Towards an agenda for open language archiving

Bird, Steven; Simons, Gary F.

Published in:
Proceedings of the International Workshop on Digital Language Archives

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

LANG ARC 2021

1st International Workshop on Digital Language Archives

PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON DIGITAL LANGUAGE ARCHIVES: LANGARC 2021

UNT

JCDL 2021

acm Association for Computing Machinery

IEEE

Proceedings of the International Workshop on Digital Language Archives:

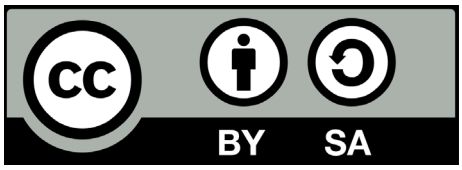
LangArc 2021

Virtual Format
September 30 – October 1, 2021

Workshop Co-Chairs:
Oksana L. Zavalina, Shobhana L. Chelliah

Proceedings Chair:
Oksana L. Zavalina

University of North Texas
Denton, Texas



Welcome from the Workshop Co-Chairs

It is our pleasure to share with you the Proceedings of the 1st International Workshop on Digital Language Archive (LangArc 2021)! The proceedings include 14 peer-reviewed accepted submissions from Asia, Australia, Europe, North and South America. The workshop, held as a virtual event on September 30, 2021, US Central time (September 30-October 1, 2021, Coordinated Universal Time UTC), is part of the ACM/IEEE Joint Conference on Digital Libraries 2021 https://2021.jcdl.org/ .

This interactive virtual workshop seeks to address a growing need. It explores a broad scope of issues related to digital language archives -- digital libraries that preserve and provide online access to language data. The objective of this workshop is to bring together researchers, practitioners, educators, and students from around the world who are currently working or are interested in working in different areas related to collecting, archiving, curating, organizing, and providing access to born-digital or digitized language data, and evaluation of digital language archives. The workshop will help foster collaborations among information professionals; library and information science, linguistics, data science, computer science, and humanities researchers; educators; representatives of language communities (including indigenous communities, refugees, speakers of under-resourced languages); and other interested audiences. The event is expected to become the first one in the series of regular workshops focused on the digital language archives.

We hope you find these proceedings interesting and useful and will consider attending or actively participating by authoring submissions for the upcoming meetings of the International Workshop on Digital Language Archives.

Dr. Oksana L. Zavalina, Associate Professor at the Department of Information Science at the University of North Texas
Oksana.Zavalina@unt.edu

Dr. Shobhana L. Chelliah, Distinguished Professor at the Department of Linguistics and the Associate Dean for Research and Development at the University of North Texas College of Information
Shobhana.Chelliah@unt.edu

Table of Contents

A website is a website is a website: Why trusted repositories are needed more than ever

Vera Ferreira
Endangered Languages Documentation Programme
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin Germany
vera.ferreira@bbaw.de

Leonore Lukschy
Endangered Languages Documentation Programme
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin Germany
lukschy@bbaw.de

Buachut Watyam
Research Institute for Languages and Cultures of Asia
Mahidol University
Thailand
buachut.bcw@gmail.com

Siripen Ungsitipoonporn
Research Institute for Languages and Cultures of Asia
Mahidol University
Thailand
siripen.ung@mahidol.edu

Mandana Seyfeddinipur
Endangered Languages Documentation Programme
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin Germany
mandana.seyfeddinipur@bbaw.de

ABSTRACT

Over the last two decades there has been a surge in activists, linguists, anthropologists, documenters digitally recording endangered language use. These unique records often are uploaded to corporate social media sites or to privately run websites. Despite popular belief, uploading these materials to a server does not mean they are archived and preserved for future generations. In this paper we discuss the differences between professional archiving systems and content management system (CMS) based approaches to making language materials accessible. Looking at the example of the *Archive of Languages and Cultures of Ethnic Groups of Thailand* we discuss the benefits of a Mukurtu based community website, and how linking it to a professional archive can ensure long-term preservation of precious and unique language materials.

CCS CONCEPTS

•Information systems ~ Information systems applications ~ Digital libraries and archives •Information systems~Information storage systems~Storage management~Information lifecycle management •Information systems~World Wide Web~Web interfaces •Information systems~Information retrieval~Users and interactive retrieval~Search interfaces

KEYWORDS

Digital archiving, Community archives, CMS, Archiving systems, Data preservation

1 Introduction

Of the 7000-7500 languages spoken today less and less are learned by children who instead learn majority languages. Once children do not learn the language of their heritage, the fate of the language is sealed.

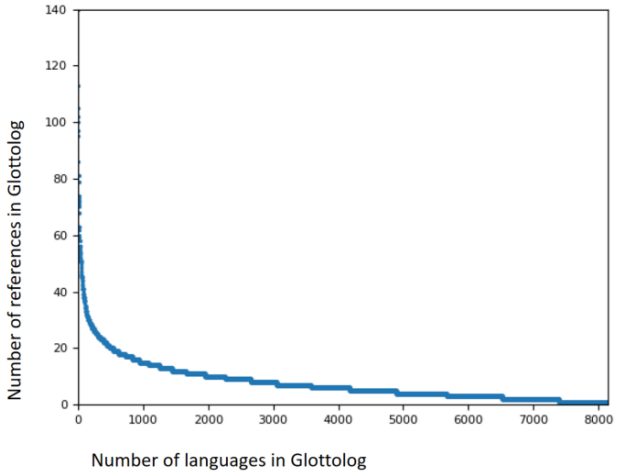


Figure 1: References in Glottolog (data extracted from [3])

Glottolog's bibliographic collection of linguistic descriptive works indicates that for about 35% of the languages there is a full grammar, for 25% there is a sketch grammar, of the remaining 40%,

12% have received some attention, meaning there is a dictionary, a translation of the New Testament or an in-depth discussion of a specific linguistic feature. For 28% there is only a wordlist or similar (see figure 1). Now, while there are some linguistic publications about languages of the world, what about recordings of language use, of people talking, chanting, praying, discussing, negotiating, narrating in situ?

OLAC data aggregator harvests metadata from around 60 language archives and it paints a grim picture with a lot of audio and video recordings and texts for some languages and nothing or very little for the majority of the languages of the world. This tells us something about the current situation of primary language materials available in archives.

Since the digital revolution many people have become active documenting languages and traditions tied to them, making recordings on their phones, on audio recorders and video cameras. If lucky, these materials do not end up on harddrives, laptops or CDs in private possession but the creators aim to make them available on the web to preserve them for posterity. Materials are uploaded to a variety of platforms, in some cases websites created for this specific purpose, in others commercial platforms such as Youtube, Vimeo or Facebook are used to publish recordings. Websites specifically created for the dissemination of language recordings need to be maintained and funded. If the person or group in charge of hosting and maintaining a website no longer has time, the interest, or runs out of funds, the website and the materials on it may be taken offline. Commercial platforms are problematic because it is at the discretion of a private company whether or not the materials stay online. Neither individual websites nor social media platforms have standardised metadata which means that even if these materials are online, they are not necessarily discoverable. And even if they are discoverable, there is no long-term preservation infrastructure. If digital materials are not migrated to newer formats, they will not be accessible in the future, which makes digital files extremely volatile (for more information on issues related to digital preservation in general see [1]; for issues in preserving language documentation data see [2]).

This is worrisome because many of these recordings are invaluable and may be the only recording of a ritual, of an elder, the holder of special knowledge, the shaman, or the singer of songs no one else remembers. Without these materials being professionally archived and preserved long-term, humanity's intangible heritage is at stake of being lost.

Another issue with individually created platforms is that they rarely rely on long-term funding. This is partly due to the academic funding cycle which is usually only for three years.

Language documentation should result in a multipurpose record, serving speakers or signers of the language documented, linguists and researchers from other disciplines, as well as the general public (see [4] and [5]). These different stakeholders need different ways of accessing materials, which is why websites geared towards specific groups can be very helpful, but it is vital to keep in mind that these websites can only offer a way of showcasing materials, and do not offer actual preservation. The same holds true for social media platforms, which might be valuable for presenting and disseminating materials, but cannot guarantee that these materials will be safeguarded in the long-term.

The fact that recordings are being uploaded to social media sites and privately run websites indicates that there is a clear need to increase the number of local archives to support local efforts in safeguarding documentary records created by a multitude of stakeholders. However, the implementation of sustainable archival infrastructures requires long-term financial and institutional commitment as well as technical expertise. In the meantime, a bottom-up approach whereby local scholars and activists set up a basic content management system and create and collections is one way to secure invaluable existing data, even though it must be clear from the very beginning that a website is not an archive, as reiterated before.

In a discussion on the differences between a website and an archive, and the need to keep the materials archived sustainably to guarantee their safeguarding, in the next sections we will present a bottom-up participatory approach for archive creation which we followed in the project *Archive of Languages and Cultures of Ethnic Groups of Thailand* supported by the Newton Fund. The major goal of the project, which was carried out in a collaboration between the Endangered Languages Archive (ELAR) and researchers from the Research Institute for Languages and Cultures of Asia at Mahidol University, was the implementation of a pilot small-case digital infrastructure for preservation and dissemination of indigenous linguistic materials and cultural heritage in Thailand.

2 CMS vs Archiving

Content management systems (CMS) available at the majority of web hosting servers have made the creation of websites available to a wide variety of users with different levels of technical know-how and are therefore well suited for crowd-sourcing materials collected by a number of individuals. However, using a CMS for adding recordings of a language to a website is not to be confused with archiving and preserving these materials. While a digital archive has both a preservation layer entailing the data conservation and maintenance workflows (like automated format migration, integrity checks, version control, etc.) illustrated in figure 2, as well as a presentation layer for displaying the data, a CMS lacks the preservation layer, focusing solely on displaying materials. A preservation layer is necessary as digital formats change rapidly, and it is key to migrate archived materials to the most up-to-date formats to guarantee their accessibility. In a professional archiving system this kind of migration can be, and normally is, automated. In a CMS, the migration of formats and their conversion needs to be done manually, which is error prone and time intensive.

Figure 2 illustrates the workflow connected to an archiving system, whereas Figure 3 highlights the components that are missing in standard CMS systems (or websites in general and social media platforms).

A website is a website is a website

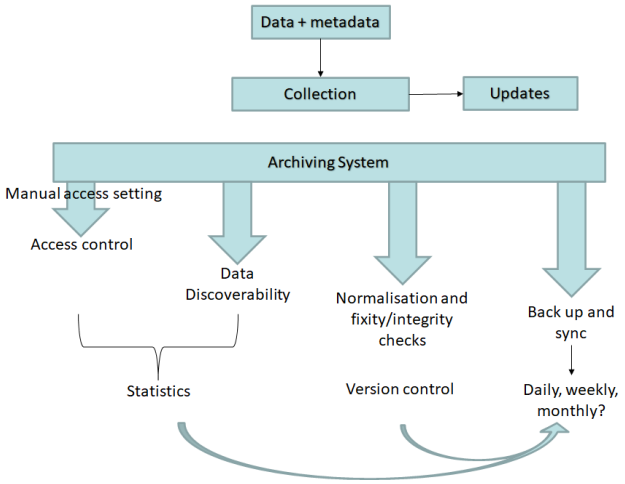


Figure 2: Workflow of an archiving system

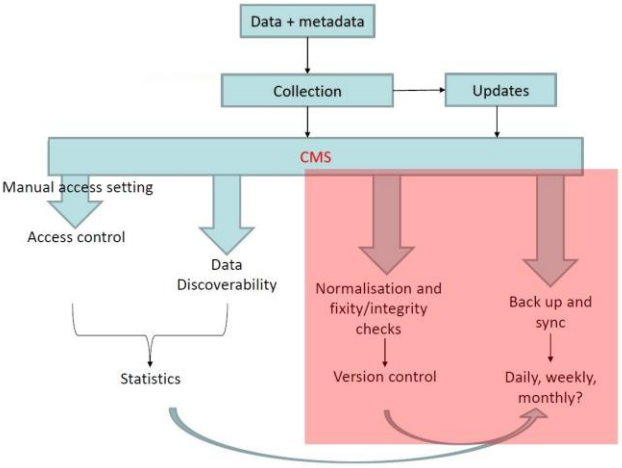


Figure 3: CMS vs archiving systems

The technical infrastructure and long-term funding necessary for archiving represent obstacles for the creation of local archives following archiving standards and best practices.

There are however intermediate solutions which combine less technical expertise and low costs with the basics of archiving, namely Mukurtu[1]. Mukurtu was developed out of the need for an easy to use out of the box system for communities to build up their own archives under their own leadership, maintaining data sovereignty. Mukurtu (meaning *dilly bag* or a safe keeping place for sacred materials in Warumungu language; see [6] and [7]) is a community-oriented CMS infrastructure based on Drupal (an open-source web content management) developed and maintained by the Center for Digital Scholarship and Curation at Washington State University. Mukurtu is a grassroots project aiming to empower local communities to manage, share, and exchange their digital heritage in culturally relevant and ethically-minded ways. It follows archiving standards by supporting and enforcing standard metadata schemas and formats; it has different levels of access, respecting data sensitivity and community wishes, in a user-friendly interface, ensuring CARE[2] and FAIR[3] data principles. It is easily customisable and localisable, allowing multilingual data presentation. Even though Mukurtu is still a CMS system without a preservation layer, it was developed based on archiving core principles and introduces its users to the basics of digital archiving.

3 Bottom-up participatory approach to archive creation

In this section we will present the project *Archive of Languages and Cultures of Ethnic Groups of Thailand*[4] as an example of an intermediate solution for digital archive creation which is based on a bottom-up participatory approach (see also [8]).

The *Archive of Languages and Cultures of Ethnic Groups of Thailand* came to fruition through a collaboration between ELAR and the Research Institute for Languages and Cultures of Asia (Mahidol University). The project was supported by the Newton Fund, with the aim to create a digital platform for the preservation and dissemination of indigenous linguistic materials and cultural heritage in Thailand. The richness of publicly unknown data collected over the years in Thailand, the activism that characterises the attitude of several language community members and scholars in the country, associated with the lack of a digital archive for language materials, led us to develop a community-oriented approach to archiving and to select Mukurtu as the digital platform. The major reason behind the selection of Mukurtu was the fact that even though Mukurtu is a CMS system, it enforces archiving best practices (like metadata consistency, file format unification, access granularity), and lays the ground for professional archiving. It is fully customisable (also in terms of language interface - the Mukurtu instance in this project was fully localised to Thai), simple to use and less academia-oriented. The resources (audio, video, pictures, texts), the languages and the speaker communities are in the foreground – which is an important feature to catch the attention of a broader audience and thus increase the usability of the archived materials.

In this particular case, Mukurtu was combined with a working and backup server, to guarantee the preservation of original primary data and the necessary format migrations.

After the digitisation of legacy materials from 15 different languages in Thailand (comprising audio, video, text, pictures), the materials were sorted according to their language and for each language and/or ethnic group a collection was created in Mukurtu. The materials that belonged together (for instance audio recordings and corresponding transcriptions) were organised in bundles and corresponding metadata was created. The metadata which followed a clearly defined structure, together with the resources, were loaded

[1] https://mukurtu.org (accessed on July 30, 2021)
[2] **C**ollective benefit, **A**uthority to control, **R**esponsibility and **E**thics
[3] **F**indable, **A**ccessible, **I**nteroperable and **R**eusable
[4] http://langarchive-th.org (accessed on July 30, 2021)

to the corresponding collections in Mukurtu and were made available for search and visualisation through the Mukurtu discovery layer. The data sets were also expanded with materials provided by researchers / community members not directly involved with the project. They were trained on data management and archiving mainly in the archiving workshops organised throughout the project. To facilitate the interaction with the archive, helpsheets on data curation and loading were created and made available through the archive website.

It is the only digital platform in Thailand which entails primary data for different ethnic groups and their languages in a consistent and methodological way. It is the first fully localised Mukurtu instance. It includes 15 collections on 15 different ethnic groups in Thailand (Hakka, Tak Bai, Gong, Pattani Malay, Chung, Saek, Chong, Urak Lawoc, Northern Khmer, Kasong, Nyah Kur, Kuy, Moken, Akha and Bisu) with more than 110 digital heritage items (5 hours of video, 7 hours of audio, 90 text files and around 140 pictures) and detailed metadata in Thai.

Processing the legacy materials and making them available digitally following best practices on data processing and metadata creation has a huge impact not only at socio-cultural level by contributing to the promotion and preservation of language diversity in Thailand but also at academic level by fostering research both on language documentation, linguistics in general and pedagogy (several teaching materials can now be created based on the data that was made available through the archive). Moreover, training community members on data curation and archiving, so that they can expand the database created within this project was key for the necessary empowerment that allows community members to have control over their language and culture and to take part in decision making processes.

However, this is only the first step towards sustainable digital archiving. As mentioned before, while Mukurtu enforces basic archiving workflows, it is merely a CMS rendering a presentation layer. Throughout the project, the users inputting data into the Mukurtu platform became aware of the importance of rich and standardised metadata, format consistency and format migration, i.e. they became aware of the core digital archiving principles and how they differ from a simple website creation. Due to the clear workflows and basic archiving principles implemented throughout the project, the shift to or the combination of Mukurtu with a professional archiving system with an automated preservation layer will be much easier in the future.

4 Conclusion

Platforms such Mukurtu offer an opportunity to break with the tradition of an extractivist North-South relationship, where data is kept securely in western academic institutions, while the rich materials compiled by language community members and activists in the Global South are not preserved and made accessible locally. Having a platform which can be easily localised, as is the case with Mukurtu, is already an essential step to make the materials discoverable by and accessible to their own authors and creators, strengthening the relationship between archives and their users – a tendency we could observe during the Thai project.

In terms of community archiving, the ideal scenario would be the combination of the functionalities offered by Mukurtu with an automated preservation layer or with the additional storage of the materials in a professional archive that guarantees their preservation and accessibility over time. The same applies to websites dedicated to individual languages or larger scale projects. While all of these efforts are important for making materials more easily accessible to communities and the general public and can be very valuable for crowd-sourced collection of materials, they need to be linked to or integrated in a professional archive to ensure that the data is preserved long-term.

REFERENCES

[1] Digital Preservation Handbook, 2nd Edition, https://www.dpconline.org/handbook, Digital Preservation Coalition © 2015.

[2] Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. LANGUAGE, 79 (3), 557-582. https://doi.org/10.1353/lan.2003.0149.

[3] Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. 'Glottolog Database 4.4'. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/ZENODO.4761960.

[4] Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. Linguistics 36:161-95.

[5] Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.), *Essentials of Language Documentation*. 1-30. Berlin: Mouton de Gruyter.

[6] Christen, Kimberly, Alex Merrill, and Michael Wynne. 2017. 'A Community of Relations: Mukurtu Hubs and Spokes'. D-Lib Magazine 23 (5/6). https://doi.org/10.1045/may2017-christen.

[7] Christen, Kimberly. 2015. Tribal Archives, Traditional Knowledge, and Local Contexts: Why the "s" Matters. Journal of Western Archives*:* Vol. 6, Iss. 1, Article 3. DOI: https://doi.org/10.26077/78d5-47cf.

[8] Ungsitipoonporn, Siripen, Buachut Watyam, Vera Ferreira, and Mandana Seyfeddinipur. 2021. Community Archiving of Ethnic Groups in Thailand. Language Documentation & Conservation 15, 267-284. Handle: http://hdl.handle.net/10125/24975.

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

Karthick Narayanan R
Independent Researcher

Madurai, India
0000-0001-5234-310X

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.

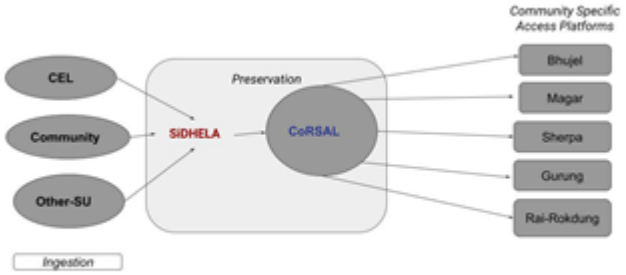
B. Moving the records from the field to the archive

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.

Track to the past: tracking workflows, versions, and citations of legacy language data

Tobias Weber
Graduate School of Language & Literature, Ludwig-Maximilians-Universität
München, Germany
tobias.weber@lrz.de

***Abstract*— This paper discusses three issues encountered with legacy language data in archives: First, the provenance of an artefact containing the data may be unclear, as well as all procedures that shaped its form(at) or contents. Second, legacy language data are often orphan data with opaque links to other versions, or texts providing more information on them and their contents. Third, these data predate methods of data citation, thus requiring retroactive ways of citation tracking. With a few adjustments to their infrastructures, digital archives can be used as a platform to track workflows, versioning, and citations of legacy language data.**

Keywords*— *legacy data, language documentation, linguistics, archiving, versioning, citation tracking, scientometrics, graph theory, anthropology

I. INTRODUCTION

Legacy materials are the outputs of past documentation projects [2]. As a result, working with these materials requires the researcher to understand the contexts of their creation, the history of their subsequent transformations, and the scientific as well as socio-cultural impact of the documentation project and its outputs. At the same time, relevant information is often scattered across different parts of archives and libraries, for example publications and raw data are kept in different repositories, while especially older documents might not be fully digitised or interoperable and metadata can be faulty. Digital language archives can offer tools and infrastructures to discover missing links, present data in context, and keep track of the histories of each artefact. This paper outlines three desiderata for language archives of legacy data. To ensure brevity, some issues pertaining to legacy materials are omitted from the discussion (for an in-depth discussion, see [19]).

