Access to recordings in the languages of the Pacific

Nick Thieberger*

* School of Languages and Linguistics, University of Melbourne, Australia.
thien@unimelb.edu.au
Abstract

With over a quarter of the world’s languages the Pacific is a particularly good place to focus on how language records can be made accessible. The creation and description of research records has not always been a priority for humanities academics and any records that are created have typically not been provided with good archival solutions. This is despite these records often being of cultural or historical relevance beyond academia.

Many academic researchers at the end of their careers despair at the task of making sense of a lifetime’s output of papers, notes, images, and recordings. Our project, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a collaboration between the University of Sydney, University of Melbourne, and the ANU, began in 2003 by digitising analogue tape collections and providing sufficient metadata to make them discoverable. These tapes belonged to retired or deceased researchers and would otherwise have been stored in a house or maybe a library, but in both cases are difficult to find and more difficult to access.

In this paper I outline how PARADISEC works and how to find information in it. I will show how we provide access to the collections we hold and how that has helped build links with people and agencies in the Pacific. We have partnered with a number of museums and cultural centres to digitise analogue tapes and are working on ways of getting information about the collection to the source communities so that they can find recordings made by their members in the past.

Introduction

With over a quarter of the world’s languages, the world’s largest ocean, the Pacific, is a particularly good place to focus on how language records can be made accessible. The creation and description of research records has not always been a priority for humanities academics, and any records that are created have typically not been provided with good archival solutions. This is despite these records often being of cultural or historical relevance beyond academia. In this paper I suggest there is a basic requirement for accessible primary language records to serve both community and research needs.

Many academic researchers at the end of their careers despair at the task of making sense of a lifetime’s output of papers, notes, images, and recordings. Our project, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), began in 2003 by digitising analogue tape collections and providing sufficient metadata to make them discoverable. These tapes belonged to retired or deceased researchers and would otherwise have been stored in a house or maybe a library, but in both cases are difficult to find and more difficult to access.
We provide access to the collections which has helped build links with people and agencies in the Pacific. We have worked with a number of museums and cultural centres to digitise analogue tapes. We are working on ways of getting information about the collection to the source communities, typically villages in the Pacific, PNG, or South-East Asia, so that they can find recordings made by their members in the past.

The Pacific Language Wonderland

A quarter of the world’s 7,000 languages are spoken in the Pacific (Lynch 2014), representing huge linguistic diversity in this region. For linguists, this displays a range of possibilities for examining how languages diverge from each other and what changes occur within languages over sometimes relatively shallow time depths of perhaps the past fifty years. We know, for example, that Polynesian settlement of some parts of the Pacific, including Aotearoa/New Zealand, is less than a thousand years old (Kirch 2000). Here I will focus on the records of these languages, and, in particular, on recordings (audio records created in the past century to capture performance of various kinds). In Australia, recordings of Pacific languages are typically found in the hands of researchers (like linguists, anthropologists or musicologists), or in their deceased estates, ending up in museums, libraries or archives which often deal with tapes as static objects to put on shelves rather than as dynamic media to be accessed. Finding these records can be difficult and require word-of-mouth knowledge if they are located in homes or university storerooms. For people in the widely dispersed islands of the Pacific it is even more difficult to locate analogue records and to play recordings, so we have prepared digital files with catalogue information and clear deposit conditions to facilitate access.

The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) provides the means to curate collections of these recordings and has built systems that can ingest metadata and primary records and then expose them for access by registered users of the catalogue. PARADISEC began in 2003 after a group of linguists and musicologists recognised the need to preserve many priceless analogue recordings that were at risk of being lost. We had great help from the National Film and Sound Archive and the National Library of Australia in determining what metadata and equipment standards were required, and we applied for Australian Research Council funding for a year to build research infrastructure.

With that, we started building PARADISEC, a completely digital archive that is a collaboration between the University of Sydney, University of Melbourne, and the ANU. We are grateful to Australian national storage and network programs like AARNet, RDSI, ANDS, and Nectar, that have supported our work over time. In 18 years the archive has grown to 86 terabytes, and has material from 1,236 languages, now including over 12,500 hours of audio recordings. While the initial focus was on the region around Australia, the collection now holds material from anywhere in the world. It is a significant collection that has been listed on the UNESCO Australian Memory of the World Program’s Register (AMW 2013).

