Author/s:
Thieberger, N;Harris, A

Title:
Be Not Like the Wind: Access to Language and Music Records, Next Steps

Date:
2020

Citation:
Thieberger, N. & Harris, A. (2020). Be Not Like the Wind: Access to Language and Music Records, Next Steps. Proceedings of the Language Technologies for All (LT4All), pp.101-103. European Language Resources Association (ELRA).

Persistent Link:
https://hdl.handle.net/11343/265940

Proceedings of the Language Technologies for All (LT4All) , pages 101–103
Paris, UNESCO Headquarters, 5-6 December, 2019.

Be Not Like the Wind: Access to Language and Music Records, Next Steps

Nick Thieberger, Amanda Harris
University of Melbourne, University of Sydney
School of Languages and Linguistics, The University of Melbourne, Parkville, VIC 3010, Australia
PARADISEC, Sydney Conservatorium of Music, C41, University of Sydney, NSW 2006, Australia
thien@unimelb.org.au, amanda.harris@sydney.edu.au

Abstract

Language archives play an important role in keeping records of the world's languages safe. Accessible audio recordings held in archives can be used by speakers of small and endangered languages, and their communities, and provide a base for further research and documentation. There is an urgent need for historical analog tape recordings to be located and digitised, as they will soon be unplayable. PARADISEC holds records in 1228 languages. We run training for language documentation and are developing technologies to localise access to language records. A concerted effort is needed to support language archives and sustain language diversity.

Keywords: archives, language diversity, PARADISEC

Em i no olsem win – Painim tok peles na musik rekod
Wanpela kain ples olsem akaiv i save lukautim ol rekod or pepa bilong ol kain kain tok ples bai stap gut long bihain taim. Planti ol liklik tok ples ol klostu dai nau. Tok ples bilong ol manmeri na komuniti mas usim akaiv long helpim wok bilong painim aut moa na raitim ol pepa bilong ol tok ples. PARADISEC i gat 1228 tok ples. Mipela painim ol rikoding long taim tumbuna we ol tok ples i stap long keset tep bilong dijitaisim nau o sapos nogat bai ol bagarap. Taim nau long sapotim ol tok ples akaiv long lukautim planti kain kain tok ples.

1. Introduction

Me, selwan ag kupi em̃ae, tiawi itraus traus traus traus, natrauswen ga itaos nlag. Itrausi pan kaipa. Me komam uta laap kin uto mau, a? Malen umat, inom.

But when you are far away [and can't record him] the old man can talk and talk and talk, his story is like the wind. He tells it and it is gone. But there aren't many of us left. When we die, it will be finished.

†Kalfap̃un Mailei, 1998, Erakor Village, Efate, Vanuatu

In Melbourne recently a speaker of a language from Papua New Guinea looked through PARADISEC's webpage and we searched for the name of his language there. He was amazed to find recordings of his grandfather, never having expected to find anything in his language at all. For most of the world's small languages there is little or nothing available on the web, with most records, if they exist, still in analog form. In the passage quoted above, Kalfap̃un Mailei, an elderly man in 1998, was conscious of the need to record oral tradition so that it is available for others to hear in the future, and not, like the wind, here now and then gone. Language archives give us a glimpse of the richness of oral tradition, while they can never be a complete view of a language, these records of performances nevertheless provide both a cultural treasure for the speakers and their communities, and a research base for study of the world's languages.

Archives typically hold outputs of fieldwork, and so can have many hours of recordings for a language, which are often the only known recordings for that language. Of the 7,000 languages spoken in the world today, there are records of only a small proportion (Thieberger, 2016). Records of endangered languages that are unlikely to be spoken by a next generation of speakers, or have ceased to be spoken at all, are particularly valuable as they may be the only recorded source of information about that language. And getting good records back to this community can also help to strengthen the language, assisting in relearning older styles or performances. It also allows current speakers to enrich archive catalogs with their memory of what is recorded and its place in their society. There is a need for a concerted effort to index what is known for each language (see Thieberger 2106), as will be discussed below. An important first step is to locate and digitise all analog recordings, as we know that analog tape will not survive for much longer.