II. WORKFLOWS

The necessity to keep track of workflows is not just linked to professional conduct or the aims of making research intelligible or even 'reproducible' [3]. Contend that full reproducibility of linguistic analyses cannot be reached. The goal should rather be to enable readers and future researchers to understand the – occasionally subjective – decisions we made and to give them the necessary information to assess, discuss, and evaluate them on the basis of the data. Computational tools can still support this endeavour, yet not in a mechanistic replication of results [14] The goal is to have data and research papers in archives or libraries which are still understandable '500 years from now' [20]. Moreover, it is an ethical consideration to allow a review of our methodology in data collection. This involves acknowledging all individuals who contributed to a dataset or a publication [1, 8]. Yet, despite a clear requirement for transparent workflows and complete a metadocumentation, some points of metadata may be missing for legacy materials. This may be due to unknown provenance of an artefact, metadata loss during copying or transcribing, lossy artefact types (including physical media like manuscripts or wax cylinders), changes in professional standards, or idiosyncratic workflows. It might appear easy to discredit past researchers whose datasets lack sufficient records of metadata but this is not always a sign of unprofessional conduct. On the contrary, there are settings where privacy concerns or insecure socio-political circumstances have had an impact on the metadata recorded by a researcher [17–19] – we can only interpret the legacy materials and the accompanying metadata if we know about the historical contexts of their creation.

One solution to these obstacles can be found in the curation of language datasets. This process can be facilitated by archive structures that make texts findable and accessible, so that curators can reestablish links between artefacts. The curation process itself should be informed by contextual information from history, anthropology, or sociolinguistics, and is less technologically focused than data curation in other disciplines. Certainly, computational tools can support the process, although it is more about the speakers [6], or the 'human in the loop' [4], and less about the data as such. Thus, individual artefacts can even be approached from the perspective of forensics [9]. Due to the textual nature of the artefacts, we can also apply skills from philology [13], a text-based science that aims to understand texts in their historical and socio-cultural contexts. It involves comparing, commenting, and questioning texts and learning more about the circumstances of their creation. Furthermore, this approach is not just occupied with real-world contexts but also with the 'linguistic context' – the *cotext* [5] – the text surrounding a word, sentence, or paragraph. As a result, the restoration of links between artefacts is necessarily involved in intertextual networks: Field diaries help us to understand audio recordings, manuscripts support their transcription, references to previous documentation frames research projects and their objectives. With the same view, we can also link raw data and publications, or individual publications as a part of the same abstract workflow [16].

Digital archives can support the tracking of workflows through several means. For recent additions, ontologies of contributor roles and persistent identifiers for individuals and

artefacts should be used for the metadata [11]. These should be necessarily thorough and supply information even on less prominent individuals, e.g. assistants who helped with transcription can leave noticeable traces in an artefact. This can also include knowledgeable scholars who are invited as external curators [21] and enrich a collection with their contextual awareness and information about research history. Community members may also be invited as curators, e.g. if they or their family members have been involved as consultants. In either case, the curator must, in turn, be credited for their work, as they leave their own traces in the dataset. These knowledgeable scholars can help to establish links between datasets and publications, and offer commentary based on the current state of research. Hyperlinking and referencing all relevant texts enables holistic treatment with a philological approach. The necessary requirements for this is transparency, including accessibility and findability of data.

III. VERSIONING

A central benefit of digital archives is their accessibility through the internet. Yet, before the internet allowed for global access to data, they have been disseminated on physical media. As a result, a recording of the same event or a transcription of the same narrative might be archived in different locations. But are those copies actually the same? Even if they were created from the same original file, they are not identical [12]. On the one hand, different technological solutions or media have an influence on the data (e.g. loss rate, localisation), and receiving archives or researchers may have contributed further edits, annotations, or transformations to the data. On the other hand, considering the importance of co- and context, seemingly identical copies of the same data in different archives cannot be identical if we consider their archiving context relevant. Different contextualisations in the respective archives might arise from tags about the dataset, the compilation into overarching collections or thematic units, or the presentational formats of each archive (e.g. scanned copies by different archives in different resolutions or with different defective pixels). If we continue this line of thought, the observation about identity also affects data in publications, thus also the work of libraries [16] – each publication has a version of the dataset that is unique to this publication. Consider, for example, formatting rules, different layers of analysis, transcription rules, translations in to various languages [7], the surrounding interpretation (in the cotext).

Versions of textual artefacts need to be collated and compared, in order to establish the contexts in which changes were administered. This comparative task is well-known as part of textual scholarship, e.g. applied to medieval manuscripts, versions of literary texts. Stemmatology creates graphs of different versions, with each node an (actual or hypothetical) original version from which all its child nodes derive. If we think of the nodes as individual research papers, publications, compiled datasets and corpora – all in their own contexts – the image of all versions becomes opaque. Yet, we cannot separate the task of identifying relationships between individual versions from their concrete use. In other words, the version of the data has an impact on their analysis and their interpretation, and, in the spirit of replicability, we need to know which version produced a result or conclusion. We need each version to contain information on its position in the tree graph, and its relationship to other nodes. This means, it needs to be aware of preceding and subsequent versions, and the transformations from the original to this version; inheritance of metadata across time (horizontally) and sub- and supersets of the fragment (vertically) [15]. In digital environments, this can be dynamically generated and displayed, yet with the change of the medium (e.g. to print), we lose access to the history behind the data.

Digital archives offer several opportunities to support the tracking of versions. First, they hold the original data and often citable with persistent identifiers. Second, archives have the infrastructure to keep full accounts of metadata – although it is debatable whether archives should bear the onus of tracking versions, they have the capacity to do so. Third, the display of data citation and different versions alongside the original data can be beneficial to the scholarly community who can access and assess different interpretations of data and identify potential discrepancies. A necessary requirement for offering this function is the availability of digital copies on the side of the publications, which may be facilitated through the inclusion of digital libraries and publishing houses [14].

IV. DATA CITATION TRACKING

As already mentioned in the previous sections, keeping track of versions is closely connected to tracking citations of data. On top of original citations, we also need to consider secondary citations, i.e. instances where data was copied from a publication and not from the original. Version tracking can help with this task and, considering its scope, highlight an important challenge to data citation tracking. Besides, we are potentially facing versions of data which are published in locations that are not accessible for citation tracking (e.g. community materials, blogs).Regarding online resources, Altmetrics [10] offer a possible approach to tracking citations in social media and on the internet. On the other hand, there are instances which may predate our infrastructure, i.e. citations before persistent identifiers were added to a dataset, publications which are not digitised or not included in databases. Since it is in the interest of the archive to keep track of the use of its data, a case for citation tracking by archives can be made. However, publishers and repositories need to support this endeavour by granting access to texts and cited references, especially for older publications which might not be fully digitised. This shows that the requirements are similar to those for version tracking, and that both procedures can be implemented alongside each other. For citation tracking, graphs can also be used to represent relationships between texts; combined with a copy of the version, its metadata, and all changes to the data itself, this becomes a powerful tool for researchers. At the same time, access to a holistic image of data use can prevent biases and misrepresentations, and allows all individuals who were part of the workflow to have their contributions appreciated and properly attributed.

V. CONCLUSION

Legacy data poses different challenges to archives than recently deposited datasets. Apart from ethical concerns about their provenance, the history of the artefact can be unclear, including processes of its creation, subsequent use, and citation. Yet, omitting legacy data from research or restricting access to them due to their unclear history should be the last

resort, as it means the loss of valuable knowledge and disregard to the creators' efforts, not least to that of the consultants' and communities'. The value of legacy materials needs to be appreciated through careful reconstruction using philological, anthropological, and historical knowledge and skills. Some of the required steps can be supported by computational methods, where the collaboration of digital archives and libraries is essential. At the same time, archives, libraries, and publishers stand to gain from transparent workflows, versions, and (data) citations of their resources. Legacy data must not be ignored and can, on the contrary, inform the design of tools that do not only work on recent data and metadata but also on historical records of our discipline. Creating this 'backwards compatibility' of legacy data with modern standards is a sign of our appreciation – the same appreciation we would want from future generations for our present-day deposits.

REFERENCES

[1] Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. https://doi.org/10.15497/rda00040.

[2] Peter K. Austin. 2013. Language documentation and meta-documentation. In Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation, Mari Jones and Sarah Ogilvie (Eds.). Cambridge University Press, Cambridge, 3–15.

[3] Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics 56, 1 (2018), 1–18.

[4] Steven Bird. 2020. Decolonising Speech and Language Technology. In Proceedings of the 28th International Conference on Computational Linguistics, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, 3504–3519. https: //doi.org/10.18653/v1/2020.coling-main.313

[5] John Cunnison Catford. 1965. A linguistic theory of translation: an essay in applied linguistics. Oxford University Press, Oxford.

[6] Lise Dobrin, Peter K. Austin, and David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 6. SOAS, London, 37–52.

[7] Jan Engh. 2006. Norwegian examples in international linguistics literature. An inventory of defective documentation. Universitetsbiblioteket i Oslo, Oslo.

[8] Alex O. Holcombe. 2019. Contributorship, Not Authorship: Use CRediT to Indicate Who Did What. Publications 7, 3 (2019), 1–11. https://doi.org/10.3390/publications7030048

[9] Gareth Knight. 2012. The Forensic Curator: Digital Forensics as a Solution to Addressing the Curatorial Challenges Posed by Personal Digital Archives. International Journal of Digital Curation 7, 2 (2012), 40–63. https://doi.org/10.2218/ijdc.v7i2.228

[10] Jean Liu and Euan Adie. 2013. Five challenges in altmetrics: A toolmaker's perspective. Bulletin of the American Society for Information Science and Technology 39, 4 (2013), 31–34. https://doi.org/10.1002/bult.2013.1720390410

[11] Steve Pepper. 2011. Ontologies in language documentation. In Language Documentation and Description, Julia Sallabank (Ed.). Vol. 9. SOAS, London, 199–218.

[12] Allen H. Renear and Karen M. Wickett. 2010. There are No Documents. Proceedings of Balisage: The Markup Conference 2010 5 (2010). https: //doi.org/10.4242/BalisageVol5.Renear01.

[13] Frank Seidel. 2016. Documentary linguistics: A language philology of the 21st century. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 13. SOAS, London, 23–63.

[14] Tobias Weber. 2019. Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics?. In 2nd Conference on Language, Data and Knowledge (LDK 2019) (OpenAccess Series in Informatics (OASIcs), Vol. 70), Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski (Eds.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 26:1–26:8. https://doi.org/10.4230/OASIcs.LDK.2019.26

[15] Tobias Weber. 2020. Metadata Inheritance: New Research Paper, New Data, New Metadata?. In Reframing Research Workshop Accepted Papers, Andrea Mannocci (Ed.). Zenodo. https://doi.org/10.5281/zenodo.4155362

[16] Tobias Weber. 2020. A Philological Perspective on Meta-scientific Knowledge Graphs. In ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, Mark Hall, Tanja Merčun, Boris Novikov, Christos Papatheodorou, Thomas Risse, Oscar Romero, Lucile Sautot, Guilaine Talens, Robert Wrembel, and Maja Žumer (Eds.). Springer International Publishing, Cham, 226–233. https://doi.org/10.1007/978-3-030-55814-7_19

[17] Tobias Weber. 2021. Consultant Identity in Historical Language Data: Anthroponyms as a Tool or as an Obstacle? In Proceedings of the International Onomastic Conference "Anthroponyms and Anthroponymic Researches in the Beginning of 21st Century", Anna Choleva-Dimitrova, Maya VlahovaAngelova, and Nadezhda Dancheva (Eds.). Bulgarian Academy of Sciences, Sofia, 165–175.

[18] Tobias Weber. 2021. Mind the Gap: Language Data, Their Producers, and the Scientific Process. In 3rd Conference on Language, Data and Knowledge (LDK 2021) (Open Access Series in Informatics (OASIcs), Vol. 93), Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, 6:1–6:9. https://doi.org/10.4230/OASIcs.LDK.2021.6

[19] Tobias Weber. forthcoming. Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian. In Language Documentation and Description, Lise M. Dobrin and Saul Schwartz (Eds.). Vol. 21. ELPublishing, London, forthcoming.

[20] Anthony C. Woodbury. 2003. Defining documentary linguistics. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 1. SOAS, London, 35–51.

[21] Anthony C. Woodbury. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving, David Nathan and Peter K. Austin (Eds.). SOAS, London, 19–36.

Linguistic Archives and Language Communities Questionnaire

Establishing (Re-)Use Criteria

Ilya Khait
Leibniz-Zentrum Allgemeine Sprachwissenschaft
Berlin Germany
ilya.o.khait@gmail.com

Leonore Lukschy
Endangered Languages Documentation Programme
Berlin-Brandenburg Academy of Sciences and Humanities
Berlin Germany
lukschy@bbaw.de

Mandana Seyfeddinipur
Endangered Languages Documentation Programme
Berlin-Brandenburg Academy of Sciences and Humanities
Berlin Germany
seyfeddinipur@bbaw.de

ABSTRACT

Digital language archives hold vast amounts of materials in endangered or marginalised languages. However, due to limitations in technical infrastructure and the design of these archives, the materials are usually not easily accessible to speakers of the languages represented or their descendants. With the goal to establish best practices for researchers archiving linguistic data, this paper presents a questionnaire designed to assess how archival materials can be made more readily available to language communities.

CCS CONCEPTS

• Information systems ~ Information systems applications ~ Digital libraries and archives • General and reference ~ Document types ~ Surveys and overviews

KEYWORDS

Language archives, Endangered languages, Heritage materials

1 Introduction

For many years scholars have conducted fieldwork and have worked with speakers and communities describing and documenting their language and knowledge. Recordings created during these projects have been archived in large digital archives usually housed in the global North in academic environments. The advent of the digital held the promise of democratisation of access to knowledge for everyone. However, 20-30 years down the line communities cannot easily access their own recordings in these digital archives. Studying and overcoming this issue falls within the scope of QUEST (Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data), a collaborative project that aims to establish curation criteria for digital language data for its subsequent use. In this paper, we present survey work on how communities access information digitally to understand how archives and researchers can ensure that the data they collect and preserve can be made accessible and discoverable to the very communities they come from.

We use the term ‘language archive’ to refer to digital archives of educational or memory institutions like universities or libraries, such as the members of the DELAMAN network, holding primary and secondary language documentation data, in the form of audio and video recordings, images, transcriptions, and other texts. The term ‘language community’ or simply ‘community’ is used to refer to speakers of endangered or marginalised languages and their descendants. When referring to researchers we use ‘outsider researcher’ for linguists or anthropologists who work with language communities they are not considered a part of.

The importance of making language materials such as video and audio recordings, as well as texts, available to the communities who provided the data has been recognised by a number of researchers (see [1]; [2]; [3]). Different attempts have been made to create archive interfaces that are designed for communities they serve [3]. AILLA, the Archive of the Indigneous Languages of Latin America, has a Spanish Interface making the materials discoverable in the most widely spoken language of Latin America. The Language Archive (TLA), holding the materials from the Volkswagen Foundation funded projects, developed portals for the public, and the Dane project developed a community portal [1].

Given the lack of electricity and digital connectivity, outsider researchers documenting languages have attempted leaving the recordings with the community on physical

storage devices such as tapes, CDs, hard drives, USB drives, or SD cards. The shortlivedness of the physical storage devices unfortunately meant that communities only have access to their own recordings for a short amount of time.

Long-term access can be provided through digital language archives, however the materials held there are not easily accessible and discoverable because of the way the interfaces are designed and the fact that the interface language is mostly in English. Several researchers have pointed out that large language archives are geared towards an academic audience rather than the broader public or the communities whose languages are represented in said archives (see [4], [5]).

Below we overview some known access biases and present the work on a questionnaire that surveys access to language materials and media usage in general by community members, then discuss the preliminary results obtained thus far and how archival materials can be made more readily accessible to communities in the future.

2 Restrictions

Informal interviews with linguists conducted between November 2020 and April 2021, who had carried out fieldwork in close collaboration with communities in Cameroon, Papua New Guinea, Peru and Colombia, and Vanuatu, highlighted the following obstacles to communities finding and accessing data in their languages.

2.1 Technical infrastructure

While internet coverage, particularly mobile, is spreading rapidly [6], communities speaking marginalised languages tend to live in areas with little access to the basic technical infrastructure needed to use digital media. The availability and affordability of electricity, digital devices and online access cannot be taken as a given. Even when the latter conditions are met at least to a certain extent, limits in web data volume, bandwidth or of the devices and software in use can also hinder access to media. There can be other obstacles as well, as e.g. in some areas, where mobile internet access is tied to certain platforms such as Facebook's Free Basics initiative [7], thus barring free surfing in the browser.

2.2 User skills and environment usability

The lack of computer literacy or the more basic written barrier, particularly in older generations, can hold language community members from discovering and accessing digital media. While generally more proficient in modern technology, younger people are likely to use mobile devices rather than computers. Practically, it means that in many cases data can be made available only through mobile-ready interfaces. It is also not uncommon that one's online experience is limited to a few popular applications.

Literacy is another barrier, discovering and accessing information takes place through reading and writing. To access information on the web literacy in a majority language is a requirement. Concerning digital language archives in particular, it is important to note the linguistic barrier with strong English bias tendency, the use of specific terminology, User Interface complexity and multilayerness that can sometimes be puzzling even for linguists. In addition, often additional steps like registering or requesting access for sensitive data makes data even more inaccessible. Although graded access has many advantages in terms of protecting speakers (see [8]), it also represents another barrier in navigating materials.

2.3 Discoverability

Discovering materials in archives is possible through text-based metadata. In addition both the catalogue and the metadata of large international digital language archives are mostly in English. Although there has been a push towards multilingual metadata in recent years, implementing this is still largely the responsibility of individual researchers. AILLA, the Archive of the Indigenous Languages of Latin America, stands out as an example of a large-scale language archive with bilingual metadata in English and Spanish.

Another obstacle to discovering materials is that the metadata categories and contents are grounded in linguistic categories which are relevant to mostly linguists. For descendants of speakers to find recordings of someone from their family, particular metadata such as a person's full name might be relevant. However, some informed consent protocols and recent data protection requirements lead to anonymisation of the speakers making their recordings not discoverable.

3 Designing the questionnaire

The primary aim is to develop an understanding what type of material communities would access through what digital medium. This understanding will allow in turn to build interfaces for e.g. social media channels serving the recordings relevant to the community as it is likely that recordings of an elder telling a story is of more value than some linguistic elicitation.

Linguistic Archives and Language Communities
Questionnaire

LangArc-2021

The secondary aim is to design a guide for researchers going into the field, so that they may address the question of how to make materials accessible to the community they work with at the start of the project rather than retroactively. This pertains to collecting community metadata which will be very different from metadata relevant to linguists.

In order to assess this, we designed a questionnaire divided into the following seven sections:

1) General Information

2) Materials of Interest

3,4) Existing Recordings

5) Recordings available online

6, 7) Connectivity, devices, and platforms

In order to obtain answers about communities who might currently not have access to the internet, we designed a slightly different version of the questionnaire for outsider researchers, such as linguists working with a community. The questionnaire targeted at language community members has thus far been translated into English, Russian, Spanish and French. It consists of 27 questions and takes ca. 10 minutes to complete. Links to the forms are provided in the appendix below.

It should be noted that there is a bias in who will answer the questionnaire as it is currently only being distributed online as a Google Form, and is only available in a limited number of major languages.

4 Preliminary results from the questionnaire

We will report preliminary data from 12 respondents (Three outsider researchers, 9 language community members (seven in English and two in Spanish). The respondents represent the following languages and regions: Yoruba, Igbo (Nigeria), Quechua (Bolivia), Bora (Peru), Punjabi (Pakistan), Shugnani (Tajikistan), Rejang (Indonesia), Khakas, Negidal (Russia, Altai and Far East respectively), Tsova-Tush (Georgia), Guernsey French (UK), and Irish (Ireland). Of the respondents who are part of a language community, five are female and four are male; five persons are in their twenties, three in their forties, and one respondent is in their late fifties. Five are native speakers, while three speak the language fluently and one speaks it a little. Below follows a brief summary with highlights of the results at hand, both from outsider researchers and community members.

4.1 Community interest in recordings

Popular genres of interest are

- dictionaries (twelve respondents)
- language learning materials (eleven respondents)
- knowledge about animals and plants, family and kinship and local history (ten respondents)
- stories, conversations and crafting knowledge (nine respondents)
- rituals (eight respondents)
- linguistic materials (seven respondents)
- knowledge about hunting, fishing and harvesting (six respondents)

As for media, text and video are convenient for most (ten each), leaving audio (nine) and images (eight) slightly behind.

4.2 Recordings shared with the community

All but one respondent report that recordings were made available to the language community. While website links are frequent (eight), digital carriers are used as often (a USB stick is mentioned four times and others such as hard drive, SD-card, DVD, CD appear once each). Materials on paper were distributed in four cases, and two respondents mention an analogue cassette. Nine respondents affirm that community members tried to access these materials and nine that they are aware of recordings of their language online (in most cases naming popular social media). Only in six cases people looked in digital archives and of those two encountered difficulties and one did not find any data.

4.3 Connectivity and communication

All respondents state that most people in the community have Internet access. Moreover, the internet is so good everywhere that they all can watch videos. It looks like most people have mobile phones, and seemingly in many cases these are actually smartphones. PCs are less common and tablets are rare.

Most popular platforms are

- WhatsApp (all twelve)
- Facebook (eleven) and Instagram (ten)
- Twitter and Skype (six)
- Tik Tok (four),
- Telegram (three)

Most language communities (ten) use phones to communicate, but almost as common are messengers and social media (nine) and text messages (eight). Somewhat

less widespread are emails (seven), video calls (six) and voice messages (four). The post is named only in two cases.

5 Discussion

The answers obtained thus far show active use of modern digital media among marginalised language communities all around the globe, with clear preference given to mobile devices and popular messengers and social networks.

As discussed in section 3, there is a bias in respondents due to the languages and means of distribution of the questionnaire as well as the small selection of responses collected so far. Moreover, community members filling in the questionnaire might be more likely to already have an interest in accessing materials in their language. In order to get a broader sample of answers, it would be ideal to distribute a printed version of the questionnaire, potentially via researchers conducting fieldwork.

Making materials accessible needs to be an ongoing process, particularly as communities' access to technical infrastructure is rapidly evolving. Rather than trying to teach speakers to access their data in complex archiving environments, materials should be made available to them via platforms they already know to navigate. A possible development in this direction would be, for example, designing language archive chat bot interfaces for popular messengers to search and deliver the data.

