anxious respect for linguistic data: The pacific and regional archive for digital sources in endangered cultures (PARADISEC) and the resource network for linguistic diversity (RNLD). In Margaret Florey (ed.), Endangered Languages of Austronesia. Oxford University Press. Oxford., 2010, 141–158.
13. Thieberger, Nicholas & Berez, Andrea L. Linguistic data management. In The Oxford Handbook of Linguistic Fieldwork, Nicholas Thieberger,2012. Oxford: Oxford University Press.
14. Nathan, D. Archives 2.0 for endangered languages: From disk space to MyS- pace. *Int. J. Humanit. Arts Comput.,* 2010, **4**(1–2), 111–124.
15. Nathan, D. Access and accessibility at ELAR, an archive for endangered languages documentation. In David Nathan & Peter K. Austin (eds.), Language Documentation and Description. In Special Issue on Language Documentation and Archiving. SOAS. London., 2014, **12**, 187–208.
16. Linn, M.S. Living archives: A community-based language archive model. In David Nathan & Peter K. Austin (eds.), Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving. SOAS. London., 2014, 53–67.
17. Wasson, C.; Holton, G. & Roth, H. Bringing user-centered design to the field of language archives. Lang. Doc. Conserv., 2016, **10**, 641-681.
18. The language archive. https://archive.mpi.nl/tla/ (accessed on 17/07/2020).
19. Endangered languages archive at SOAS University of London. https://elar.soas.ac.uk/ (accessed on 17/07/2020).
20. Pacific and regional archive for digital sources in endangered cultures. 2016. https://www.paradisec.org.au/ (accessed on 17/07/2020).
21. The archive of the indigenous languages of Latin America. https://ailla.utexas.org/ (accessed on 17/07/2020).
22. Nathan, David. "Progressive archiving: theoretical and practical implications for documentary linguistics". 2013. https://scholarspace.manoa.hawaii.edu/handle/10125/26115 (accessed on 17/07/2020)
23. Holton, G. Mediating language documentation. In David Nathan & Peter K. Austin (eds) Language Documentation

and Description, vol 12: Special Issue on Language Documentation and Archiving. London: SOAS., 2014, 37-52.

CONTRIBUTOR

Dr R. Karthick Narayanan, is a Research Associate cum Digital Archivist, Centre for Endangered Languages, Sikkim University. He holds Ph.D from Jawaharlal Nehru University, New Delhi. His research areas are Socio-linguistic of endangered languages, documentation and description of lesser known languages of india, lexicography in lesser known languages of india, language maintenance and revitalisation, linguistic anthropology, digitisation of linguistic data, data management and archiving for linguistics.