2. Language and music archives

The Open Languages Archive Community (OLAC) lists 60 language and music archives (http://www.language-archives.org/archives.php). 15 of these are represented by the Digital Endangered Languages and Musics Archives Network (DELAMAN). Each archive typically represents a particular geographical region. Community-agreed metadata standards enable harvesting by services like OLAC, which then creates a single page for each language that has a language code (ISO-639-3) thereby increasing the discoverability of language and music collections. Ideally, more fine-grained codes would be incorporated into such searches, for example, glottolog,[1] and more detailed regional codes, where they exist (like Austlang[2] in Australia).

Language archives provide this information as well as: licensing content; making content available; digitising analog materials; quality assurance; promotion of collections to source communities; conversion to archival

[1] https://glottolog.org

[2] https://collection.aiatsis.gov.au/austlang

and delivery formats; enforcing minimal descriptions of the items; providing citable forms of primary data.

The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) was established in 2003 as a collaboration between linguists and musicologists at the University of Sydney, the University of Melbourne, and the Australian National University (Thieberger and Barwick, 2012). Critically, it was built by researchers who saw a need for a discipline-specific archive to deal with an inherited backlog of audio recordings that had no other prospect of being digitised.

3. What PARADISEC has done

The collection currently holds:
1228 languages
529 collections
26,395 items
287,511 files
7,397 users
11,500 hours
(At 8 November 2019)

Our catalog provides feeds that are harvested by external services, like the Open Language Archives Community, which increases the findability of an item in our collection. This means that even the most remote user who has internet access can find records.

We have received various awards and recognition for our work. In 2013, PARADISEC was added to the UNESCO Australian Memory of the World Register. In 2019 we received the World Data System[3] data seal, signifying we conform to all necessary standards.

Our initial motivation was preserving heritage recordings, and that continues to be an important part of our daily work. However, having built a relatively simple system for accessioning new items and collections, we are now also receiving numerous digital collections, some deposited in the course of fieldwork or soon after recordings were made. We are keen to help current fieldworkers to adopt methods that give them greater access to their own recordings, and, at the same time, make their collections 'archive-ready', reducing the amount of work required for their accession into our archive.

This leads to a focus on training in new methods so that the process of recording, transcription and annotation of transcripts all result in records that can be reused later on. We train academics and we train community members to do their own recording. We have encountered examples of recordings made on poor equipment, or where the microphone was too far from the speaker, so that little is audible. Low resolution recordings can be difficult to use for other purposes, like phonetic analysis or in creating teaching materials for the language. Transcripts made on paper can't be searched on a computer, and transcripts typed in a word-processor don't have timecodes that link back to the media. Current tools for transcription[4] insert timecodes for each chunk of the transcript and this means that this chunk can be played immediately and so allow you to cite down to the level of a word or sentence, strengthening research practice, and all the more so if the media is stored in a public online archive, like PARADISEC, and the media can be played directly from there. Articles referencing a story or a sentence can include a link for readers to follow to hear that item.

There is only a narrow window of time before analog tapes arising from historical field research become unplayable, both because they are on fragile media, and because of the increasing scarcity of playback machines. We are part of an international network of archives that is running an ongoing survey, called "Lost & Found" which asks for information about tape collections that need to be digitised. As a result of responses to this survey we have digitised fifteen collections of tapes. For example, we arranged for a collection of six hundred tapes from Madang in Papua New Guinea held at the Basel Museum (Switzerland) to be sent to our colleagues in the Netherlands for digitisation. A small collection of eight tapes in Yonggom (Papua New Guinea) were sent to our sister archive in the USA who digitised the tapes and sent the files to us to accession. Similarly, we arranged for a collection of 44 tapes in the Wampar (Papua New Guinea) language, recorded in the years between 1958 and 1972 and held on cassettes in Switzerland, to be digitised by another archive in London who then sent us the files.

While we work within the academy, we recognise that many of our colleagues do not take seriously the need to create lasting records in the course of their fieldwork, evidenced by the number of academic works published about languages over the past 30 years and the lack of archived records for those languages.[5] Accordingly, we make an effort to run regular training workshops and to advocate for the adoption of new methods that will increase the archivability of primary research data.

Further, there is an increasing amount of documentation being produced by speakers, some intentional, and some incidental to using social media. Both kinds of recordings risk being lost if there is nowhere for them to be housed, but social media is especially difficult to capture without a concerted effort. It is beyond our current ability to capture this, but it would be useful to have an automated service to recognise non-mainstream languages in social media, and then harvest that material into an archive.