To all effects and purposes, researchers should discuss making materials available at the beginning of a documentation project to best serve community needs.

6 Appendix: Questionnaire links

Community members:

- English
  https://forms.gle/dqvyGNmVHA5uoBCMA
- French
  https://forms.gle/f5uxQuYKZpMhCDHW7
- Russian
  https://forms.gle/BNLm7BHe2CgYJecA9
- Spanish
  https://forms.gle/LKGWw2DHTzjYNcDq6

Outsider researchers (English):

- https://forms.gle/44HXXkGrcJJCq4xA8

ACKNOWLEDGMENTS

The QUEST project is funded by the German Federal Ministry of Education and Research (BMBF). We are very thankful to the respondents for dedicating their time and information to assist us. We are grateful to our colleagues, Dr. Jocelyn Aznar, Prof. Dr. Manfred Krifka, and Dr. Frank Seifart for fruitful discussions and valuable suggestions. Dr. Aznar is also to be especially thanked for assistance with translation.

REFERENCES

[1] Trilsbeek, Paul, and Dieter van Uytvanck. 2009. 'Regional Archives and Community Portals'. *IASA*32: 69–73.

[2] Wasson, Christina, Gary Holton, and Heather S. Roth. 2016. 'Bringing User-Centered Design to the Field of Language Archives', December, 641–81.

[3] Nordlinger, Rachel, Ian Green, and Peter Hurst. 2019. 'Working at the Interface: The Daly Languages Project'. Edited by Linda Barwick, Jennifer Green, and Petronella Vaarzon-Morel. *LD&C Special Publication No. 18: Archival Returns: Central Australia and Beyond*, Indigenous music of Australia, , 193–216..

[4] Holton, Gary. 2012. 'Language Archives: They're Not Just for Linguists Any More'. *Language Documentation & Conservation*Special Publication No. 3: Potentials of Language Documentation: Methods, Analyses, and Utilization: 105–10.

[5] Woodbury, Anthony C. 2014. 'Archives and Audiences: Toward Making Endangered Language Documentations People Can Read, Use, Understand, and Admire'. Edited by David Nathan and Peter K. Austin. Language Documentation and Description, no. 12: Special Issue on Language Documentation and Archiving: 19–36.

[6] Roser, Max, Hannah Ritchie, and Esteban Ortiz-Ospina. 2020. 'Our World in Data: Internet'. Our World in Data. 2020. https://ourworldindata.org/internet.

[7] Henning, Maximilian. 2019. 'How the Global South Can Protect Itself from Digital Exploitation'. *LATITUDE*, 2019. https://www.goethe.de/prj/lat/en/dis/21670998.html.

[8] Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Johnathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, et al. 2019. 'Public Access to Research Data in Language Documentation: Challenges and Possible Strategies'. *Language Documentation* 13: 545–63.

Leveraging Digital Library Infrastructure to Build a Language Archive

Mark Edward Phillips
Libraries
University of North Texas
Denton, Texas
mark.phillips@unt.edu
0000-0002-9679-6730

Mary Burke
College of Information
University of North Texas
Denton, Texas
mary.burke@unt.edu
0000-0002-6498-6820

Hannah Tarver
Libraries
University of North Texas
Denton, Texas
hannah.tarver@unt.edu
0000-0003-2344-9268

Oksana L. Zavalina
College of Information
University of North Texas
Denton, Texas
oksana.zavalina@unt.edu
0000-0002-3354-4923

***Abstract*—Building a digital language archive requires a number of steps to ensure collecting, describing, preserving, and providing access to language data in effective and efficient ways. The Computational Resource for South Asian Languages (CoRSAL) group has partnered with the University of North Texas (UNT) Digital Library to build a series of interconnected digital collections that leverage existing UNT technical and metadata infrastructure to provide access to data from and for various language communities. This article introduces the reader to the background of this project and discusses some of the important for representing language materials areas where UNT metadata has needed flexibility to better fit the needs of intended audiences. These areas include a workflow for standardized language representation (the Language field), defining roles for persons related to the item (Creator and Contributor fields), and representing interconnections between related items (the Relation field). Although further work is needed to improve language data representation in the CoRSAL digital language archive, we believe the model adopted by our team and lessons learned could benefit others in the language archiving community.**

***Index Terms*—metadata, language archives, digital libraries, controlled vocabularies**

I. Introduction

Over the past three years the collaborators at the University of North Texas (UNT) – the College of Information's Department of Linguistics and the Department of Information along with the Digital Libraries Division of the UNT Libraries – have worked to create the Computational Resource for South Asian Languages or CoRSAL. This program seeks to collect, describe, preserve, and provide access to language data and related artifacts from the South Asian region of the world. Initially, two collections from UNT Linguistics faculty (Lamkang Language Resource and Burushaski Language Resource) were uploaded; CoRSAL now accepts deposits from researchers and language community members. A key component of the CoRSAL program is a digital archive that has been built upon existing technical and metadata infrastructure in the UNT Libraries' Digital Collections. While creating the CoRSAL collection in the UNT Digital Library (one component interface of the Digital Collections), the project team has discovered information about metadata modeling and creation that we believe would be beneficial to the wider community. A selection of these lessons learned are presented below.

II. Background

The UNT Libraries' Digital Collections use a uniform locally-developed metadata scheme (UNTL) to describe items regardless of material type, owner, or collection. UNTL is based on the Dublin Core standard, with additional local fields for a total of 21 fields, 14 of which are locally-qualified. Over time, we have developed extensive guidelines (https://library.unt.edu/digital-projects-unit/metadata/input-guidelines-descriptive/) providing usage information and example data values for each of the fields across different material types. For some large, ongoing collections, we also create more specific metadata guidelines that clarify specifics of field usage, state which information applies from the general guidelines, and provide relevant examples.

As we built the CoRSAL collections, it became apparent that specialized metadata instructions would be useful. These were subsequently developed by CoRSAL staff, based on experiences describing the first two CoRSAL collections, with input from those who have archived language data in the past. Because CoRSAL prioritizes deposits from community language documenters, guidelines are intended to be readily understood by first-time metadata creators. Depositors are given a template with examples of completed metadata from other collections. The metadata guidelines development process took into account the relevant attributes of the data typical for language archive deposits: language(s), genre, roles of contributors and creators, and the relationship between items (e.g., between an audio and its transcript; original text and its translations). Though subject representation is not typically emphasized in language archive metadata [1], the CoRSAL metadata creation guide does encourage depositors to include keywords about the content or topic of the items. Finally, templated content descriptions are included to provide examples to depositors.

III. Language-Specific Metadata Usage

Currently there are twelve distinct CoRSAL collections in the UNT Digital Library. This integration process has provided structure to the wide range of language data that is being deposited as well as providing a process for unifying resource description across collections to improve discovery.

For this workshop paper we will focus on three primary areas: language representation, agent roles, and item relations.

A. Language Field

The UNTL metadata schema has a locally-developed controlled vocabulary for language codes (https://digital2.library.unt.edu/vocabularies/languages/), displayed as a drop-down list for editors. Designating new languages and codes has happened organically as material was added to the collection. Because the scope of content being collected and digitized was primarily focused on English-language resources, the language vocabulary grew slowly. Language codes were historically based on ISO 639-3 three letter codes and designated names. As the UNT Digital Library began adding CoRSAL collections, it became clear that this approach would not always work for language data, especially language documentation data. There were two challenges that came up with the existing approach to language codes and names. First, what happens when the language is not present in the ISO 639-3 language list, and second, what should happen when the "standard" language name assigned by the organization maintaining the standard is not preferred or accepted by the language community?

After discussion, a new process was developed and has been in place for the past year. First, administrators check the ISO 639-3 language code and add that version to the local vocabulary. If the ISO 639-3 language code is not present, the Glottolog (https://glottolog.org/) is used a source of language code. Glottolog describes itself as the "Comprehensive reference information for the world's languages, especially the lesser known languages." Languages in Glottolog have unique identifiers, called Glottocodes, which are added to the local vocabulary with the primary language name [2]. If a language code is not present in either source, the CoRSAL archive team will work with the depositing researcher to submit a request and provide supporting documentation to register the language with Glottolog.

As an example, CoRSAL archive added the Azamgarhi Language Resource collection, however, "Azamgarhi" did not yet have an established identifier in any standard language list. To avoid future confusion, instead of using a near match or the code for the larger group of languages (east2875), CoRSAL staff applied for a Glottocode for this variety (azam1235). This provided a standard code so that the materials could be ingested with a controlled form of the language representation.

In situations where the language community does not recognize the "standard" language name used in the ISO 639-3 documentation [3], the UNTL system has the flexibility to use more acceptable technology and document multiple versions of the name. For instance, the ISO 639-3 code *lus* is based on the language name 'Lushai' which is now called Mizo. While the UNTL language code is *lus* to match the ISO 639-3 code, the language name is represented as 'Mizo.'

B. Creator and Contributor Roles

In representing creators and contributors, the UNTL metadata scheme takes an agent-based approach (i.e., "who made this item") rather than a role-based approach (i.e., "who filled each of these roles in creating an item"). Each agent is assigned a primary role describing their specific contribution to create or steward the item, based on MARC Code List for Relators with some local additions (https://digital2.library.unt.edu/vocabularies/agent-qualifiers/). This makes sense given the wide array of material types and roles, but it means that an entity (person or organization) can only be listed once per record across the creator/contributor fields.

For materials where individuals have multiple roles, it may be challenging to determine which role is the "primary" way that they contributed to the item. For example, the same individual may have transcribed an audio recording and then translated the content into English. In this case, both the Transcriber and Translator roles are applicable. It is possible to represent both, because additional roles and clarifications can be added in an optional Info subfield of the Creator and Contributor fields that displays to users and is searchable, so no information is lost.

Also, there is not always a consensus on role terminology between the information professionals and depositors in the language archiving community. For instance, the term Analyst is defined by MARC Code List for Relators as "a person or organization that reviews, examines, and interprets data or information in a specific area." However, this term is commonly understood by documentary linguists as referring to a person or group that specifically provided linguistic analysis of language data. This difference in interpretation highlights the need for collaboration and development of common understanding of terminology, and possibly extensions to existing OLAC (Open Language Archives Community) controlled vocabularies.

C. Relationships between Items

The CoRSAL collections provide access to a wide range of linguistic data, represented in formats such as audio and video recordings; transcriptions; translations; photographs of cultural events, local flora and fauna; field notes; and collected publications and writings in a given language. Practice in the Digital Collections is to describe each discrete component piece as a separate object – allowing for clear and accurate description of creation information – however, the UNTL system has a robust process for describing relationships between resources, leveraging the Relation metadata field and the available qualifiers (https://digital2.library.unt.edu/vocabularies/relation-qualifiers/). This allows researchers to find specific types of items (e.g., only videos) as well as intellectually-related materials (e.g, a transcript, translation, etc.). Additionally, the UNT Digital Library interface provides features to draw attention to related resources with visual cues (see Figure 1 and Figure 2).

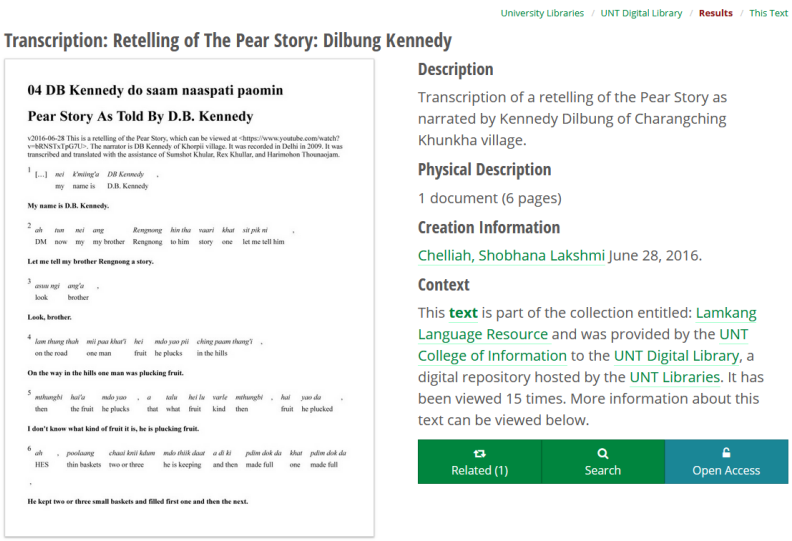


Fig. 1. Example item-level metadata with related items indicated.

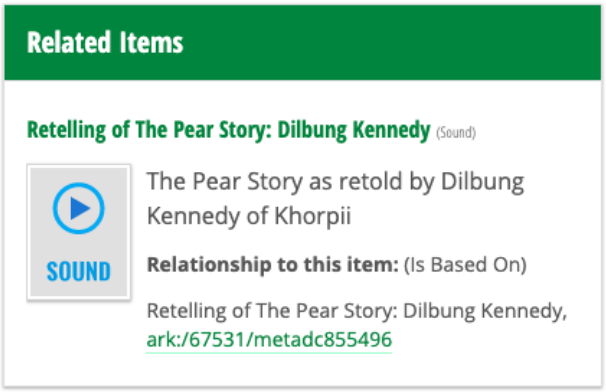


Fig. 2. Example representation of relationship from transcription to original recording.

IV. Conclusion

Although the UNTL metadata scheme is not always a perfect match for the CoRSAL digital language archive collections, since it is not specific to language-based data, it has been easily adapted to these kinds of materials in most cases. We have been able to develop new processes to address specialized concerns (e.g., those related to language names) and are engaging in continuing discussions regarding the best way to handle other issues to ensure robust description that meets the needs of both researchers and the wider, global internet audience.

With any metadata implementation, there is the need for user studies to determine the level of usability for the end-users and the areas of weakness to be addressed. A study focusing on the CoRSAL interface and metadata will help develop a robust understanding of the users' experience when interacting with the digital language archive, and get ideas for potential improvements to future metadata.

Overall, adding CoRSAL collections to the UNT Digital Library has provided a relatively easy way to make materials findable and available to other users while making use of the existing infrastructure and the UNTL metadata schema. While it does require some flexibility and logistical planning, this model and the general success in providing access to these materials show that a similar approach may allow more language researchers to make their materials available online for reuse.

References

[1] M. Burke, O. L. Zavalina, M. E. Phillips, and S. Chelliah, "Organization of knowledge and information in digital archives of language materials," *Journal of Library Metadata*, pp. 1–33, 2021. [Online]. Available: https://doi.org/10.1080/19386389.2020.1908651

[2] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank, "glottolog/glottolog: Glottolog database 4.4," May 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4761960

[3] S. Morey, M. W. Post, and V. A. Friedman, "The language codes of iso 639: A premature, ultimately unobtainable, and possibly damaging standardization," 2013-01-01. [Online]. Available: http://hdl.handle.net/2123/9838

Collaborating with Language Community Members to Enrich Ethnographic Descriptions in a Language Archive

Mary Burke†
Department of Information Science

University of North Texas
Denton TX, USA
Mary.Burke@unt.edu

ABSTRACT

Language archives connect users such as language communities, linguists, and other researchers, to language data. As the language archiving community develops, concerns have been raised about the ethics, ownership, accessibility, and context of archival materials. While there are no simple solutions to these questions, many language archives are seeking ways to involve language community members in these conversations as they continue. This presentation describes a pilot project undertaken at the Computational Resource for South Asian Languages (CoRSAL) which explores a collaborative archiving approach to enable language community members to tell their own stories by adding contextual information to archival materials.

KEYWORDS

ethnographic metadata, collaborative archiving, contextual metadata, unknown-provenance language materials, language community partnerships

1 Introduction

Language archives are cultural heritage institutions serving as repositories of primary language data: material about or in a set of languages (audio and video recordings, transcriptions, translations, and linguistic annotations). Archival materials are meant to serve as a lasting record of the language, and the starting point for further linguistic analysis or creation of pedagogical materials [1]. Language archives connect users such as language communities, linguists, and other researchers to language data. As the language archiving community develops, concerns have been raised about the ethics, ownership, accessibility, and context of archival materials [1, 2, 3]. While there are no simple solutions to these questions, many language archives are seeking ways to involve language community members in these conversations as they continue. This presentation describes a pilot project undertaken at the Computational Resource for South Asian Languages (CoRSAL) which explores a collaborative archiving approach to enable language community members to tell their own stories by adding contextual information to archival materials.

2 Related Work

Language documentation is the subfield of linguistics dedicated to preserving linguistic diversity. The process of language documentation is closely related to other subfields of linguistics, such as language description, language revitalization, language archiving, and to other disciplines, like information science, archival studies, anthropology, and ethnobiology. This section briefly reviews recent work concerning the relationship between language communities and language archives.

As the field of language documentation re-reorients to prioritize the needs of language communities, language archives too are seeking ways to be maximally useful to language communities as well as academic audiences. We see this incorporated into the design of language documentation projects themselves (see [4] on Community Based Language Research) and in the way we think about language archives. For example, [5] encourages depositors to consider the potential audiences of their archival collection, and to describe the material in a way that is appropriate for those audiences, taking into consideration factors like their primary languages and domain knowledge. More recently, [6] recommend for language documenters to discuss language communities' unmet needs during the documentation process, and work with language archives to make collections accessible despite the target community's specific barriers to access (e.g., minimal internet access).

Recent work in this area has noted the integral role that community engagement and rich contextual descriptions play in facilitating access to archival materials [7, 8, 9, 10]. Through discussions of the ethics of appropriating materials and framing community stories in non-community perspectives, many have called for increased involvement of language communities in the archiving process [11, 12]. However, linguists and their research team are often the ones responsible for managing the data, creating metadata, and depositing material into an archive. Language archive metadata records typically include the following elements: Identifier, Title, Contributor/ Depositor/ Creator, Language, Date, Description, Format, Notes, Rights, and Related items [13], with a high degree of variability in free-text descriptions noted by [14]. So, the

metadata that accompanies items in a language archive is based on the information that the research team might use to identify an item. This includes information like the names of those in the video, the date it was recorded, a genre, or the name of a story or song, but may exclude crucial cultural context, like that this song is only sung at a particular festival or by certain individuals. See, for example, [15] for a recent project where a language community representative was hired to identify gaps and errors in metadata in PARADISEC legacy material from Papua New Guinea.

3 Project Description

In light of these developments in language archiving, we saw an opportunity to test out a workflow which allows language community members to add in cultural context to already existing metadata. In the summer of 2020, two students at the University of North Texas (UNT) were hired to add cultural information to two collections in CoRSAL, briefly described here.

3.1 Burushaski Language Resource

Javid Iqbal, a Linguistics Masters student, is a Burushaski speaker. Before coming to UNT, he worked at the Burushaski Research Academy (BRA) as a research officer documenting cultural events and coordinating community meetings to raise awareness about the status of the language. He engaged with the Burushaski Language Resource, developed by Dr. Sadaf Munshi, which contains audio and video recordings of traditional, historical, and personal narratives, songs and poems, conversations, and recipes.

3.2 Lamkang Language Resource

Sumshot Khular, from the Lamkang community, is currently a PhD student in Environmental Studies, and earned her Masters in Linguistics from UNT in 2018. She has been supporting the Lamkang language for decades in numerous capacities (e.g., translating the Universal Declaration of Human Rights into Lamkang; organizing workshops, documenting natural speech and community events). Khular contributed metadata to the Lamkang Language Resource, which was developed over the course of the Lamkang project (2008-present) by Dr. Shobhana Chelliah, Sumshot Khular, Rex Khullar, Daniel Tholung, among other Lamkang community members. The Lamkang Language Resource contains digitized printed material on Lamkang culture, and primary audio and video recordings of traditional narratives, procedural narratives, semi-guided narratives (pear stories), conversations, and songs.

Both collections include a large proportion of photographs of cultural events, community members, and significant places or items (e.g., plants, utensils, churches). Because Khular and Iqbal have both experience in language documentation work and expertise in their respective cultures, they were uniquely positioned to add contextual information to the material in these collections. Though the items in these collections were already accompanied by metadata, they added metadata in the Description and Subjects fields with cultural significance in ways beyond the ability of the original collectors or current CoRSAL staff. After the pilot project, the metadata was reviewed and copy-edited by CoRSAL staff to ensure consistency. The following section summarizes the improvements that were made to the collections, including examples of the metadata records before and after the project.

4 Contributions Made

4.1 Cultural Context

The community consultants identified the salient aspects of items to highlight in the Description field, such as the significance of the colors or weaving pattern used in a shawl, or which occasions a garment might be worn on. Their additions are particularly invaluable for those items which were initially contributed with little or no metadata by community members. See for example, Figure 1, which compares the original metadata record and the record after the Lamkang community consultant updated it. The earlier version of the Description field, for example, states that this is a photo "illustrating Lamkang culture," but it is not clear what aspect of culture is intended, or what context the photo was taken in. The new description expands on the traditional clothing items worn by the young dancer, the materials used to make them, and the event where the photo was taken.

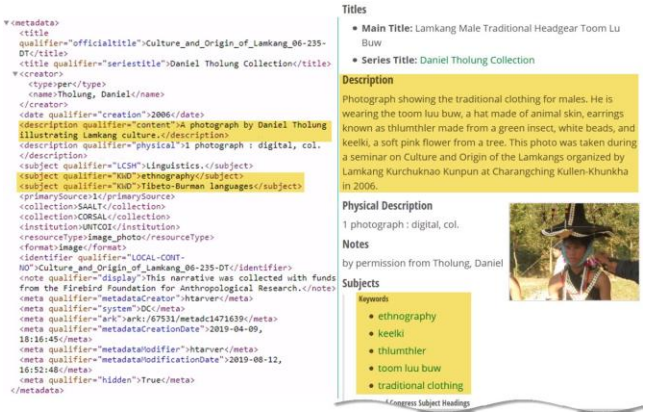


Figure 1: Example of a record before (left) and after (right) changes to the Keywords and Description fields

With this added information, the photograph is now connected to several others demonstrating the traditional clothing of the Lamkangs with the keyword 'traditional clothing'. Further, the updated description notes that these photos were taken during an event for celebrating and educating about Lamkang culture in 2006.

4.2 Target Language Metadata

Language community consultants also added information in the target language; for example, for the photograph of Lamkang children in traditional dress (Figure 1 above), the Lamkang community consultant explained that the dancer is wearing a *toom luu buw* (hat) and *thlumthler* (earrings). For the same photograph, another metadata creator would likely have used generic subject

terms (e.g., 'Ethnic costume' from the Library of Congress Subject Headings). Though not inaccurate, a term like 'Ethnic costume' may not be as helpful as a Lamkang term to users whose primary language is Lamkang.

In some cases, CoRSAL staff were able to identify the scientific name for the plants and animals featured in the photographs with the help of community consultants, such that the final record contains the Lamkang name of the animal, the English name, and the scientific name. Compare again the metadata records before and after the pilot project for this photograph of the antlers of a hog deer.

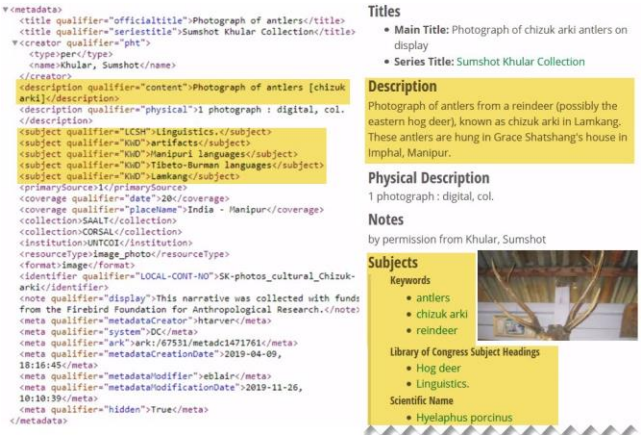


Figure 2: Example of a record before (left) and after (right) changes to the Keywords and Description fields

While the initial metadata included only 'antlers' and the Lamkang term, adding the English term for the animal allowed us to identify the scientific name. The resulting record contains useful information for Lamkang speakers, as well as users interested in the wildlife of Northeast India.

In the Burushaski collection, the community consultant added the names of the recipes and dishes featured in photographs and audio recordings. See for example, Figure 3, which compares the metadata record for the same photograph in March 2020 and July 2021.

Figure 3: Example of a record before (left) and after (right) changes to the Keywords and Description fields

Note how the Burushaski name of the dish is included in multiple varieties of Burushaski, while the previous version of the metadata record did not have a complete Content Description.

5 Summary and Future Plans

As a result of this pilot project, the metadata for the Lamkang and Burushaski collections is more accurate, complete and culturally relevant. The contextual information added by the community consultants will improve the experience of Lamkang or Burushaski speakers using these collections, as well as those interested in the respective cultures. Given this positive experience, we intend to replicate this process with future incoming collections whenever possible.

REFERENCES

[1] Henke, Ryan, & Berez-Kroeker, Andrea. (2016). A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation and Conservation, 10,* 411–457.

[2] O'Meara, Carolyn, & Guadarrama, O A. G. (2016). Accessibility to results and primary data of research on indigenous languages of Mexico. In Language Documentation and Revitalization in Latin American Contexts (pp. 59–80). De Gruyter Mouton. doi: 10.1515/9783110428902-003

[3] Wasson, C., Holton, G., & Roth, H. (2016). Bringing user centered design to the field of language archives. *Language Documentation and Conservation, 10,* 641–671.

[4] Czaykowska-Higgins, Ewa. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. *Language Documentation & Conservation 3*,15-50.

[5] Woodbury, Anthony. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description, 12,* 19-36.

[6] Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, & Anthony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation, 13,* 545-563.

[7] Greyling, Elizabeth, & Zulu, Sipho (2010). Content development in an indigenous digital library: A case study in community participation. *International Federation of Library Associations and Institutions*, *36*, 1, 30–39. doi: 10.1177/0340035209359570

[8] Roeschley, Ana, Kim, Jeonghyun, & Zavalina, Oksana L. (2020). An exploration of contributor-created Description fields in participatory archives. In Sundqvist A., Berget G., Nolin J., Skjerdingstad K. (Eds), Sustainable Digital Communities. iConference 2020. Lecture Notes in Computer Science, 12051 (pp. 638-648). Springer, Cham. https://doi.org/10.1007/978-3-030-43687-2_54

[9] Sullivant, Ryan. (2020). Archival description for language documentation collections. *Language Documentation & Conservation, 14,* 520-278.

[10] Burke, Mary, Zavalina, Oksana L., Chelliah, Shobhana, & Phillips, Mark (in press). User Needs in Language Archives: Findings from Interviews with Language Archive Managers, Depositors, and End-Users. *Language Documentation and Conservation.*

[11] Ormond-Parker, Lyndon, Corn, Aaron, Fforde, Cressida, Obata, Kazuko, & O'Sullivan, Sandy. (2013). (Eds.) Information Technology and Indigenous Communities. Canberra: AIATSIS Research Publications.

[12] Shepard, Michael A. (2016). The Value-Added Language Archive: Increasing Cultural Compatibility for Native American Communities. *Language Documentation & Conservation, 10,* 458-479.

[13] Burke, Mary, & Zavalina, Oksana L. (2019). Exploration of information organization in language archives. *Proceedings of the American Society for Information Science and Technology, 56,* 1.

[14] Burke, Mary, & Zavalina, Oksana L. (2020). Descriptive richness of free-text metadata: a comparative analysis of three language archives. *Proceedings of the American Society for Information Science and Technology, 57,* 1.

[15] Harris, Amanda, Gagau, Steven, Kell, Jodie, Thieberger, Nick, & Ward, Nick. (2019). Making meaning of historical Papua New Guinea recordings:

Collaborations of speaker communities and the archive. *International Journal of Digital Curation 14*, 1, 136-149.

Challenges in heritage language documentation: BraPoRus, spoken corpus of heritage Russian in Brazil

Anna Smirnova Henriques
LAEL, Pontifícia Universidade Católica de São Paulo/PUC-SP
São Paulo, Brazil
anna.smirnova.liaac@gmail.com

Aleksandra S. Skorobogatova
FFLCH, Universidade de São Paulo/USP
São Paulo, Brazil
as.skorobogatova@gmail.com

Svetlana Ruseishvili
Universidade de São Carlos/ UFSCar
São Carlos, Brazil
s.ruseishvili@gmail.com

Sandra Madureira
LAEL, Pontifícia Universidade Católica/PUC-SP
São Paulo, Brazil
sandra.madureira.liaac@gmail.com

Irina A. Sekerina
College of Staten Island, The City University of New York/CUNY
New York, USA
irina.sekerina@csi.cuny.edu

***Abstract*—The Bolshevik revolution in 1917, followed by the Civil War, induced a big wave of emigration from the ex-Russian Empire. These emigrants created their "Russia Abroad". Many Russians stayed in Europe or China, but, in the 1940s and 1950s, many of them went to the USA, Latin America and other destinations. The importance of preserving the memories and documents of the old waves of the Russian emigration is crucial. Our group is collecting a corpus of heritage Russian in Brazil, the BRAzilian POrtuguese RUSsian Corpus (BraPoRus). While the history of Russian immigration in Brazil is to some extent studied, their remarkably preserved Russian has not been described. Our current aim is to describe the BraPoRus, a corpus that consists of multiple speech samples of older Russian heritage speakers in Brazil, and to discuss the best ways to make these data available in the forms that satisfy the requirements both for the linguistic and sociological research.**

Keywords—Russian as a heritage language; heritage speakers; language archives; oral history; bilingualism

I. INTRODUCTION

The Bolshevik revolution in 1917, followed by the Civil War, resulted a large emigration wave from the ex-Russian Empire: approximately 1.5 to 2 million Russians (this term frequently includes the Russophones or people who belong to "culturally russified communities") left the country [1]. In their new countries, they founded churches, schools, local journals, and artistic groups, creating their "Russia Abroad". Many Russians stayed in Europe [2], but the World War II forced them to emigrate as refugees to the USA, Latin America, and other destinations.

A special group of Russians is the "Russians from China": most of them come from the village of Harbin which at the end of the 19th century began to receive Russian engineers in charge of the construction of the Chinese Eastern Railway, and later, after the Bolshevik Revolution and the arrival of many Russian immigrants, became an important destination for the Russian emigré community [3]. The Chinese Communist Revolution in 1949 brought important changes in the immigration politics, and the majority of Russians left China in the 1950s, either moving back to the USSR or to the USA, Australia, and Brazil. Preserving the memories and documents of the old waves of the Russian emigration is crucial, but many archives were lost [1, 4]. It is also important to document and preserve speech samples of the Russian language spoken by these emigrants and their descendants. The digital collection of the Columbia University Libraries contains a number of oral history interviews with Russian immigrants, recorded in the 1960s [5], but this is a unique collection of such materials on this issue.

Brazil accepted the Russian refugees from the former Russian Empire between 1921 and 1941 [6]. In the post-World War II period, many Soviet Displaced Persons (DPs) and the families of the "white" Russian community in Europe also arrived to Brazil. In the 1950s, it was the turn of the Russian "white" stateless refugees from China [7]. Rough estimate based on the Brazilian Institute of Geography and Statistics 1950 census [8] is that 1,500 of these Russophone immigrants from China could be still alive. While the history of Russian immigration in Brazil is studied to some extent, their remarkably preserved Russian has not been described. Our group is collecting a corpus of moribund heritage Russian in Brazil, the BRAzilian POrtuguese RUSsian Corpus (BraPoRus). D'Alessandro and colleagues [9] define a heritage language as moribund when it is spoken by elderly people who are a final generation of proficient speakers whose production and comprehension of heritage language are preserved; when they die, the language dies with them. Our current aim is to present the BraPoRus, a corpus that consists of 160 hours of speech samples of elderly Russian heritage speakers in Brazil (M_{age} = 77 years), and to discuss the best ways to make these data available in the forms that satisfy the requirements for the linguistic and sociological research in heritage language documentation.

II. PARTICIPANTS

The participants were selected according to the following criteria: 1) age 59 years and older (range: 59-98); 2) living in Brazil for most part of their life or being born in Brazil, speaking Portuguese in a nativelike way; 3) proficiency in Russian as a heritage language, sufficient to maintain a conversation for an hour; 4) no long-term residence in Russia; 5) no cognitive impairment. Currently, 31 participants (12 men and 19 women) are enrolled in the study.

III. METHOD

The protocol for data collection includes: 1) a brief demographic questionnaire; 2) a working memory test in Russian and Brazilian Portuguese using the Month-Ordering task [10]; 3) a semi-spontaneous narrative about the history of the participants' family and their immigration to Brazil; 4) the Bilingual Language Profile [11]; 5) a sociolinguistic interview with 139 questions adapted from the long HLVC (Heritage Language Variation and Change, Toronto) questionnaire [12]; 6) an assessment of narrative abilities in Russian and Brazilian Portuguese using Multilingual Assessment Instrument for Narratives (MAIN) [13]; 7) unscripted dialogues between participants in Russian; 8) intonation tasks; and 9) reading tasks. The data are being collected in 6-8 online sessions, through phone calls or videoconference by Zoom. All the speech samples recorded at the steps 3-7 will be transcribed and annotated using ELAN.

IV. RESULTS

The sociodemographic data that describe profiles of the BraPoRus participants are presented below.

TABLE I. SOCIODEMOGRAPHIC DATA

Age		
	Mean	77.3 years (SD = 8.7)
	Range	59-98
Residence		
	São Paulo city	24
	São Paulo state	2
	Rio de Janeiro	4
	Curitiba	1
Place of Birth		
	China (Harbin)	13 (9)
	Brazil	11
	Europe	4
	Russia	2
	Belarus	1
Age of Arrival to Brazil (20 out of 31 participants)		
	Median	10.8
	Range	1-24

Only the two oldest participants (97 and 98) arrived in Brazil over the age of 18 (20 and 24, respectively), both born in Europe.

We have recorded the family history stories and, partially, sociolinguistic interviews in Russian from 21 participants. The total duration of these recordings currently is 160 hours. In addition, we have recorded 5.8 hours of speech in dialogue interactions.

V. DISCUSSION

In this project, we have recorded 160 hours of Russian speech samples produced by elderly Russian heritage speakers that reside in Brazil, mainly descendants of the emigrants that left the Russian Empire directly after the Bolshevik revolution. We plan to describe and characterize the bilingual speech of the elderly heritage Russian speakers. These will include attrition effects from Brazilian Portuguese, code-switching, intonation profiles, and interaction between the working memory and narrative abilities.

The sociolinguistic interviews collected in this study contain unique information about the history of Russian emigration, and the immigration experience after the Bolshevik revolution and World War II. Our questionnaires address the family history, the immigration paths, the places of residence and description of the houses and daily routine in Russophone communities of Europe, China and Brazil, the adaptation difficulties of Russophone families in Brazil, the life of the Russophone communities in Brazil, including education and cultural events, the religious traditions (some of the participants are from Old Believers families), food habits, and many other issues. In addition, some participants provided artifacts in the form of documents and old photos kept by the family.

As far as ethics is concerned, the common way of addressing ethical concerns in speech corpus is anonymization in ELAN and submission to an open database, such as the TalkBank [14]. The TalkBank Code of Ethics establishes that only age and location of the recordings could be annotated, but no personal data about individual participants. Many interviews contain sensitive data and should be anonymized; following the gold standards of the speech corpus construction, all the names and places mentioned in the recordings should be replaced with silence. From the other side, the names and places are necessary for the historical and sociological analysis, and the collections of oral history interviews, as in [5], contain these types of data.

VI. CONCLUSION

The rich data obtained in our project, which focuses on the study of the elderly Russian heritage speakers in Brazil, can be framed in two ways: as an annotated and anonymized speech corpus, and as a database of oral history interviews. This dualism raises many questions. Does it make sense to make the same interviews available in two forms, one as an anonymized corpus for linguistic research proposes and, at least with restricted access, as oral history interview database for use in history and sociology studies? If so, how could it be done? How to balance the participant anonymization and the preservation of the memories about the immigration history? How to guarantee that the linguistic data in specialized databases could be efficiently accessed by historians and sociologists in a useful format? The answers to these questions are important in order to improve the documentation of heritage languages and relate them to the historical and sociological studies of immigration.

ACKNOWLEDGMENT

Dr. Smirnova Henriques is supported by postdoctoral fellowship PNPD/CAPES (Programa Nacional de Pós-Doutorado da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

REFERENCES

[1] M. Raeff, "Recent Perspectives on the History of the Russian Emigration (1920-1940)," Kritika: Explorations in Russian and Eurasian History, vol. 6 (2), pp. 319-334, 2005. DOI: https://doi.org/10.1353/kri.2005.0025.

[2] A. J. Cohen, "Our Russian Passport': First World War Monuments, Transnational Commemoration, and the Russian Emigration in Europe, 1918–39," Journal of Contemporary History, vol. 49 (4), pp. 627–651, 2014. DOI: https://doi.org/10.1177/0022009414538469.

[3] O. M. Bakich, "Emigre Identity: The Case of Harbin," The South Atlantic Quarterly, vol. 99 (1), pp. 51-73, 2000. Available at: https://muse.jhu.edu/article/30644/. Accessed on: 05 September 2021.

[4] N. Saul, "American Collections on Immigrants and Émigrés from the Russian Empire," Slavic & East European Information Resources, vol. 4 (4), pp. 49–61, 2003. DOI: https://doi.org/10.1300/J167v04n04_05.

[5] Columbia University Libraries, Digital Collections, Radio Liberty, 2021. Available at: https://dlc.library.columbia.edu/catalog?utf8=%E2%9C%93&search_field=all_text_teim&q=Radio%20Liberty. Accessed on: 05 September 2021.

[6] A. Bytsenko, Imigração da Rússia para o Brasil no início do século XX. Visões do paraíso e do inferno. São Paulo: Universidade de São Paulo, 2006. Available at: https://www.teses.usp.br/teses/disponiveis/8/8155/tde-12112007-132926/pt-br.php. Accessed on: 05 September 2021.

[7] S. Ruseishvili, Ser russo em São Paulo. Os imigrantes russos e a reformulação de identidade após a Revolução Bolchevique de 1917. São Paulo: Universidade de São Paulo, 2016. Available at: https://teses.usp.br/teses/disponiveis/8/8132/tde-13022017-124015/pt-br.php. Accessed on: 05 September 2021.

[8] IBGE (Instituto Brasileiro de Georgrafia e Estatística), Censo demográfico, 1956. Available at: https://biblioteca.ibge.gov.br/visualizacao/periodicos/67/cd_1950_v1_br.pdf. Accessed on: 05 September 2021.

[9] R. D'Alessandro, D. Natvig, M. T. Putnam, "Addressing challenges in formal research on moribund heritage languages: A path forward," Frontiers in Psychology, vol. 12, art. 700126, 2021. DOI: https://doi.org/10.3389/fpsyg.2021.700126.

[10] D. Kempler, A. Almor, L. K. Tyler, E. S. Andersen, M. C. Macdonald, "Sentence comprehension deficits in Alzheimer's disease: a comparison of off-line vs. on-line sentence processing," Brain and Language, vol. 64 (3), pp. 297-316, 1998. DOI: https://doi.org/10.1006/brln.1998.1980.

[11] D. Birdsong, L. M. Gertken, M. Amengual, Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism, COERLL, University of Texas at Austin. Available at: https://sites.la.utexas.edu/bilingual/. Accessed on: 05 September 2021.

[12] N. Nagy, "Heritage languages as new dialects," In: M. Côté, R. Knooihuizen, J. Nerbonne (Eds.), The future of dialects. Berlin: Language Science Press, 2016. DOI: https://doi.org/10.17169/langsci.b81.81.

[13] N. Gagarina, D. Klop, I. M. Tsimpli, J. Walters, "Narrative abilities in bilingual children," Applied Psycholinguistics, vol. 37 (1), pp. 11-17, 2016. DOI: https://doi.org/10.1017/S0142716415000399.

[14] B. MacWhinney, The TalkBank system, 2021. Available at: https://www.talkbank.org/. Accessed on: 05 September 2021.

Towards an Agenda for Open Language Archiving

Steven Bird
Charles Darwin University
Australia
steven.bird@cdu.edu.au

Gary F. Simons
SIL International
United States of America
gary_simons@sil.org

***Abstract*—The Open Language Archives Community (OLAC) provides a comprehensive infrastructure that has allowed our community to index and discover language resources over the past 20 years. However, OLAC infrastructure has fallen behind as the digital libraries community has continued to evolve. New investment is required in order to move OLAC into the digital libraries mainstream. This paper reports on the first 20 years of OLAC and on an agenda leading to a more sustainable future for open language archiving.**

Keywords—Russian as a heritage language; heritage speakers; language archives; oral history; bilingualism

I. INTRODUCTION

OLAC was founded in 2000 as an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by developing consensus on best current practice for the digital archiving of language resources, and developing a network of interoperating repositories and services for housing and accessing such resources. We take a language resource to be "any physical or digital item that is a product of language documentation, description, or development or is a tool that specifically supports the creation and use of such products" [29, p88].

OLAC infrastructure is built on Dublin Core metadata [14] and the Open Archives Initiative Protocol for Metadata Harvesting [17]. At the time of writing, OLAC catalogues over 440,000 items from 62 participating language archives (http://www.language-archives.org/archives). These items cover all of the living languages recognised by the ISO 639-3 standard (see Fig. 1; http://www.language-archives.org/documents/coverage.html). For the most recent month, we logged 8,600 record views on the OLAC site, with 2,172 click-throughs to individual archives (does not include traffic to the search service hosted at the University of Pennsylvania).

Users access the OLAC catalog in a variety of ways: via any search engine, since OLAC exposes everything as pages that Web crawlers can index; via faceted search which exploits the controlled vocabularies to give search with complete recall and precision (http://search.language-archives.org); via links from language-related sites like Ethnologue (https://ethnologue.com/language/aaa: see link to "Language Resources"); via services such as WorldCat, CLARIN, Linguist List which harvest OLAC metadata from the OLAC Aggregator (http://www.language-archives.org/cgi-bin/olaca3.pl); by consuming the XML or RDF/XML nightly dumps of the entire OLAC metadata catalog (http://www.language-archives.org/xmldump/ListRecords.xml.gz; http://www.language-archives.org/static/olac-datahub.rdf.gz); or by accessing the RDF/XML of any metadata record via HTTP content negotiation (http://www.language-archives.org/item/oai:paradisec.org.au:AA1-001)

Alongside this technical infrastructure, OLAC has a document infrastructure: defining OLAC metadata standards [23]; specifying processes around repositories [24]; and laying out the process for managing the document lifecycle through a Council and Board [25].

This paper reviews the first 20 years of OLAC and identifies new opportunities to support long-term growth and viability of open language archiving.

TABLE. 1. COVERAGE OF OLAC ITEMS INDEXED BY ISO 639-3 IN RELATION TO LANGUAGE SIZE

Population range	Lan-guages	OLAC has data	(%)	Items
1-9	133	133	100	3,563
10-99	339	339	100	13,372
100-999	1,038	1,038	100	29,605
1,000-9,999	2,014	2,014	100	62,791
10,000-99,999	1,824	1,824	100	46,813
100,000-999,999	895	895	100	29,235
1,000,000-9,999,999	304	304	100	14,892
10,000,000-99,999,999	77	77	100	49,008
100,000,000-999,999,999	8	8	100	47,233
Unknown	277	277	100	7,909
All living languages	6,909	6,909	100	304,421
Extinct languages	626	599	96	7,247

II. OLAC VISION FOR THE OPEN LANGUAGE ARCHIVING

The original vision for OLAC was set out in a document entitled The Seven Pillars of Open Language Archiving [22]. According to this vision, the individuals who use and create language documentation and description are looking for three things: Data, information that documents or describes a language of interest; Tools, computational resources that facilitate creating or using language data; and Advice, help in knowing what data sources to rely on, what tools to use, and what practices to follow. Despite this need, potential users of language resources did not have ready access to the data, tools, and advice that they needed. We explored these shortcomings through a "gap analysis", as follows: some archives (e.g. Archive 1, in Fig. 1A) have a site on the Internet which the user is able to find, so the resources of that archive are accessible; other archives (e.g. Archive 2) are on the Internet, so the user could access them in theory, but the user has no idea they exist so they are inaccessible in practice; still other archives (e.g.