4. Return of archival files

We have built a catalog that makes it relatively easy for material to be found in PARADISEC, assuming an internet connection and literacy in English. The normal kinds of search terms are provided: language, country, person, role, data, geography. To get the files to people with little or no internet access we have explored ways of sending copies of archival collections to source communities or nearby regional centres. An obvious way of doing this is to send all items for a given language or place on a hard disk to the local cultural centre or museum. This can work well, but also requires a catalog of those files to be created so that the contextual information in the catalog can be seen

[3] https://www.icsu-wds.org

[4] http://www.dynamicsoflanguage.edu.au/research/resources-for-linguistic-tools/

[5] see the analysis here: http://www.paradisec.org.au/blog/2016/07/finding-what-is-not-there/

together with the files. PARADISEC has a system that writes a text file (in XML) to the items in the collection each time the catalog entry is saved In this way, each item or set of files is self-describing so we can aggregate all of these text files for any given set of items and create a simple (html) catalog of just that collection.

But what about those places that don't have computers and so can't use a hard disk? We have built local wifi transmitters with hard disks that can be used in this situation[6]. The wifi transmitter is called a Raspberry Pi and costs less than AUD$100. It can be set up to transmit within a small area and so make this material accessible on mobile phones or tablets.

We are currently developing an OCFL[7] based version of our collection and that uses a similar principle of including the metadata with the object exported as RO-Crates[8]. The combination of RO-Crate described items stored in a standardised format could be a means to stepping into a modern, scaleable catalog application able to support many communities and many terabytes of data.

5. Collaboration with regional agencies to digitise their tapes

PARADISEC has established working relationships with agencies in our region like the Vanuatu Cultural Centre, Institute of Papua New Guinea Studies, University of French Polynesia, and the University of New Caledonia, among others. In 2014 we received funding to digitise hundreds of tapes held by the Solomon Islands Museum in Honiara. As we continue to run workshops in the Pacific region on issues around language recording methods, transcription, and data management, we continue to be offered tapes to digitise from local language authorities who no longer have the means to play tapes they created in the past. In November 2019 we collaborated in a seminar at the University of French Polynesia, the third we have run over the past few years. Present at the seminar were representatives of the various local language academies: Tahitian, Pa'umotu, and Marquesan. Earlier, we had worked to digitise 50 cassettes with some of these agencies, and at this event we received 19 cassettes, 6 minidisks, and 10 compact disks, entrusted to us to digitise and return. There are many such collections that we are yet to find, but each requires a relationship of trust with the owners, to let us take such important material to our offices where it can be properly assessed, cleaned, and digitised. In part, our longevity provides the status that allows us to build deeper relationships with these agencies and with the owners of other collections.

6. Conclusion

Without our work at PARADISEC over 11,500 hours of audio records would not have been preserved. However, we know there are many more records that have not yet been located by collecting agencies, and that there are not enough language and music archives in the world to deal with the quantity of material that has yet to be discovered. There needs to be a concerted effort to support new archives so that, in the words of Kalfaƀun Mailei, these words will not be like the wind, but will continue to enrich cultural practices, and help to show the value of the world's many small languages.

7. Acknowledgements

Thanks to Steven Gagau for translating the abstract for this paper into Tok Pisin. PARADISEC acknowledges funding support from the Endangered Archives Programme grant 693 (Preservation of Solomon Islands analogue recordings), Australian Research Council (ARC) LIEF program (2003, 2004, 2006, 2011), ARC Centre of Excellence for the Dynamics of Language, ARC Future Fellowship FT140100214.

8. Bibliographical References

Nick Thieberger. 2016. What remains to be done – Exposing invisible collections in the other 7000 languages and why it is a DH enterprise. In *Digital Scholarship in the Humanities*. 32(2), 1:423–434 . http://dx.doi.org/10.1093/llc/fqw006

Nick Thieberger and Linda Barwick. 2012. Keeping records of language diversity in Melanesia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), pp. 239-253 in Nicholas Evans and Marian Klamer (Eds). *Melanesian languages on the edge of Asia: Challenges for the 21st Century.* [LD&C Special Publication No. 5]. Honolulu: University of Hawai'i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/4567

[6] Discussed further here: http://www.paradisec.org.au/blog/2018/07/local-wifi-versions-of-paradisec/

[7] https://ocfl.io/)

[8] https://researchobject.github.io/ro-crate/