Archive 3) are not even on the Internet. There are potentially hundreds of archives (Archive n) that the user should know about. Finally, tools and advice reside in many places, and are not indexed in a way that allows users to discover them, or relate the available tools to the available data.

OLAC was established in order to address these issues. According to the vision, OLAC would do this by offering four things: Gateway, a single portal through which users can access all available data, tools, and advice; Metadata, uniform descriptions of all available data, tools, and advice; Reviews, peer evaluations of available data, tools, and advice; and Standards, processes and protocols that enable the operation of the gateway and ensure the quality of metadata and reviews. We then articulated an overall solution having the structure shown in Fig. 1B.

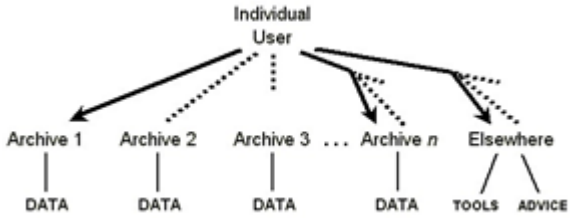


FIG. 1A. The Vision of the Open Language Archives Community: Gap Analysis for People Attempting to Access Language Archives

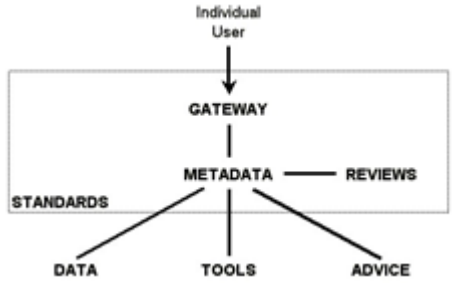


FIG. 1B. The Vision of the Open Language Archives Community: The Seven Pillars of Open Language Archiving

We have fleshed out this vision in further detail: requirements, for users, creators, archivists, developers and sponsors [21]; a survey of the state of the art in digital language documentation and description [3]; a later analysis with best practice recommendations [6]; and a white paper on establishing the infrastructure for open language archiving on the framework of Dublin Core Metadata and the Open Archives Initiative [4], subsequently implemented and reported in a series of publications [5, 19, 26, 27].

III. Taking Stock of OLAC Today

The present state of OLAC can be summarised as follows. The OLAC document process [25] has been established and used by the community in a series of workshops over several years to create many OLAC documents (http://www.language-archives.org/documents.html). The repositories and metadata standards have functioned continuously for 20 years [23, 24]. The community has used the OLAC Process to develop and refine vocabularies for linguistic data type [2]; discourse type [16]; contributor roles [15]; and linguistic field [1]. The most significant of these vocabularies is linguistic data type, though it only has three items corresponding to the Boasian trilogy [7]: lexicon, primary text, and language description. We specified the OLAC Language Extension [28] which standardised OLAC metadata to use ISO 639-3 codes for representing the names of languages. We developed a MARC to OLAC crosswalk [12]. We compiled best practice recommendations for the use of OLAC Metadata [31]. We established guidelines for metadata quality and provided automatic evaluation of quality and a quarterly report emailed to the repository coordinator, in order to motivate effort to improve metadata quality [18], an area where OLAC has been considered exemplary [11]. We have articulated sustainability conditions for language resources [29], chiefly, the conditions that ensure a resource will be usable—it is discoverable, available, interpretable, and portable. We have established and continue to maintain core infrastructure hosted at the Linguistic Data Consortium, and a search service hosted at the University of Pennsylvania Library (http://search.language-archives.org). Library.

Alongside these contributions of OLAC is the response from the community, including over 5,000 publications that cite OLAC (https://scholar.google.com/scholar?q="OLAC"+language). There is evidence that OLAC is enabling research that accesses language resources (e.g. [13]), and that OLAC is supporting ongoing scholarship on language archiving itself (e.g. [8, 9]).

Aside from these successes, there are various ways in which OLAC has not yet achieved our aspirations for long-term sustainability: the OLAC Council and Board have fallen inactive; the software infrastructure has not been refreshed in over a decade, and it is being maintained by volunteers and could fail catastrophically at any time (the website and search functionality would still operate, but new content coming from participating archives would not be harvested); of the 62 registered archives, 27 have not been updated in the past five years, and an overlapping 19 archives are failing to harvest. Also, the original vision for OLAC identified potentials which have not yet been realised: the indexing of tools and advice (cf. [34]), and using a formal document process in defining best practices in language archiving beyond resource description and discovery (cf. [6, 32]). More fundamentally, OLAC has not had the resources to keep up with current best practices of the digital libraries community. Funding has always been project-based. Advice from program managers has been that we add a research piece and compete for research funding, or that we objectively quantify the value of OLAC and seek infrastructure funding.

Since the founding of OLAC, the space for defining best practices in language archiving more broadly has been filled by the establishment of DELAMAN—the Digital Endangered Languages and Musics Archives Network (https://www.delaman.org/). We have initiated a process that is bringing OLAC under the governance structure of DELAMAN, with a narrowed scope of "developing consensus on best current practice for the interoperable description of archived language resources."

IV. Towards an Agenda for Open Language Archiving

Much remains to be done across the space of language archiving [33]. In considering the opportunities offered by OLAC in particular, we begin with what OLAC already offers: a community that has grown up around the participating archives; a suite of documents that define OLAC's operation; a process for updating these documents; an archive registration process; an aggregation infrastructure; a federated search service; a focus on documenting subject language and linguistic data type in language resource metadata; and automated encouragement for archives to improve metadata quality.

In looking to the future, we envisage improvements in coverage. There are significant collections not yet participating, both archives and special collections within libraries. However, it is evident that implementing a data provider for OLAC metadata is too high a bar for some organisations. Some archives only expose an index page per language, and instead need to expose metadata for the individual resources so that they can be indexed centrally. Finally, scholars need to be able to report language resources they discover in places that would never join OLAC (such as isolated texts in endangered languages).

We also envisage improvements in access. Many archives need to improve metadata quality so as to improve the discoverability of their holdings. At the time of writing, 22 out of 62 archives score below 70% on OLAC's metadata quality metric. Subject language is only used in 65% of records. Linguistic Data Type is used in a mere 21% of records. In addition, sub-communities could make OLAC more directly relevant for themselves, by cataloguing holdings according to their own system, e.g.: <dc:type>Sociolinguistic corpus</dc:type>, <dc:format>text/x-eaf+xml</dc:format>.

Finally, we envisage mainstreaming language archives, by replacing our parochial metadata format with a generic application profile, effectively steering OLAC and the cataloging of language resources into the library and information systems mainstream. Observing the trend in library automation toward Linked Data in cataloging, we have taken a first step by mapping OLAC metadata for Linked Data [30]. We envision OLAC's idiosyncratic metadata format being superseded by an application profile [10] for describing language resources. This would be anchored by a Language Resource Type vocabulary, enlarged from Linguistic Data Type to encompass the full range of resources held by language archives [20]. In this way, we hope to shift from an idiosyncratic community-specific infrastructure to a mainstream infrastructure that interoperates with the global Web of Data. At the same time, we would hope to influence mainstream cataloging practices to embrace the Language Resource Type vocabulary, along with ISO 639-3 for greater precision in language identification, so that their catalog records would conform to the application profile for language resources.

Acknowledgment

We acknowledge the financial support of the US National Science Foundation during the early days of OLAC. We are grateful to Mark Liberman and Christopher Cieri at the Linguistic Data Consortium, and to Martha Brogan and Lauris Olson at the University of Pennyslvania Library for their support in hosting OLAC infrastructure. We are indebted to Helen Dry and Anthony Aristar of the Linguist List, to participants of several OLAC and EMELD workshops, and to OLAC's advisory board and council, for many discussions about the design and operation of OLAC. Special thanks to Haejoong Lee for maintaining OLAC software over many years.

References

[1] Helen Aristar Dry and Michael Appleby. 2003. OLAC Linguistic Subject Vocabulary. http://www.language-archives.org/REC/field.html.

[2] Helen Aristar Dry and Heidi Johnson. 2002. OLAC Linguistic Data Type Vocabulary. http://www.language-archives.org/REC/type.html.

[3] Steven Bird and Gary Simons. 2000. A Survey of the State of the Art in Digital Language Documentation and Description. http://www.languagearchives.org/docs/survey.html.

[4] Steven Bird and Gary Simons. 2000. White Paper on Establishing an Infrastructure for Open Language Archiving. http://www.languagearchives.org/docs/white-paper.html.

[5] Steven Bird and Gary Simons. 2003. Extending Dublin Core metadata to Support the description and discovery of language resources. Computers and the Humanities 37 (2003), 375–388. http://arxiv.org/abs/cs.CL/0308022.

[6] Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79 (2003), 557–82.

[7] Franz Boas (Ed.). 1911. Handbook of American Indian languages. Smithsonian Institution Bureau of American Ethnology Bulletin, Vol. 40. Washington: Government Printing Office.

[8] Mary Burke and Oksana Zavalina. 2019. Exploration of information organization in language archives. Proceedings of the Association for Information Science and Technology 56 (2019), 364–367.

[9] Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Discovery of language resources. In Linguistic Linked Data. Springer, 263–279.

[10] Karen Coyle and Tom Baker. 2009. Guidelines for Dublin Core application profiles. https://www.dublincore.org/specifications/dublin-core/profileguidelines/

[11] Diane Hillmann. 2008. Metadata quality: From evaluation to augmentation. Cataloging and Classification Quarterly 46 (2008), 65–80.

[12] Christopher Hirt, Gary Simons, and Joan Spanne. 2009. Building a MARC-to-OLAC crosswalk: repurposing library catalog data for the language resources community. In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. ACM, 393–394.

[13] Russell Hugo. 2015. Constructing online language learning content archives for under-resourced language communities. Technical Report. University of Washington.

[14] Renato Iannella and Rachel Heery. 1999. Dublin Core Metadata Initiative – Structure and Operation. http://dublincore.org/documents/dcmi-structure/

[15] Heidi Johnson. 2003. OLAC Role Vocabulary. http://www.language-archives.org/REC/role.html.

[16] Heidi Johnson and Helen Aristar Dry. 2002. OLAC Discourse Type Vocabulary. http://www.language-archives.org/REC/discourse.html.

[17] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[18] Gary Simons. 2009. OLAC Metadata Quality. http://www.language-archives.org/NOTE/metrics.html .

[19] Gary Simons. 2014. The role of metadata in the infrastructure for archival interoperation. Language and Linguistics Compass 8 (2014), 486–494.

[20] Gary Simons. 2016. From Linguistic Data Type to Language Resource Type: Laying the groundwork for a metadata application profile.https:

//scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/simons-language_resource_type_vocabulary.pdf.

[21] Gary Simons and Steven Bird. 2000. Requirements on the Infrastructure for Open Language Archiving. http://www.language-archives.org/docs/requirements.html.

[22] Gary Simons and Steven Bird. 2000. The Seven Pillars of Open Language Archiving: A Vision Statement. http://www.language-archives.org/docs/ vision.html.

[23] Gary Simons and Steven Bird. 2001. OLAC Metadata. http://www.language-archives.org/OLAC/metadata.html

[24] Gary Simons and Steven Bird. 2001. OLAC Repositories. http://www.language-archives.org/OLAC/repositories.html

[25] Gary Simons and Steven Bird. 2002. OLAC Process. http://www.language-archives.org/OLAC/process.html

[26] Gary Simons and Steven Bird. 2003. Building an Open Language Archives Community on the OAI Foundation. Library Hi Tech 21 (2003), 210–218. http://www.arxiv.org/abs/cs.CL/0302021

[27] Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. Literary and Linguistic Computing 18 (2003), 117–128

[28] Gary Simons and Steven Bird. 2008. OLAC Linguistic Data Type Vocabulary. http://www.language-archives.org/REC/type.html

[29] Gary Simons and Steven Bird. 2008. Toward a global infrastructure for the sustainability of language resources. In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation. De La Salle University, Manila, Philippines, 87–100.

[30] Gary Simons and Steven Bird. 2020. Expressing language resource metadata as Linked Data: The case of the Open Language Archives Community. In Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences, Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos (Eds.). MIT Press, 117–130....

[31] Gary Simons, Steven Bird, and Joan Spanne. 2008. Best Practice Recommendations for Language Resource Description. http://www.languagearchives.org/REC/bpr.html.

[32] Nick Thieberger. 2012. Using language documentation data in a broader context. In Potentials of Language Documentation: Methods, Analyses, and Utilization, Frank Seifart, Geoffrey Haig, Nikolaus Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (Eds.). University of Hawai'i Press, 129–34.

[33] Nick Thieberger. 2017. What remains to be done: Exposing invisible collections in the other 7,000 languages and why it is a DH enterprise. Digital Scholarship in the Humanities 32 (2017), 423–434.

[34] Hans Uszkoreit, Brigitte Jörg, and Gregor Erbach. 2003. An Ontology-based Knowledge Portal for Language Technology. In Proceedings of ENABLER/ELSNET Workshop "International Roadmap for Language Resources". ELRA.

Linguistic Repositories as Asset: Challenges for Sociolinguistic Approach in Brazil

Raquel Meister Ko. Freitag
Vernacular Languages Department
Federal University of Sergipe
São Cristóvão, Sergipe, Brazil
rkofreitag@academico.ufs.br

***Abstract*—This paper provides remarks for a management plan for Brazilian linguistic documentation repositories in order to contribute to their conservation. The depreciation, authorship, sharing, and financing problems are discussed, pointing solutions.**

***Index Terms*—Linguistic repositories management, Brazilian Portuguese, linguistic repository**

I. Introduction

The sociolinguistic approach is characterized by the use of "databases" of authentic speech samples collected through interviews with speakers of a given speech community. The collection of sociolinguistic interviews constitutes a repository of linguistic documentation because the database conception presupposes automated search in a system for storage and organization [1]. Authentic and aligned transcribed linguistic data is an expensive product of interest to those working in data mining, machine learning, and artificial intelligence. In Brazil, this product is a result of larger sociolinguistic projects, with the objectives of:

- Providing resources for the description of Brazilian Portuguese
- Developing and testing linguistic theories
- Training new researchers
- Providing resources for educational programs

This is the case, for example, of NURC [2, 3, 4], PEUL [5, 6], and VARSUL [7, 8]. The intangible assets of the Brazilian sociolinguistic projects were constituted by actions 1 and 4, which constitute the tangible assets. In accounting terms, the difference between intangible and tangible assets is related to depreciation. Tangible assets are those that physically exist; in the case of sociolinguistic projects, their repositories of linguistic documentation are supported. The conservation of these repositories needs to overcome some challenges (depreciation, authorship, sharing, and financing), which are the objective of the discussion in this paper.

II. Challenges

A. Depreciation

In accounting, a tangible asset is depreciable, which means that it loses value over time. While the content of the linguistic corpus is stable, the support for this content is subject to depreciation. In order to beat depreciation, further resources are required. Transferring the audio collection stored on magnetic tapes to digital media, for example, is a procedure that prevents obsolescence (achievement of NURC-Recife [3]), since magnetic tapes have an expiration date and today's devices for this type of media are outdated. Even in digital repositories, routine backup procedures are requested, either in local physical storage media or in cloud storage. There are operational costs involved, both with the storage service and with specialized human resources to carry out this procedure.

B. Authorship

In Brazil, authorship and copyright are regulated by federal law 9610/1998. From a legal point of view, the repository of a sociolinguistic project is assigned as an intellectual property, with copyrights. Thus, the repositories of linguistic documentation from Brazilian sociolinguistic projects are a result of collective construction [9].

From the academic point of view, authorship and contribution are different: a researcher may have contributed to the data collection but may not be considered an author of it [10, 11]. One way of recognizing the types of contribution in science is presented by CRediT taxonomy (Contributor Roles Taxonomy), which names 14 roles that can be assigned to those who contribute to the construction of a scientific product, such as repositories of linguistic documentation. CRediT taxonomy does not attribute authorship, but only formalizes the type of contribution to the scientific product [12, 13]. Also, the CRediT taxonomy specification is more precise than the copyright law.

New linguistic documentation projects have to provide in their design the roles of contributors and copyright. These definitions impact sharing.

C. Sharing

The goal of providing resources for the description of spoken and written Portuguese in Brazil and for educational programs makes the product resulting from the collective undertaking of Brazilian sociolinguistic projects a tangible asset that is not exhausted in itself: sharing is one of the inherent characteristics in the constitution of a linguistic documentation repository [1, 9]. However, although ideally shareable, the circulation of the product takes on barriers associated with

For their support. National Council for Scientific and Technological Development (CNPq), Brazil.

copyrights and ethical aspects, which must be considered in the data management plan.

Since a linguistic repository is an intellectual property product, the legally responsible author (collective work) or the coauthors hold the copyright, which can be Copyright (©) type, which protects the author's exclusive right to take advantage of their product, whether for commercial purposes or not, or Creative Commons (CC), a range of open licenses that encourage reuse and free circulation of authorial products, which involve acknowledging authorship (BY), sharing the product as it is made available (SA), allowing only non-commercial use (NC), or not allowing derivative works from the original (ND).

A linguistic documentation repository may, for example, have a less open license with all rights reserved, or more open licenses that allow reuse but prevent commercial use, or allow unrestricted use as long as the authorship is acknowledged. The data management plan needs to provide the type of license to be assigned to the final product.

D. Funding

Even though speech is free, there are costs involved in making a set of linguistic data systematically organized available in linguistic documentation repositories. To start linguistic documentation, institutional conditions are required: physical space for project allocation and human resources (researchers and assistants). For researchers to be able to elaborate a plan for the documentation and management of the linguistic data, it is necessary to have time allocated for this purpose. In addition, the management of a linguistic documentation project requires specific technical expertise, especially in audiovisual technology, where research assistants available to the project would be the ideal situation (with appropriate pricing in the final product).

After conception, the work team for the constitution of the linguistic samples needs to be trained (which requires the mobilization of a specific structure for this purpose) to develop the specific activities, providing a highly specialized technical service for language documentation. Once the constitution period is over, the data management plan needs to consider the maintenance of the repositories for a long time, or at the project level, which involves annual maintenance costs for as long as the sample remains available, whether the access is more or less open.

Discontinuity of funding accelerates the depreciation of the linguistic documentation repository, which without personal investment becomes obsolete. Loss of tapes, hacking into unprotected servers, and lack of data back-up are all risks arising from the absence of a funding policy for the maintenance of linguistic documentation holdings. It should be emphasized that this is not a problem exclusive to the area of linguistic documentation, but a broader and more systemic problem for all forms of cultural heritage preservation in Brazil.

E. Management plan for Brazilian linguistic documentation repositories

A data management plan is a document that describes the procedures for collecting, processing, organizing, storing, and preserving data, at all stages of a research project. However, not all projects present this plan, not only because it has not been a requirement, but also because some issues still need further discussion.

The Open Science movement for research replicability empowers the repositories of Brazilian sociolinguistic projects as a privileged source for linguistic descriptions. However, the policy of access to the data from these projects is not always explicit to the community. This restriction policy has implications for Open Science requirements, such as those of publications that have as a submission requirement access to the dataset. In Brazilian sociolinguistics, the arguments in favor of a restrictive access policy evoke the waste of time and financial resources in the constitution of the sample, which would generate intellectual property and the right to primacy in the description of linguistic phenomena.

On the one hand, the arguments in favor of an expanded access and sharing policy evoke the nature of public funding of research projects that give rise to products. A linguistic documentation repository is a product subject to all intellectual property laws, and as a product, it should circulate in the community for transparency in research and equity of access to the results, promoting social justice.

The funding argument needs to be relativized because not all costs are covered by project funding, and, even when funding exists (which is not always the case), it is not enough to cover all the steps of the data collection process. Accountability and social justice arising from public funding guarantee the right of access to data, which is not to be confused with total and unrestricted availability; after all, the responsibility for the use and reuse of data is on the authors (responsible researchers, controllers, and organizers).

On the other hand, the starting point of the Open Science movement is transparency and replicability of analysis: does the data actually exist? Will another researcher replicate the same procedures and achieve the same results? Due to this principle, journals have been stimulating the availability of repositories.

Thinking about the sustainability of projects to build linguistic documentation repositories, partnerships with the information technology area, or even companies, could minimize problems of obsolescence and safeguarding of data, by promoting the circulation and automation of analysis through natural language processing algorithms.

These planning actions may help to promote the longevity of the linguistic documentation repositories of Brazilian sociolinguistic research.

REFERENCES

[1] R. M. K. Freitag, et al., "Challenges of Linguistic Data Management and Open Science", CadLin, vol. 2, no. 1, pp. 01-19, Apr. 2021.

[2] L. A. Silva, "Projeto NURC: histórico," Linha D'Água, vol. 10, pp.83–90, 1996.

[3] M. Oliveira Jr., "NURC Digital Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC)," CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, vol. 3, n. 2, pp. 149–174, 2016.

[4] M. Oliveira Jr., NURC - 50 anos: 1969-2019. São Paulo, SP: Parábola Editoral, 2019.

[5] M. C. Paiva, and M. M. P. Scherre, "Retrospectiva sociolingüística: contribuições do PEUL," DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, vol. 15, n. spe, pp 201–232, 1999.

[6] M. C. Paiva, and C. A. Gomes, "Grupo PEUL: passado, presente e futuro de uma agenda de pesquisa," Cadernos de Estudos Lingüísticos, vol. 58, n. 3, pp.503–519, 2016.

[7] G. Collischonn, and V. O. Monaretto, "Banco de dados VARSUL: a relevância de suas características e a abrangência de seus resultados, " Alfa: Revista de Linguística, vol. 56, n.3, pp.835–853, 2012.

[8] L. Bisol, and V. O. Monaretto, "VARSUL e suas origens, uma história sumariada," Revista virtual de estudos da linguagem–ReVEL, vol. 14, n. 13, pp. vi–xi, 2016.

[9] R. M. K. Freitag, M. A. Martins, and M. A, Tavares, "Bancos de dados sociolinguísticos do português brasileiro e os estudos de terceira onda: potencialidades e limitações," Alfa: Revista de Linguística, vol. 56, n. 3, pp.917–944, 2012.

[10] A. Brand, et al., "Beyond authorship: attribution, contribution, collaboration, and credit," Learned Publishing, vol. 28, n. 2, pp.151–155, 2015.

[11] A. Liz, et al., "Publishing: Credit where credit is due," Nature News, vol. 508, n. 7496, pp.312, 2014.

[12] A. Liz, A. O'Connell, and Veronique Kiermer, "How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship," Learned Publishing, vol. 32, n.1, pp.71–74, 2019.

[13] A. Holcombre, "Farewell authors, hello contributors," Nature, vol. 571, n. 7763, pp. 147–148, 2019.

Best Practices for Information Architecture, Organization, and Retrieval in Digital Language Archives within University Institutional Repositories

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

***Abstract*— This report presents a case study about building a working digital language archive in a hosted university institutional repository. Best practices in language documentation regarding information architecture, organization, and retrieval are considered in relation to university library commitments to resource acquisition/preservation and online cataloging/delivery systems. Despite challenges, findings suggest that constructing digital language archives in university institutional repositories may offer viable collaborative solutions for researchers unable to find suitable, pre-existing archives in which to deposit their language documentary materials. The report concludes that, in such situations, the ability to satisfy best practices may respond to the strengths/weaknesses of particular software implementations as much as it reflects the design team's vision, as theory and method in language documentation increasingly become matters of library and information science.**

Keywords—institutional repository, university library, digital language archive, language documentation theory, best practices, information architecture, information organization, information retrieval, spoken language corpora, Spanish, DARDOSIPCAT

I. Introduction and Statement of Questions

The 21st century has seen an explosion in digital scholarship in the humanities and social sciences. With regard to language sciences, digital recording/dissemination technologies have allowed researchers to largely disentangle language description and language documentation [1], leading to an evolution in our understanding of what constitutes a modern linguistic record. Traditional print outputs are no longer the gold standard [2], and expectations grow annually for language researchers to provide wider electronic access to our data. With regard to library and information sciences, Dublin Core metadata formats and The Open Archives Initiative Protocol for Metadata Harvesting have paved the way for the rise of institutional repositories (IRs) at universities worldwide.

In 2021, it is now highly desirable for language researchers to archive language documentary materials digitally, and digital language archives (DLAs) abound these days. Problems may arise, however, when researchers attempt to find an appropriate DLA in which to deposit their language documentary materials [3]. Such problems may include, among others: languages archived, type of language data archived, type of speech data available, levels of access available, fees for service, audience design, degree of archive user-friendliness, and degree of archive sustainability. When, due to such problems, researchers are unable to find a suitable, pre-existing DLA in which to deposit their language documentary materials, can university IRs (UIRs) offer a viable solution? What are the challenges of building a working DLA within a UIR? In discussing such challenges, this report focuses on best practices in language documentation in relation to DLA information architecture, organization, and retrieval in UIRs.

II. Review of Literature

Over the last two decades, language documentation theory has emerged in various key publications [1, 4-8] that outline guiding principles for the field. These works demonstrate the need for digitally archiving multipurpose records of language in the form of primary data (recordings of spoken language) and apparatus (metadata and transcriptions) to responsibly preserve/disseminate such records for future uses yet unknown. These publications agree that DLAs hold the key to contemporary language documentations. In this regard, Bird and Simons [9] analyzed seven dimensions of data portability for digital language resources, with best practice recommendations that have since become seminal.

Often under the aegis of university libraries, UIRs provide the infrastructure/tools necessary for depositing, preserving, and delivering digital language resources. Consequently, UIRs provide a compelling means of archiving digital language data. Universities worldwide increasingly use repository software, which may be free or licensed and university-integrated or commercially-hosted. The two largest IR software implementations in US universities are dSPACE (120 research institutions/departments) and bepress (419 research institutions/departments) [10]. While dSPACE [11] is an open-

This research was partially funded by Western Michigan University.

source project of LYRASIS, a non-profit organization, bepress provides for-profit, hosted services licensed to universities [12].

Rapidly expanding UIRs now include DLAs. The Virtual Linguistics Lab at Cornell [13] is an example of a dSPACE implementation that, since 2010, has housed a DLA in a UIR. Pilot initiatives like Cornell's are valuable insofar as they provide proof-of-concept that UIR infrastructure suits the information architecture, organization, and retrieval needs of modern DLAs. In this regard, pre-existing commitments to resource acquisition/preservation, online cataloging/delivery systems, and access make UIR-based DLA solutions easier than DLA solutions outside UIRs. Moreover, UIRs also demonstrate strong sustainability potentials for digital resource curation [14]. Unlike "standalone" DLAs whose sustainability may be uncertain due to dependences on grant renewals or sufficient user fees, DLA sustainability in UIRs is relatively dependable. Recent research [15] suggests a trend in sustainable support for UIRs due to their frequent absorption into regular university budgets.

III. Project Design

This section focuses on best practices in language documentation for building working DLAs within UIRs, based on the author's experience designing the UIR-based DLA known as the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT. DARDOSIPCAT is dedicated to archiving/disseminating language resources from the Països Catalans. Resources on deposit include audio recordings of interviews about language and society and orthographic transcriptions of these recordings.

Essentially, best practices for building working DLAs within UIRs stem from epistemological responsibilities with regard to the disposition of language data. These responsibilities include documenting data digitally, archiving language documentary materials, and incorporating language documentation theory into information architecture, organization, and retrieval. These responsibilities relate to both the making of DARDOSIPCAT (how it came to be) and the makings of DARDOSIPCAT (the nuts and bolts that hold it together).

A. The Making of DARDOSIPCAT

My early experiences in digital language documentation trace to the Electronic Metadata for Endangered Languages Data [16] language digitization project conferences of the early 2000s, where the seeds of DARDOSIPCAT were planted. These years coincided with the rise of language documentation theory [1, 7, 17]. From these works, it became clear that good documentations should be diverse, large, ongoing, distributed, opportunistic, transparent, preservable, portable, and ethical. Accordingly, mission-critical elements of working DLAs within UIRs would have to focus on creating an architecture for collecting/transcribing/archiving primary data digitally, developing transparent corpus-level and resource-level metadata to create lasting, multipurpose records of observable community linguistic behavior, and promoting maximum accessibility. These best practices would guide the making of DARDOSIPCAT.

A grant awarded in 1998 funded transcribing a corpus of spoken language that would become DARSOSIPCAT's first deposit. Sabbatical leave in 2002-2003 enabled digitizing 20+ hours of audiotaped conversations and editing 500+ pages of transcriptions. Once the UIR at my university, a bepress implementation called ScholarWorks at WMU, became fully operational under university library administration, I collaborated extensively with the UIR's director (a university librarian) and bepress representatives to create architecture within the UIR for the DLA that would become DARDOSIPCAT. This collaboration involved (a) organizing respective DLA series for audio recordings, video recordings, and orthographic transcriptions; (b) strategically mapping Dublin Core elements to customized metadata fields for optimal harvesting/subsequent information retrieval; and (c) designing/testing extensive resource deposit forms leveraging the customized metadata.

DARDOSIPCAT's first recordings were deposited in 2013. Additional resources were added in 2015, 2018, and 2020 as successive grants were obtained in support of various RA collaborations. These collaborations involved, inter alia, digitizing additional recordings, anonymizing recordings with "bleep" tones, creating associated transcriptions, and developing resource pages with catalog metadata and hyperlinks to related resources.

B. The Makings of DARDOSIPCAT

ScholarWorks is a hosted UIR that runs on Digital Commons software. Advantages of being hosted include a proven track record, an "off-the-shelf" product for quick startup, customizable options, strong tech support, and built-in human resources at bepress. Disadvantages include two layers of UIR-related administration and "one size fits all" software limitations that in practice lead to inevitable design compromises.

Currently, DARDOSIPCAT features two (longitudinally-related) corpora on deposit, with resources distributed among three active series: two primary data series (audio recordings and video recordings) and one transcriptions series. Together the collections comprise 253 distinct catalog pages with highly robust metadata and cross-referenced hyperlinks that promote resource discovery/access within the archive by connecting metadata about individual resources to metadata about all related resources throughout the DLA, longitudinally between corpora and by modality within each corpus. Present resources include 113 audio recording resource pages, 18 video recording resource pages, 113 transcription resource pages, and a 9 page contributors galley. Many audio recording resource pages have back-end WAV archival master files (56) and front-end MP3 downloads available (37). Such audio files are encrusted with Dublin Core metadata for harvest. Future work will include uploading more audio files as well as PDF transcriptions and general access resources including usage conditions, user authorization instructions, and explanations of metadata terminology/mapping.

IV. Findings

This section reports my findings regarding the ability of my UIR to accommodate a DLA in compliance with Bird and Simons' [9] best practice recommendations (henceforth BSBPRs) regarding incorporating language documentation theory into information architecture, organization, and retrieval.

BSBPR for FORMATTING involve information retrieval, organization, and architecture issues such as ***openness***, ***markup***, and ***rendering*** respectively. ***Openness*** refers to using language resources without special software. DARDOSIPCAT's access copies are served in MP3 and PDF formats. Such published, proprietary formats are preferred over secret proprietary formats, though nonproprietary formats are considered ideal. ***Markup*** refers to how (meta)data are represented. ScholarWorks's bepress engine supports XML and the Open Archives Initiative Protocol for Metadata Harvesting (V2). ***Rendering*** refers to presenting materials in conventionally formatted displays. DARDOSIPCAT serves MP3 audios that play on common media players and orthographic transcriptions that open in conventional PDF readers.

BSBPR for DISCOVERY involve retrieval issues such as ***existence*** and ***relevance***. ***Existence*** refers to finding resources easily. DARDOSIPCAT resources are full-text indexed in major search engines like Google thanks to metadata cataloging/optimization. ***Relevance*** refers to judging appropriateness of language resources. DARDOSIPCAT's resource pages' extensive catalog metadata facilitate such judgments.

BSBPR for CITATION involve information organization and architecture issues such as ***bibliography***, ***persistence*** of electronic resource identifiers, and ***immutability*** of citable materials. ***Bibliography*** refers to making referential citations of electronic resources. ScholarWorks's bepress implementation lists complete bibliographic information for resources in catalog metadata and provides recommended citations. ***Persistence*** refers to URL breakage. DARDOSIPCAT boasts persistent URLs. ***Immutability*** refers to URL versioning; DARDOSIPCAT's required catalog metadata includes last-revision dates.

BSBPR for PRESERVATION involve information retrieval issues such as ***longevity*** and ***safety***. ***Longevity*** refers to digital resource lifespans/degradation. DARDOSIPCAT stores archival resources in formats likely to endure for generations (WAV and PDF/A). ***Safety*** refers to potential catastrophic loss. DARDOSIPCAT boasts redundant off-site backup copies.

BSBPR for RIGHTS involve information retrieval issues such as ***public benefits***, which refer to user rights to fair use. Enterprise-level login/password safeguards restrict access to DARDOSIPCAT resources. Unfortunately, however, ScholarWorks's bepress implementation provides no automated way to ensure that only bona fide researchers obtain DARDOSIPCAT accounts.

V. Significance

The finding that DARDOSIPCAT has been able to meet BSBPRs as described above is significant in terms of the ability of UIRs with bepress implementations to incorporate language documentation theory into DLA information architecture, organization, and retrieval. This finding suggests that, given the right circumstances, such UIRs can indeed offer viable solutions for researchers to build their own DLAs. The challenges involved in building working DLAs within such UIRs are also significant, however. Developing such DLAs depends on the prior existence of such UIRs and the willingness of UIR directors and bepress employees to collaborate on such projects. Moreover, while the robust metadata handling in bepress UIR implementations is significant for optimal harvesting, indexing, and retrieval of DLA documentary resources, this significance is constrained by the fact that bepress services are merely licensed. One wonders what would become of DLA metadata in a bepress UIR implementation were that UIR to give up its bepress licensure. Similarly, one wonders to what degree the significance (and limitations) of the results presented here may extend to dSPACE UIR implementations as well. Further research is warranted in this regard.

VI. Conclusions

This report has explored a case study of constructing a working DLA inside a UIR to determine the viability of such a solution for researchers unable to find a suitable, pre-existing DLA in which to deposit their language documentary materials. The approaches, methods, and techniques for collection development described above lead to the inevitable conclusion that such an endeavor is a long-term proposition involving multiple, distributed collaborations with university librarians, RAs, and IR software consultants. This conclusion is consistent with prevailing wisdom in language documentation theory [18].

Informed by such theory, construction of DARDOSIPCAT followed BSBPR for language documentation inasmuch as resources allowed. In the end, the ability to satisfy best practices in relation to DLA information architecture, organization, and retrieval in DARDOSIPCAT was conditioned by the strengths/weaknesses of my institution's bepress UIR implementation. The creative vision of the design team, though realized in large part, was not fully achieved.

Since language documentation today depends so heavily on library and information science, one could conclude that the future is bright for DLAs in bepress implementations of UIRs managed by university libraries. In particular, the administration of such DLAs can build on university library strengths in resource acquisition/preservation, online cataloging/delivery systems, and resource retrieval. In this regard, future research regarding archiving language documentary materials in such DLAs may come to see language documentation theory and method as epistemologically indistinguishable from library and information science.

References

[1] N. P. Himmelmann, "Documentary and descriptive linguistics," Linguistics, vol. 36, no. 1, pp. 161-195, 1998, doi:10.1515/ling.1998.36.1.161.

[2] J. Good, "Valuing technology: Finding the linguist's place in a new technological universe," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 111-131.

[3] R. E. Vann, "Frustrations of the documentary linguist: The state of the art in digital language archiving and the archive that wasn't," in Proceedings of the 2006 E-MELD Workshop on Digital Language Documentation (Tools and Standards: The State of the Art), Michigan State University, East Lansing, MI, June 20-22, 2006, [Online]. Available: http://emeld.org/workshop/2006/proceedings.html

[4] P. Austin, "Language documentation 20 years on," in Endangered Languages and Languages in Danger: Issues of Documentation, Policy, and Language Rights, L. Filipović and M. Pütz, Eds. Amsterdam: John Benjamins, 2015, pp. 147–170.

[5] L. A. Grenoble and N. L. Furbee, Eds. Language Documentation: Practice and Values. Amsterdam/Philadelphia: John Benjamins, 2010.

[6] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1-30.

[7] A. C. Woodbury, "Defining documentary linguistics," in Language Documentation and Description, vol. 1, P. Austin, Ed. London: Hans Rausing Endangered Languges Project, SOAS, 2003, pp. 33-51.

[8] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159–186.

[9] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557-582, 2003, doi: 10.1353/lan.2003.0149.

[10] Registry of open access repositories, University of Southampton, School of Electronics and Computer Science, July 30, 2021. [Online]. Available: http://roar.eprints.org/information.html

[11] "About DSpace." About DSpace - DSpace. https://duraspace.org/dspace/about/ (accessed July 30, 2021).

[12] "Digital Commons." Digital Commons - bepress. https://bepress.com/ (accessed July 30, 2021).

[13] B. Lust, S. Flynn, M. Blume, E. Westbrooks, and T. Tobin, "Constructing adequate language documentation for multifaceted cross-linguistic data: A case study from the Virtual Center for Study of Language Acquisition," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 89-107.

[14] J. McGann, "On creating a usable future," in Profession 2011, R. J. Feal, Ed. New York, NY: Modern Language Association of America, 2011, pp. 182-195.

[15] C. S. Burns, A. Lana, and J. M. Budd, "Institutional repositories: Exploration of costs and value," D-Lib Magazine, vol. 19, no. 1/2, Jan./Feb. 2013, doi: 10.1045/january2013-burns.

[16] "LINGUIST List Projects." E-MELD Homepage. http://emeld.org/ (accessed July 30, 2021).

[17] J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Essentials of Language Documentation. Berlin: Mouton de Gruyter, 2006.

[18] A. Dwyer, "Models of successful collaboration," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 193-212.

How Software Features and Linguistic Analyses Add Value to Orthographic Markup in Transcriptions of Multilingual Recordings for Digital Archives

Enrique Rodríguez
Department of Spanish & Portuguese
Indiana University
Bloomington, USA
enrodri@iu.edu

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

***Abstract*— This report discusses the importance of accounting for language contact and discourse circumstance in orthographic transcriptions of multilingual recordings of spoken language for deposit in digital language archives (DLAs). Our account provides a linguistically informed approach to the multilingual representation of spontaneous speech patterns, taking steps toward documenting ancestral and emergent codes. Our findings lead to portable lessons learned including (a) the conclusion that transcriptions can benefit from a bottom-up approach targeting particular linguistic features of sociocultural relevance to the community documented and (b) the implication (for researchers developing transcriptions for other DLAs) that the principled implementation of particular software features in tandem with systematic linguistic analysis can be helpful in finding and classifying such features, especially in multilingual recordings.**

Keywords—language documentation, digital language archives, spoken language corpora, orthographic transcription, multilingualism, linguistic analysis, transcodic markers, discourse patterns, emergent codes, Spanish, Catalan, DARDOSIPCAT

I. Introduction and Research Questions

This report focuses on accounting for language contact and discourse circumstance in markup of orthographic transcriptions of multilingual recordings of spoken language (henceforth MRSL) for subsequent deposit in digital language archives (henceforth DLA). We address two research questions. First, what linguistic phenomena should such orthographic transcriptions account for in terms of language contact and discourse circumstance? Second, what sort of linguistic analyses may aid in classifying such linguistic phenomena?

In exploring these research questions, we discuss the need to account for words that were unambiguously spoken in languages other than the base-language of the transcription as well as words that were spoken transcodicly [1] in some way. Moreover, we also discuss the need to account for basic discourse phenomena such as overlapped words and interrupted words and turns. Subsequently, we report on useful software features and linguistic analyses that can help in the accurate representation of such linguistic phenomena. Specifically, we discuss strategic use of standard functions in ExpressScribe and Microsoft Word and we explain the importance of phonetic-phonological, morphological, and discourse-pragmatic analyses in identifying and categorizing particular contact phenomena and discourse circumstances.

This research was partially funded by Western Michigan University.

Discussions throughout this paper are based on transcriptions and digital recordings on deposit in the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT. DARDOSIPCAT is a DLA dedicated to collecting, preserving, annotating, cataloging, and disseminating language resources from The Països Catalans. Resources on deposit include longitudinal audio recordings of spoken language made in both Spanish and Catalan as well as orthographic transcriptions of these recordings. The audio recordings represent interviews about language and society in Barcelona. The fact that most research participants spoke in both Spanish and Catalan during their interviews presented multiple challenges for transcribing the recordings, including how to identify and represent potential discourse and contact phenomena perceived in the recordings. With the goal of producing transcriptions that represented the original recordings as faithfully as possible while maintaining easy readability, the DARDOSIPCAT research team innovatively exploited particular software features and carried out diverse linguistic analyses.

II. Literature Review

In our view, all DLAs are a form of language documentation, an interdisciplinary endeavor that aims to create lasting, multipurpose records of language [2]. Given that a central goal of all language documentations is the archiving of the linguistic practices of specific speech communities, systematic recordings of spoken language collected in appropriate sociocultural contexts are vital, as are transcriptions that apply linguistic knowledge to create practical representations, adding value to such primary data [3]. Best practice recommendations regarding the content of such samples [4, p. 571] advocate for comprehensive digital language resources that are "sufficiently broad in scope, rich in detail, and authentic in portrayal that future generations will be able to experience and study the language, even if no speakers remain". Accordingly, for multilingual communities in which the dynamic interaction of languages may lead to all manner of translanguaging, best practices in language documentation include the archiving of transcriptions of spoken language samples of "ancestral and emergent codes" [5] whose very existence may depend on such usage.

Contact-induced language phenomena can manifest in different linguistic systems, hence the need for implementing

linguistic analyses in the transcription of MRSL on deposit in DLAs. For example, while codeswitching as defined by Jakobson, Fant and Halle [6] can be either intentional or spontaneous, Poplack [7] has argued that this practice may be governed by morphosyntactic and phonotactic constraints from either language. Nevertheless, linguistic boundaries between two or more languages often blur, leading to situations in which speakers produce utterances that can be interpreted in multiple languages simultaneously. In this regard, the term *bivalency* refers to "the use by a bilingual of words or segments that could 'belong' equally, descriptively, and even prescriptively, to both codes" [8]. More generally, the term *transcodic marker* [1], a catchall for linguistic innovations that occur in language contact situations, may denote codeswitching, bivalency, borrowings, calques, semantic extensions, and/or spontaneous speech innovations.

Following Vann [9], digital recordings and orthographic transcriptions on deposit in DLAs represent our best hope of finding such phenomena in contact dialects, as well as our best way to reference them, and multifaceted linguistic analyses provide the best way to identify them and the discourse-pragmatic strategies they may represent. Correspondingly, such transcriptions also need to address relevant discourse phenomena that deal with turn sequence and organization in social interaction, such as overlapped and simultaneous talk, interruptions, and realignment across turns and sequences [10, 11], whose discovery and identification is greatly facilitated by linguistic analyses that take into account context-dependent conversational actions.

III. Methodology

A. Software

Audio recordings were played in ExpressScribe while corresponding transcriptions were written in Microsoft Word. Functions of these two software applications were key to our accounting for both language contact and basic discourse circumstance in orthographic transcriptions of MRSL on deposit in DARDOSIPCAT. Practical implementations are discussed below.

ExpressScribe is freeware that features constant-pitch, variable-speed playback on the fly, as well as user-configurable options for rewinding and advancing playback. These features, particularly the ability to play audio smoothly even at speeds as slow as 25%, were critical to the discovery of the linguistic phenomena under investigation, as the research assistant (henceforth RA) was able to listen meticulously and repeatedly to segments of each recording. Moreover, in a small window within the ExpressScribe interface itself, the RA was able to annotate potential contact and discourse phenomena observed in each recording. These notes were later exported as text files and stored for future research and transcription-related discussions between the Principal Investigator (henceforth PI) and the RA.

As the RA listened to the audio playback in ExpressScribe, the RA typed into Word the utterances that the RA perceived as the RA perceived them, spelling all utterances the way native speakers of Castilian Spanish would typically write down the spoken language heard on the recordings. Word dictionaries set to Spanish were then used to spellcheck the transcriptions. Utterances that the spellchecker flagged as spelled incorrectly in Spanish were considered as potential transcodic markers or discourse phenomena such as false starts or interrupted words. These utterances were then spellchecked in Word dictionaries of other languages to ascertain whether they were in fact words in a language other than Spanish.

B. Linguistic Analyses

Once we had uncovered potential language contact and discourse phenomena thanks to strategically implementing the functions of these two software applications, we used different levels of linguistic analysis to categorize these linguistic phenomena accordingly. Corresponding transcriptional markup followed. Once transcripts were finalized, they were converted to PDF format for deposit in DARDOSIPCAT, where master copies are stored in PDF/A format and access copies are served in basic PDF format.

1) Phonetics and Phonology: In transcribing DARDOSIPCAT interviews, phonetic and phonological features were used to distinguish between Catalan and Spanish words in potential situations of bivalency. Bivalent words and expressions such as *Esquerra Republicana* 'Republican Left' (the name of a political party in Catalonia) and *Polònia* 'Poland', despite ostensibly being Catalan words, were deliberately not always regarded as such by the speakers. Though the decision to determine whether an utterance was being spoken in Spanish or Catalan was a principled one based on phonological criteria, determining the language in which a particular word or expression was being spoken was not always straightforward even when the RA and the PI strongly agreed on the phonology used, because many people in Catalonia speak Spanish with a phonology that may reflect varying degrees of influence from Catalan.

2) Morphology: Morphological criteria were used to determine the language to which a given word or expression belonged in situations of speech innovations due to language contact between Spanish and Catalan. Two examples that illustrate how such criteria were used in transcription are *bastoneres* and *foguerones*. In both cases, we have Catalan-based words, *bastoners* 'emcees' and *foguerons* 'bonfires', that have been borrowed into Spanish with concomitant morphological change (adoption of the Spanish plural agreement suffix *-es*) that make them appear as Spanish to the transcriptionists. These two cases reflect the sort of contact-induced linguistic innovations that abound among bilingual speakers of Spanish and Catalan in the Països Catalans. Without transcodic morphological analyses, such borrowings might have been deemed spontaneous speech errors or, worse, gone undetected entirely.

3) Discourse and Pragmatics: In the transcription of DARDOSIPCAT's audio recordings, pragmatic and prosodic analyses played a role in determining how discourse was organized and co-constructed across turns by participants in the conversations transcribed. Extract (1) illustrates an interrupted question in Spanish (an English translation follows):

R:*¿Puede haber gente castellanohablante* (1)

en las manifestaciones <de>-

X: *<Claro.> Sí, sí.*

R: *De independencia?*

X: *La hay, de hecho.*

R: Is it possible to find Spanish-speaking people

at the demonstrations <for>-

X: <Of course.>. Yeah, yeah.

R: For independence?

X: In fact there are.

Extract (1) begins with the PI asking a question to the interviewee, who answers before the question is concluded, thus interrupting the turn in which the question originated. In this and similar examples, pragmatic analysis was used to determine how to categorize such discourse patterns and mark them accordingly in the transcriptions in a principled way that respected the illocutionary force of the utterances spoken despite subsequent discursive interruptions and force abandonments.

IV. Findings and Significance

The software and linguistic analyses carried out in aid of the transcriptions uncovered extensive contact and discourse phenomena that may be of interest to future users of our DLA. In terms of contact, our linguistic analyses revealed transcodic markers including codeswitching, codemixing, bivalency, borrowings, calques, semantic extensions, and spontaneous speech innovations. In terms of discourse, linguistic analyses revealed numerous performance errors, overlapped words, and interrupted words and turns.

To determine which of these phenomena to include in DARDOSIPCAT transcriptions, we considered the best practice recommendations described in Section II. In light of these recommendations, our findings regarding Question 1 are that orthographic transcriptions of MRSL on deposit in DLAs should account for ALL perceivable instances of language contact and spontaneous discourse patterns to the extent that they can do so in a user-friendly way. Simple orthographic conventions, easy-to-read formatting, and minimal markup should prevail so all users can easily understand the transcriptions without linguistic training. This finding is significant as it highlights the importance of transcriptions with sufficient detail to represent linguistic phenomena salient to the community under documentation. In DARDOSIPCAT, accounting for contact and discourse phenomena is key to faithful documentations of significant linguistic patterns in the community's ancestral and emergent codes [5]. These patterns may hold evidence of potential changes in progress. Additionally, this finding provides a straightforward way for researchers to locate and identify such phenomena within the transcripts themselves.

With regard to Question 2, we found that linguistic analysis in the areas of phonetics, phonology, morphology, and pragmatics was useful in identifying and categorizing relevant linguistic phenomena in the transcription of DARDOSIPCAT interviews. Given their compartmentalized nature, we believe such analyses could be useful in creating transcriptions for other DLAs as well, separately or in combination, depending on the phenomena of interest to the transcriptions' audience design. Incorporating linguistic analyses into the transcription process is important for accountability in research resources [2] insofar as accurate rendering of linguistic transcripts strengthens the empirical foundations of those branches of linguistics and related disciplines whose work depends on quality documentary resources.

V. Conclusions

The present report set out to discover what linguistic phenomena orthographic transcriptions of MRSL on deposit in DLAs should account for in terms of language contact and discourse circumstance and to describe software features and linguistic analyses that support the accurate representation of these linguistic phenomena in such transcriptions. While the issues addressed here relate to linguistic phenomena that are particularly pertinent to DARDOSIPCAT, our research questions and methods have implications for other DLAs, especially those that also document MRSLs. Our findings suggest that such DLAs can benefit from a bottom-up approach that targets specific linguistic features relevant to the speech community at hand. Accordingly, the principled implementation of particular software features in tandem with systematic linguistic analysis can be most helpful in this regard.

References

[1] G. Lüdi, "Les marques transcodiques: regards noveaux sur le bilinguisme," in Devenir Bilingüe - Parler Bilingüe. Actes du 2e Colloque sur le Bilinguisme, Université de Neuchâtel, Niemeyer, Tübingen: Max Niemeyer Verlag, 1984, pp. 1–19.

[2] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1–30.

[3] P. Austin, "Language documentation 20 years on," in Endangered Languages and Languages in Danger: Issues of Documentation, Policy, and Language Rights, L. Filipović and M. Pütz, Eds. Amsterdam: Benjamins, 2015, pp. 147–170.

[4] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557–582, 2003, doi: 10.1353/lan.2003.0149.

[5] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159–186.

[6] R. Jakobson, G. Fant, and M. Halle, Preliminaries to Speech Analysis: The Distinctive Features and their Correlates. Cambridge, MA: MIT Press, 1952.

[7] S. Poplack, "Sometimes I'll start a sentence in Spanish y termino en español": Toward a typology of code-switching," Linguistics, vol. 18, no. 7/8, pp. 581–618, 1980, 10.1515/ling-2013-0039.

[8] K. Woolard. "Simultaneity and bivalency as strategies in bilingualism," Journal of Linguistic Anthropology, vol. 8, no. 1, pp. 3–29, 1998.

[9] R. E. Vann, "On the importance of spontaneous speech innovations in language contact situations," in Convergence and Divergence in

Language Contact Situations, K. Braunmüller and J. House, Eds. Amsterdam: Benjamins, 2009, pp. 153-182.

[10] H. Sacks, E. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking,” Language, vol. 50, no. 4, pp. 696–735, 1980.

[11] E. Schegloff, Sequence Organization in Interaction: A Primer in Conversation Analysis. Cambridge, MA: Cambridge University Press, 2007.

Challenges to Representing Personal Names and Language Names in Language Archives: Examples from Northeast India

Mary Burke†
Department of Information Science
University of North Texas
Denton TX, USA
Mary.Burke@unt.edu

Shobhana Chelliah
Department of Linguistics
University of North Texas
Denton TX, USA
Shobhana.Chelliah@unt.edu

ABSTRACT

Language archives are not only a valuable resource for language communities to tell their stories and to create lasting records of their ways of life, but also for those interested in anthropology, linguistics, agriculture, or art history. This recent emphasis on archiving primary datasets in linguistics has resulted in an abundance of datasets online; however, of the languages of South Asia, only a small percentage are represented in digital language archives or described thoroughly. Though several of these languages are being documented, this material is at risk of being lost or inaccessible without concerted attention paid to long-term preservation. There are several obstacles to documenting and archiving language materials from this area, including political instability and lack of access to infrastructure. This submission reviews one particular challenge to data management relevant to South Asia, which is the complexity of names (of individuals, groups, and languages). We provide examples from Northeast India and recommendations based on experience from CoRSAL (Computational Resource for South Asia).

KEYWORDS

Personal names, language names, name authority control, information organization in language archives

1 Introduction

Language archives are not only a valuable resource for language communities to tell their stories and to create lasting records of their ways of life, but also for those interested in anthropology, linguistics, agriculture, or art history. This recent emphasis on archiving primary datasets in linguistics has resulted in an abundance of datasets online; however, of the languages of South Asia, only a small percentage are represented in digital language archives or described thoroughly [1]. Though several of these languages are being documented, this material is at risk of being lost or inaccessible without concerted attention paid to long-term preservation. There are several obstacles to documenting and archiving language materials from this area, including political instability and lack of access to infrastructure. This submission reviews one particular challenge to data management relevant to South Asia, which is the complexity of names (of individuals, groups, and languages). In this paper, we provide examples from Northeast India and recommendations based on experience from CoRSAL (Computational Resource for South Asia).

Name authority control is a crucial component of information organization whereby names are represented uniformly across metadata records despite having multiple possible variants. For example, the Library of Congress Name Authority File (LCNAF) cites 'Zin, Hauard, 1922-2010' as a variant of the authorized name form 'Zinn, Howard, 1922-2010'. Implementing name authority control improves the recall and accuracy of search results. This can be done with a public name authority file, such as the LCNAF or the Virtual International Authority File (VIAF), or using a local name authority file (e.g., a local spreadsheet). Few linguists, students, and language community members have records in a public name authority files, so, in the case of language archives, it is more common to use locally-developed name authority files [2].

2 Variation in personal names

Common practice in metadata records is to represent personal names in an inverted structure with the name parts appearing in the order "Family, Given Middle" -- as in, Turing, Alan Mathison. However, for many communities worldwide, personal names include village names, caste names and/or clan names, and come with a host of variations including abbreviations, nicknames, and differing orders of name parts based on convenience or context. Examples of variation in personal names are described in literature on this topic [3, 4, 5]. This section reviews some examples of variation in personal names through examples from Northeast India.

Name structure may change based on personal preferences, but could also be influenced by large-scale factors like political or religious affiliations, and the impact of colonization. [6] describes

how naming conventions of the Meitheis changed as a result of the ruler Pamheiba (1709-48) introducing Hindu traditions to Manipur, and again during the Manipur renaissance to assert ethnic identity. Beyond just Manipur, the effects of colonization can be seen on personal names when a Western convention is imposed on a previously uncodified system.

2.1 Ordering of name parts

[6] reviews some examples from Northeast India which are relevant here. Following the effects of the Hindu influence of the 1700s, in Manipur, Kshatriya caste names were adopted. For women, 'Devi'; for men, 'Singh'. Thus, the typical name structure for the Meithei community would be:

FAMILY NAME - GIVEN NAME - CASTE NAME
Thounaojam - Harimohon - Singh

However, these name parts may appear in any of the following orders:

FAMILY NAME - GIVEN NAME - CASTE NAME
GIVEN NAME - FAMILY NAME - CASTE NAME
FAMILY NAME - GIVEN NAME
GIVEN NAME - FAMILY NAME

Each order is possible, and the choice seems to depend on personal preference given a certain situation or purpose. Thounaojam Harimohon Singh, for example, might sign an informal letter *Harimohon Thounaojam*. He signed the acknowledgments section of his dissertation *Th. Harimohon Singh*. When buying a plane ticket, however, one needs to ensure the name matches what is written on their passport, and this may yield yet another ordering of the name parts.

2.2 Inclusion, omission, and abbreviation of name parts

Another situation is illustrated by nicknames used by the Lamkangs of Manipur. In this community, individuals are sometimes identified by nicknames or birth order names (e.g., first, second, third born). They may also have Lamkang given names and additional Christian names given at the time of baptism.

Take, for example, a young Lamkang speaker recorded as part of the Lamkang project in 2017. In conversation, he introduced himself as 'Koko,' a nickname assigned based on birth order. On the IRB paperwork, he signed his full name, 'William Shilshi.' In preparing the recordings for archiving, the research team had difficulty choosing the 'correct' name to include in metadata records because this speaker was identified only as 'Koko' except for on the formal paperwork.

It is useful to know that these conventions and styles of identification exist so that documenters can properly keep track of individuals' personal names. For example, for Lamkang names, it might be useful to provide a known nickname in parentheses after the given name: *William (Koko) Shilshi*. Both the given name and the nickname are needed to identify the individual, as many people may have the same nickname.

Another example from the Lamkang community involves individuals taking on additional names for the purpose of affirming identity (e.g., *Swamy Tholung* ***Ksen***). In this example, *Ksen* is neither a given name nor part of the birth order naming system described in this paper, but was adopted to represent the individual's pride to be Lamkang—the Lamkang refer to themselves as *ksen mi*, 'the red people.'

Alternatively, some individuals who take part in language documentation projects do not provide their full name to the collector of the material, but rather only a first name. This may cause confusion when depositors prepare for archiving because it is not clear whether the individual is distinct from another with the same first name.

Name parts may also be excluded for convenience. For instance, because all Meithei males assumed the caste name '*Singh*', one may choose to omit the '*Singh*' to avoid confusion. Also for convenience, or to save space, the family name may be abbreviated, as seen in the following examples ([6], p. 3):

- *L. Bhima Singh*
- *N. Promodini Devi*
- *W. Ibemu Devi*
- *M. Kirti Singh*
- *Ch. Yashawanta Singh*

It is important to note that the same abbreviation might be used for different family names (e.g., both *Layrikyengbam* and *Leimapokpam* are abbreviated 'L.').

This section has illustrated the complexity of naming systems with two examples from Northeast India: the Meithei community, and the Lamkang community. Still, this is by no means a comprehensive review of every possibility—every group will have their own practices surrounding personal names. The following section briefly reviews the multiple forms of language names.

3 Language names

Much of the early recorded history in Northeast India was created by colonial officials, resulting in inaccurate or incomplete names for languages (called *glossonyms* in [7]) and the groups of people who speak them (*ethnonyms*). [8] notes that a language name may be falsely recorded as the name of the village where it is spoken, or as an outsiders' name for the group (*exonym*). There may be one term used for the group of speakers, the village where the language is spoken, and the language itself (*glossonym*), or multiple names for each.

Further, any one language name is likely to have multiple spellings or pronunciations (*allonyms*) or variants (*allograms*) [7]. Spelling differences can be attributed to Westernizations, or differences in transliterations and transcriptions. Take, for example, pairs like *Tsáirel* and *Chairel*, or *Hlota* and *Lhota*. The name used for a language may change depending on individual preferences, or over

time. See [7, 8] for extensive discussion of this phenomenon specific to the Tibeto-Burman language family.

Generally, one would rely on a standard to navigate the variants and select the 'correct' name for a language. The primary standards for representing language names are the ISO 639-2 and 639-3 codes, developed by the Summer Institute of Linguistics (now called SIL International), a Bible translation organization. Though this standard is used by the linguistics community at large, it is maintained only by SIL International. Many take issue with the ISO 639-3 standards [9, 10, 11, 12] and, depending on the languages in focus, prefer to use Glottocodes or a regional standard like AUSTLANG instead.

4 Discussion and Recommendations

This paper has briefly demonstrated the ways in which personal names and language names may vary across contexts. To summarize, we offer some basic recommendations to maximize consistency when representing these names in archival collections.

For archival staff:

- When reviewing metadata, ask depositors about personal names
- In the case that two names look similar, it is better to ask for clarification than to assume this is a typo. (e.g., Abdullah and Abdulla may refer to two separate individuals.)

For language documenters/ depositors:

- Understand naming conventions used in the area. What strategies do people use to identify themselves and each other, and what factors affect this?
- Make note of all possible name forms, and make these clear in the consultant or language contributor records (e.g., spreadsheets, SayMore).
- Include pictures of contributors as recommended in [13].
- If abbreviations are common in the naming convention, be aware of the expanded names from which these abbreviations arise.

5 Conclusion

Given that there is such variation, when it comes time for archiving, it is necessary for depositors and archival staff to be familiar with different naming conventions that may apply to the language or region in question. These issues need not prevent the languages of South Asia from the level of digital accessibility that archives provide. With a better understanding of the structure of personal names, we can represent personal names as consistently and accurately as possible.

REFERENCES

[1] Post, M. 2020. The Eastern Himalayan ethno-linguistic diversity hotspot: Where is it, why is it significant, why is it endangered, and what should "we" do about it? ISSN: 2249-1511.

[2] Burke, Zavalina, O. L., Phillips, M, Author. Organization of knowledge and information in digital archives of language materials. Journal of Library Metadata. doi: 10.1080/19386389.2020.1908651

[3] Akinnaso, F. N. 1980. The Sociolinguistic Basis of Yoruba Personal Names. Anthropological Linguistics 22:275—304.

[4] Abd-el-Jawad, H. 1986. A Linguistic and Sociocultural Study of Personal Names in Jordan. Anthropological Linguistics 28:80—94.

[5] Hong, B. 1985. Politeness in Chinese: Impersonal Pronouns and Personal Greetings. Anthropological Linguistics 27:204—13.

[6] Chelliah 2005. Asserting nationhood through personal name choice: The case of the Meithei of northeast India. Anthropological linguistics, 169-216.

[7] Matisoff, J. A. 1996. General introduction. In Matisoff, J. A., Baron, S. P., & Lowe, J. B. eds. (1996). Languages and dialects of Tibeto-Burman. Sino-Tibetan Etymological Dictionary and Thesaurus Project, Center for Southeast Asia Studies, University of California, Berkeley.

[8] Marrison, G. E. 1967. The classification of the Naga languages of North East India. (Doctoral dissertation, SOAS, University of London).

[9] Dobrin, L. M., & Good, J. 2009. Practical language development: Whose mission? Language, 85(3), 619-629.

[10] Epps, P. et al. 2006. Letter of opposition to adopting Ethnologue's language codes for ISO 639-3. SSILA Bulletin, 246.

[11] Errington, J. 2008. Linguistics in a colonial world. Oxford: Blackwell.

[12] Morey, S., Post, M. & Friedman, V. A. 2013. The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization. Paper presented at the Workshop Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures. Melbourne, University of Melbourne https://ses.library.usyd.edu.au/handle/2123/9838

[13] Everett, D. & Sakel, J. 2012. *Linguistic Fieldwork*. Cambridge University Press, Cambridge.

Creating Workflow for Mediated Archiving in CoRSAL

Merrion Dale
Department of Information Science

University of North Texas
Denton TX, USA
JessicaDale@my.unt.edu

ABSTRACT
Language archiving involves the collection and curation of a variety of language materials. As an emerging language archive, CoRSAL caters to a range of different language depositors with different research needs. As such, we have developed a workflow process that can accommodate this diversity.

CCS Concepts
• Information systems~Information systems applications~Digital libraries and archives • **Human-centered computing~Interaction design~Interaction design process and methods~User centered design • Human-centered computing~Collaborative and social computing~Collaborative and social computing theory, concepts and paradigms~social media • Human-centered computing~Collaborative and social computing~Collaborative and social computing theory, concepts and paradigms~Social content sharing**

KEYWORDS
Language archiving, Participatory archiving, Community archiving, Archival workflow, Archival engagement, Social media

1 Introduction

The Computational Resource for South Asian Languages (CoRSAL) is an archive housed within UNT's Digital Library that specializes in the curation and preservation of source audio, video, and textual language data. CoRSAL is a dedicated repository for linguists to archive grammars and related grammatical information of South Asian languages, many of which are endangered and/or under-resourced. Since its inception in 2016, a primary objective for CoRSAL has been to engage closely with community depositors in each step of the archiving process, from pre-archival decision making to post-archival collection promotion. The ingest process is purposely mediated and participatory because of our diverse depositor base. This paper offers an exploration of CoRSAL's curation and ingest workflow and our findings at this fairly early stage in the growth of CoRSAL. I also include a brief discussion of our desired direction for the future. The results and feedback from our mediated approach have been encouraging so far. For this reason, we will explore additional opportunities to involve speaker-community members in the archiving process, and in increasing their engagement with archived collections.

2 Background

The emergence of a participatory archiving framework in the twenty-first century is contributing to a shift in archival power dynamics [10]. The participatory framework privileges the power of those connected through ownership or cultural heritage to archival materials so that they may play integral roles in the archiving process as archive designers, researchers, depositors, and users [7, 15]. As outlined in Cook as well as in Roeschley and Kim [7, 14] participatory archives may, depending on the wishes and needs of the community, be facilitated, and established by institutional archives.

CoRSAL's approach to archiving falls within the participatory framework by implementing a mediated workflow that is very much a conversational exchange between depositor and curator and resulting in a collection that is representative of each depositor/community's research needs. Additionally, community depositors can engage in post-archival participatory measures in the sharing and promotion of their collections. It is an ongoing concern that general engagement with archived language collections is low among members of the language communities. One contributing factor, based on Burke and Zavalina's [2] interviews with linguistics researchers/depositors, is that some depositors, especially those from Indigenous communities, access information through social media. The researchers also found that many of the Indigenous have limited access to reliable, high-speed internet. As a result, the community's individuals use mobile devices in place of computers and find streaming media platforms like Facebook, Youtube, and Instagram more realistic to access than potentially downloading an archive's large files. This contributes to an overall preference for a "social media-type interface over the archival access point" [2, 3].

Thus, as a final piece of the mediation between the depositor and the archive, and to facilitate the use of the archive as framed by the depositor, CoRSAL has worked with recent depositors to use that existing interest [4, 5, 6, 8, 9, 11, 12, 13] as a means to increase engagement with language archives. Social media has become a tool for Indigenous communities to utilize for purposes of connection, outreach, and activism. As of October 2020 the CoRSAL team has been exploring how to utilize Facebook's private group feature as a tool for depositors to share items with

their community and for the community to discuss the items with the depositors and one another.

3 Methods

Since 2016 the CoRSAL team has done a great deal of outreach work in the form of lectures and workshops. The purpose of this was to explain the process and the purpose for archiving with CoRSAL and offer examples of the ways archived language data can be useful for language revitalization and pedagogy development. As a result of this type of work the team has often received interest from young community researchers who are interested in discussing their own potential collections in greater detail. These conversations have recently resulted in several new language collections in CoRSAL, such as Azamgarhi and Bhojpuri language collections deposited by Maaz Shaikh, and the Boro Language Collection deposited by Prafulla Basumatary.

At present, CoRSAL has eleven published language collections which reflect its diverse depositor pool. For instance, some of the collections, like the Burushaski Language Resource and Lamkang Language Resource, are the result of funded documentary linguistic projects. Others are legacy collections, that is, materials collected in pre-digital times. CoRSAL uses a mediated approach here as well as researchers in possession of a lifetime's worth of fieldwork notes and recordings are often overwhelmed by the tasks of data management and file transfer, and metadata creation. To create metadata for legacy collections, CoRSAL curators collaborate with depositors to aid with data management and metadata recall.

When a depositor is interested in archiving a collection with CoRSAL we work with them hand-in-hand to assist with every step of the process. The workflow consists of communicating regularly over Zoom and email to discuss selection and collection of content, file transfer, the process of filling in the CoRSAL metadata spreadsheet, and general data management. This is the longest part of the process, taking less or more time depending on the state of data management when we begin the process. When there are gaps in the metadata, the CoRSAL team attempts to help fill those. For example, in the case of the Azamgarhi collection, we lacked an identifying language code. A member of the CoRSAL team applied for and obtained a Glottocode for the language on the depositor's behalf. Depositors also have several opportunities to provide context for their collection. For example, we work with depositors to write a brief description of the language and the collection for the collection's landing page in the CoRSAL archive. We assist and edit only as needed. Depositors are also given the option to contribute a thumbnail image for their collection.

Once the digital files and their metadata are transferred to a dedicated CoRSAL drive, the materials are made available to staff at the UNT Digital Library to upload the items to the archive. Despite being published, collections can still be edited after this stage. When a depositor wants to modify their metadata, they can inform us of changes they have made to the spreadsheet and a member of the team implements the changes on their behalf.

Again, keeping our diverse depositor pool in mind, CoRSAL does expect that deposits will be incrementally added to the CoRSAL drive. This possibility of incremental growth of what is to be deposited supports weaker infrastructure for collectors - lack of backup media, unstable computers, a possibility for loss of data. If depositors need to move files into the CoRSAL drive in smaller batches, we allow and encourage this. We don't ask for all of the metadata at once, but rather the key pieces, which depositors can then build upon. We also believe that it is not essential to immediately expect depositors to be ready or able to provide intellectual access to the individual items, i.e., through transcription and translations. Our curators acknowledge that the source data is the most precious and under the most pressing danger of loss. We encourage depositors to approach archiving by starting small and holding the view that the other important items can be obtained as long as we have adequate metadata for the source files. As part of the metadata, we do encourage at least a rough translation. It is our goal to provide training so that depositors are ultimately able to provide different levels of access through transcription and translation using more interlinear glossed texts, and through encouraging the publication of text collections.

Another way in which CoRSAL encourages and then provides help for depositors to "own" their collections is through depositors and associated community members improving metadata, especially in enriching keywords with ethnographic information. In the case of two language collections, Lamkang and Burushaski, two community researchers contacted their peers requesting additional context on a number of photos and videos. The community researcher for the Burushaski collection was able to crowdsource helpful feedback by posting in an existing Burushaski Facebook group [1].

We encourage community depositors to promote their collections on social media as they are interested/able. For example, we have assisted several depositors in creating private Facebook groups. Within these groups the depositor/group admin shares items from their collections with members of their communities. Some depositors share their collection in existing Facebook groups, namely the Burushaski Research Academy group. Maaz Shaikh, the depositor of the Azamgarhi and Bhojpuri collections, recently created his own Azamgarhi Facebook group after observing the models of existing groups.

For each group created as part of this promotional initiative the depositor acts as the primary group admin and is responsible for inviting new members to join. They are also responsible for accepting/rejecting any membership requests the group receives. The admin shares items directly from their collection which they believe will be of specific interest to group members. Members are then able to click on the shared link and view the item in the archive.

4 Findings

Depositors who have completed the archiving and publication process with CoRSAL have given very positive feedback on the workflow and their resulting collections and publications. CoRSAL is also seeing encouraging growth and engagement from the newly created Facebook groups. By monitoring the usage

statistics feature embedded in each collection we are able to see that the shared items are being used via their Facebook referral links. We can thus track the number of uses an item has in general, and more granularly, over the last 30 days. Additionally, Boro group admin/depositor Prafulla Basumatary has reported that he has received useful fieldwork recommendations from community members who have commented on his shared posts.

5 Statement of Significance

A participatory approach to archiving corrects for imbalances in power that unjustly take intellectual, cultural, and material information from one set of people for inquiry by another. The goals of participatory archiving can be greatly supported through a mediated archiving workflow. This helps non-professional collectors and depositors develop an understanding of the archiving process and of how to enhance the usability and accessibility of each collection. While this mediated methodology is time consuming and labor intensive, it gives depositors the power to frame and contextualize a collection. It also requires a sympathetic connection with speakers.

6 Conclusion and Future Work

The CoRSAL team feels encouraged by depositors' responses to the workflow. Boro depositor Prafulla Basumatary is currently preparing a collection guide which will assist users in how best to utilize the collection for revitalization efforts, particularly relating to development of pedagogy. We expect to see a similar guide for the Azamgarhi and Bhojpuri collections come to fruition this year. CoRSAL curators are actively in the process of curating more legacy collections and would like to have them reviewed and vetted by community members before their publishing. We are hoping that inviting community members to promote these collections on Facebook will lead to even greater community viewership and feedback.

Additionally, as the community Facebook groups continue to grow, we are actively researching more ways we can potentially. use social media to enhance the community experience engaging with the archive. Azamgarhi depositor Maaz Shaikh has recently taken the initiative of creating an Azamgarhi Instagram account to promote his Facebook posts to a wider audience and we are excited to see the response this will receive. We are also interested in exploring other social media platforms like WhatsApp and TikTok that may be more appealing to younger community members.

ACKNOWLEDGMENTS

I'm very grateful to Dr. Shobhana Chelliah for her helpful reviews and feedback.

REFERENCES

[1] Mary Burke, Sumshot Khular, Javid Iqbal, and Mark Phillips. 2021. Enriching ethnographic metadata with the help of native speakers Poster presentation. *7th International Conference on Language Documentation and Conservation (ICLDC).* March 4-7, 2021, Honolulu, HI. URI: http://hdl.handle.net/10125/74477.

[2] Mary Burke and Oksana Zavalina. 2020. Identifying Challenges for Information Organization in Language Archives: Preliminary Findings. In Anneli Sundqvist, Gerd Berget, Jan Nolin, and Kjell Skjerdingstad (eds) *Sustainable Digital Communities. iConference 2020.* March 23-26, 2020, Boras, Sweden. Springer, Cham, Switzerland, 622-629. DOI: https://doi.org/10.1007/978-3-030-43687-2_52.

[3] Mary Burke. In press. User Needs in Language Archives: Findings from Interviews with Language Archive Managers, Depositors, and End-Users.

[4] Laura Buszard-Welcher. 2001. Can the Web Help Save My Language? In Leanne Hinton & Ken Hale (eds.) *The green book of language revitalization in practice*, 331–345. Brill, Leiden, Netherlands. DOI: https://doi.org/10.1163/9789004261723

[5] Morgan Cassels. 2019. Indigenous languages in new media: Opportunities and challenges for language revitalization. *Working Papers of the Linguistics Circle of the University of Victoria* 29, 1 (Sept. 2019), 25-43.

[6] Coppelie Cocq. 2015. Indigenous Voices on the Web: Folksonomies and Endangered Languages. Journal of American Folklore 128, 509 (Aug. 2015), 273-285.

[7] Terry Cook. 2013. Evidence, memory, identity, and community: four shifting archival paradigms. *Archival Science* 13, 2 (Jun. 2013), 95–120. DOI: https://doi.org/10.1007/s10502-012-9180-7.

[8] Dario De Falco and Alfonso Cesarano. 2016. Endangered Languages in the Era of Social Media: The Case of the Kenyah Lebu' Kulit Language. In Joseph Cru (Ed.) *Linguapax Review 2016: Digital Media and Language Revitalization*. 55-65. Linguapax International, Barcelona, Spain.

[9] Gabriel Djomeni and Etienne Sadembouo. 2016. African Languages and Digital Media: Practice, Challenges, and Perspectives in Cameroon. In Joseph Cru (Ed.) *Linguapax Review 2016: Digital Media and Language Revitalization*. 33-54. Linguapax International, Barcelona, Spain.

[10] Ryan Henke and Andrea L. Berez-Kroeker. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation & Conservation* 10, (Dec. 2016), 411-457. URI: http://hdl.handle.net/10125/24714.

[11] David Lee. 2011. Micro-blogging in a mother tongue on Twitter. (April 2011). Retrieved June 16, 2021 from http://news.bbc.co.uk/2/hi/programmes/click_online/9450488.stm.

[12] Brook Danielle Lillehaugen. 2016. Why write in a language that (almost) no one can read? Twitter and the development of written literature. *Language Documentation & Conservation* 10, (Sept. 2016), 356-393. DOI: http://hdl.handle.net/10125/24702.

[13] Brook Danielle Lillehaugen. 2019. Tweeting in Zapotec: Social Media as a Tool for Language Activists. In Jennifer Carolina Gómez Menjívar and Gloria E. Chacón, (eds.) *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*. University of Arizona Press, Tucson, Arizona.

[14] Ana Roeschley and Jeonghyun Kim. 2019. "Something that feels like a community": the role of personal stories in building community-based participatory archives. *Archival Science* 19. (March 2019), 27-49. DOI: 10.1007/s10502-019-09302-2.

[15] Christina Wasson. 2021. Participatory Design of Language and Culture Archives. *Oxford Research Encyclopedia of Anthropology*, (Mar. 2021). DOI: https://doi.org/10.1093/acrefore/9780190854584.013.234

Linguistic Analysis, Ethical Practice, and Quality Assurance in Anonymizing Recordings of Spoken Language for Deposit in Digital Archives

Diana Sofia Ovalle Lopez
Department of Spanish Immersion
Fremont Christian School
Fremont, USA
slopez@fremontchristian.org

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

***Abstract*— This report considers linguistic analyses as matters of ethical practice and quality assurance in the anonymization of recordings of spoken language for deposit in a digital language archive. Ethically, researchers must be committed to protecting the identities of primary data providers. Accordingly, conducting pragmatic analyses before initiating technical anonymization procedures can aid in determining exactly what discourse, in what contexts, might constitute identifying information in need of anonymization. Qualitatively, one of the main goals of language documentation is to preserve as much primary data as possible for future research. Accordingly, conducting phonotactic analyses with the help of computer software can aid in determining precise chronometer readings for each tonal insertion to excise as little primary data as possible during anonymizations. These findings warrant further research on anonymization protocols in digital language archive projects.**

Keywords*— *language documentation, digital language archives, spoken language corpora, anonymization practices, linguistic analysis, pragmatics, phonotactics, research ethics, quality assurance, Spanish, DARDOSIPCAT

I. Introduction and Statement of Problem

Language documentation is a growing field of study that continues to evolve with the advancement of technology. As in most research with human subjects, participant identities must be protected. Typically, participant names are usually left out of written publications. However, in the world of digital language archives, language data can be found in the form of audio recordings in which, potentially, the uniqueness of participants' own voices or their ways of speaking could identify them. Given this potential, in practice, audio recordings on deposit in digital language archives can never truly be completely anonymous. Nevertheless, to protect the privacy of research participants, reasonable efforts can and should be made to minimize identifying information in such recordings. In many cases, names and any other identifying information may be "bleeped" out to protect an individual's identity. This anonymization practice involves replacing spoken language in the soundwave with an audible tone. As a matter of professional ethics, the anonymization of audio recordings was one of the quality assurance steps taken in the development of the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT, a language documentation project that aims to preserve and disseminate spoken language corpora of Spanish from The Països Catalans. This report addresses linguistic analyses involved in DARDOSIPCAT anonymization practices. Pragmatic, phonetic, and phonological analyses were crucial in developing principled anonymization practices. These practices involved (a) determining exactly what could be potentially identifying information and (b) separating coarticulated sounds across word boundaries.

II. Review of Relevant Literature

Following [1], while the act of collecting data should be seen as distinct from that of analyzing it, language description can indeed inform language documentation. In this sense, "descriptive techniques are part of a broad set of techniques applied in compiling and presenting a useful and representative corpus of primary documents of the linguistic practices found in a given speech community" [1, p. 2]. Accordingly, linguistic analyses can be a key component of language documentation. In DARDOSIPCAT, such analyses have been nothing less than necessary to ethically archive anonymized access copies of the primary data collected.

Moreover, the process of anonymization is, by nature, an editing process that may compromise the accountability of the work in question, leading to problems of interpretation. According to [2, p. 563], "heavy editing of recorded materials may give an artificial or even misleading impression of the original linguistic event." Therefore, as a matter of quality assurance, the process of anonymization must be meticulous yet considered in order to preserve as much primary data as

This research was partially funded by Western Michigan University.

possible. Among others, [3] and [4] have both pointed out the importance of distributed and redundant collaboration in this regard. It would be exceedingly difficult for one researcher alone to carry out all the tasks involved in anonymizing the individual recordings of multiple corpora on deposit in a given documentation project. Because we need to account for human error, best practice is for multiple individuals to revise the work that others have done.

III. Methodology

Our work with digital recordings on deposit in DARDOSIPCAT mainly concerned interviews that were originally recorded on analog cassette tape in Barcelona, Spain in 1995. In 2015, a previous research assistant (henceforth RA) digitized the tapes, creating digital audio files in AIFF format. In 2018, another RA listened carefully to the AIFF files, cataloging potentially identifying information on an Excel spreadsheet that included approximate chronometer readings for each stretch of discourse that might possibly contribute to the identification of research participants.

The anonymization process started by using the suggested chronometer readings to isolate each individual stretch of potentially identifying discourse for each audio recording using Audacity software. At times the approximate chronometer readings turned out to be spot-on and no fine tuning was necessary. In most cases, however, the approximate readings needed further specification to be precise. Determining accurate timing involved both an ear for detail and subsequent visual analysis of soundwaves to precisely identify the starting points and endpoints of identifying information. When human hearing alone was not reliable enough to discern an acceptable split point for a diphthong, for example, use of Audacity, Praat, and ExpressScribe software facilitated finding the most accurate starting points, endpoints, and volumes for the tones to be inserted. This process required careful linguistic analysis, which RAs carried out with the help of the Principal Investigator (henceforth PI) and the software mentioned above.

During the anonymization process, we found that some audio recordings contained potentially identifying information that had not been initially included on the Excel spreadsheet. As well, the PI determined that some of the stretches of potentially identifying information initially included on the Excel spreadsheet did not actually correspond to identifying information. For example, in one instance, the stretch of discourse “Hola, hola” was initially mistaken for “Hola, Laura”; such entries were removed from the spreadsheet.

For the purposes of quality assurance and research ethics, the PI determined that, before depositing anonymized access recordings in the archive, an RA should review each audio recording a second time to search for additional potentially identifying information that might previously have been missed. Thus, the phase of the anonymization process during which we cataloged potentially identifying information was recursive.

This added attention to detail was intended as a measure to help safeguard the anonymity of individuals whose spoken language is on deposit in DARDOSIPCAT. Nevertheless, because one must be very focused while listening to each audio recording in order to “catch” any potentially identifying information, names in particular, one may begin “fishing” for names where there are none as in the example above. Any human error in this regard that were to lead to unnecessary bleeping, though well-intentioned, could hinder the authenticity or richness of the primary data.

A. *Pragmatic Analysis*

One of the gaps in language documentary literature concerns best practice recommendations for exactly what (and how much) to anonymize. In DARDOSIPCAT, we turned to pragmatic analyses to determine whether or not certain information was identifying. As more identifying information was discovered, the PI established anonymization policies for certain cases.

One of these policies concerned the names of places in which the speakers and their parents had been raised. We determined that, with the gender and age of each speaker given in resource metadata, if users of the access recordings were to learn from the recordings themselves that speakers and their families were from particular places outside Barcelona, speaker identities could potentially be ascertained. Consequently, our best practice policy was to anonymize the name of any place of speaker or family origin that was not located within the Barcelona metropolitan area. Importantly, these same place names were fine to leave un-anonymized when mentioned in discourse contexts other than those of speaker/family origin.

The second anonymization policy concerned the speakers' majors and the universities they attended. Again, because resource metadata include the gender and age of each speaker, if users of the access recordings were to learn from the recordings themselves both the university that speakers attended and the major they pursued there, speaker identities could potentially be ascertained. Sometimes, there was mention of just the major but not the university; however, because in Barcelona in 1995 some majors were offered at only one university, the PI determined that in such cases the mention of the major should be anonymized. Subsequently, it became policy to always anonymize the major if the university was previously mentioned or if that specific major was only offered at one university. When the major was not specific to one university, we decided not to anonymize the major, but rather the university. Although at times we questioned whether anybody would ever purposely analyze such information just to identify a participant, we determined that our ethical duty as researchers requires we do everything we can reasonably do to protect the anonymity of the participants.

B. *Phonological and Phonetic Analysis*

Isolating identifying information sometimes proved difficult in context due to the formation of diphthongs and synalephas across word boundaries in Spanish. In such cases, the anonymization process required further phonological and phonetic analysis for quality control. Such analysis was often required when the suggested chronometer readings included a word ending in a vowel before the utterance to be anonymized. For example, one of the anonymizations was for the name “Elizabeth”, and the chronometer readings included the stretch of discourse “La Elizabeth” ‘The Elizabeth’. In order to anonymize the name but leave the article, chronometer readings had to be set to insert a tone exactly where the /a/ in “La” ends, but before the beginning of the /e/ in Elizabeth.

Dealing with consonantal coarticulations and nasalized vowels was another challenge we encountered when isolating identifying information. For example, in the phrase “en Molins de Rei” ‘in Molins de Rei’, as the hometown of one of the speakers' parents, “Molins de Rei” represented potentially identifying information. Due to an obligatory process of nasal assimilation in Spanish, the initial nasal consonant was bilabial before the nasal stop. Moreover, the articulation of the vowel was nasalized before the initial nasal consonant.

In complex cases like those described above, we used Praat to visually analyze the spectrogram of the contextualized audio fragment. In the case of “La Elizabeth”, via Pratt we were able to discern the speech formants, which represent concentrations of energy based on frequency. In Spanish, the second formant represents the highest amplitude that a soundwave reaches, and each vowel reaches a different amplitude in the wave. Looking at the spectrogram of “La Elizabeth”, we were able to see the rising of the second formant from [a] to [e] and thus identify the precise chronometer reading at which to insert the anonymization tone. In the case of “en Molins the Rei”, we used Praat to determine the onset of nasality in this sequence based on acoustical measures.

IV. Findings and Significance

One finding of our research is that pragmatic analyses are needed to discern potentially identifying information in a contextually-appropriate manner. This finding is significant to the accountability of documentary work; without such analyses, researchers have no principled way of knowing the extent of potentially identifying information that a recording may include. Accordingly, pragmatic analyses of spoken language corpora should be a prerequisite to the insertion of anonymization tones in digital recordings on deposit in language archives. Once such analyses have been completed, researchers can implement principled anonymization policies uniformly throughout the documentation to deal with context-dependent identifying information that might otherwise remain undetected.

Another finding of our research was that phonotactic analyses conducted with the help of computer software are needed to accurately isolate potentially identifying information in the phonetic phrase for later anonymization. Given that Spanish syllabifies discourse irrespective of word boundaries, without computer-mediated phonotactic analyses, researchers have no principled way of determining precise chronometric readings at which to begin and end anonymization tones in digital recordings of spoken Spanish. This finding is significant to documentary accountability insofar as it improves quality in anonymization processes. Once such analyses have been completed, researchers can be assured of high-quality tonal insertions in the development of anonymized language resources.

V. Conclusion

Language documentation is a science composed of various types of linguistic analysis. This report has described practices and protocols utilized in DARDOSIPCAT to demonstrate how and why linguistic analyses may be useful in the anonymization of spoken language resources for deposit in digital language archives. Given its goal of preserving as much primary data as possible for future research, the anonymization process is arduous, meticulous, and iterative. Accordingly, we have discussed pragmatic and phonotactic analyses as matters of both professional ethics and quality assurance.

As technology continues to advance and digital language archives grow richer in content, it is important for everyone involved in language documentation to stay committed to protecting participant identities. Because language documentation is an ongoing, ever-evolving process best achieved in continued collaboration, further research on anonymization procedures is warranted. Such research could improve ways of maintaining high-quality recordings of spoken language while also protecting the privacy of primary data providers.

References

[1] N. P. Himmelmann, "Documentary and descriptive linguistics," Linguistics, vol. 36, no. 1, pp. 161-195, 1998, doi:10.1515/ling.1998.36.1.161.

[2] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557-582, 2003, doi: 10.1353/lan.2003.0149.

[3] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1-30.

[4] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159-186.