Building PARADISEC

To begin with, we surveyed linguists, initially at the ANU in Canberra, and from that work identified hundreds of hours of recordings made since the 1950s in PNG, Indonesia, other parts of Melanesia, and further afield. We negotiated with the collectors, or with the executors of their wills, to digitise the tapes, and we developed a deposit form in which they can license how the recordings could be used. We also developed access conditions that are agreed to by users of the collection. We employed an audio technician who built the system to digitise the tapes, with suitable analogue-digital converters and using the Quadriga system to create archival Broadcast Wave Format (BWF) files.

As digital language archives started to operate in the early 2000s, the Open Language Archives Community (OLAC) 2011) proposed a metadata set to be used by language archives. Based on the Open Archives Initiative and Dublin Core, this minimal set of metadata terms was designed to allow interoperability of metadata so that OLAC could aggregate metadata from all archives
and present it in a unified format. PARADISEC has been an OLAC-compliant archive since 2004
and a great benefit of this is that all public catalogue items in PARADISEC are harvested every
day by OLAC and made available via their system to other services, maximising the reach of our
catalogue. OLAC provides a page for every language in the world, listing everything that is in
each archive for that language (see OLAC 2011).

PARADISEC is one of several digital archives with a focus on the Pacific that are also part
of OLAC. Three others are: the University of Hawai`i’s Pacific collection (discussed further
below), Pangloss at the CNRS/LACITO in Paris, (CNRS/LACITO n.d.), and CoCoOn, also in Paris
(CoCoON n.d.).

We wrote a database to catalogue the PARADISEC collections that uses controlled vocabularies
(for languages, countries, roles of participants and so on) and standardised entries via popup
menus, populated from existing terms (for people names). The catalogue entry for each item
enforces the deposit conditions provided by a depositor who is able to assign access rights to
particular individuals, even if a collection is otherwise closed. There is provision in our catalogue
to mark items as being ‘private’, so they are not published or visible to anyone except the
depositor, which is useful for the period in which a collection is being developed.

We based our metadata set on the OLAC recommendations and provided for export of a
snippet of XML (customised markup language) from the catalogue to be included inside the
wrapper of each wave audio file in order to make a BWF (broadcast wave format) file that
identifies it and its provenance in our collection. A key feature of modern digital language
archives is the use of a standard language identifier, ISO-639-3, which provides three-letter
codes for each language in the world, avoiding the problem of using language names or
descriptors which can vary in spelling. This three-letter code underpins the services that OLAC
provides as it aggregates records from participating archives.

OLAC is one service that uses PARADISEC’s catalogue feeds. Our catalogue’s APIs (application
program interface) (http://catalog.paradise.org.au/apidoc) export collection-level (RIf-CS)
and item-level (OAI-PMH) metadata that is also picked up by Research Data Australia (https://
researchdata.edu.au/) and the National Library of Australia’s TROVE, as well as Google. All of
this provides the incentive to researchers that their work will become locatable and accessible,
and above all, citable, through international networks.

However, even if an item is in a collecting institution, it is unlikely to have its content language
noted in the catalogue in a standard format. In Australia, the National Library of Australia and
most State Libraries do not use ISO-639-3 in their catalogues, so it remains difficult to find items
in these collections for anyone searching for language records. As an example of how this can
be done, the University of Hawai`i Library’s Pacific Collection has recently had its catalogue
updated to include language identifiers that are now findable via OLAC (see Kleiber 2015).

We use the PARADISEC catalogue to index existing resources that have no language identifiers.
This allows the OLAC system to harvest that information from our catalogue, and to include it
in their online database of resources for each language. A website may include excellent
language material—perhaps a transcript of missionary records, or a dictionary—but it is not
registered in OLAC’s directory. Because websites are at risk of being moved or lost, we find the
version of the same site that is hosted in the Internet Archive and point to this preserved version
from our catalogue, ensuring persistence of the link. So, for example, there is a wonderful site
of information about Rotuma (http://www.rotuma.net/) that includes typed versions of early
sources in the language, so we have included a link to the persistent form of this site
(http://catalog.paradise.org.au/collections/External/items/rotuma) and this means it is also now found in the OLAC page for Rotuman (http://www.
language-archives.org/language/rtm).
The materials we deal with have personal connections to the communities they were recorded in, being stories and songs recorded with a range of people, from the 1950s through to the present. They always represent a time slice of performance, recording how people spoke at a particular time, and sometimes recording stories or songs that are no longer known today. A primary motivation for our work is ensuring that the people recorded, or their families, can access these pieces of their cultural heritage.

A Race Against Time

Analogue recordings are soon going to become unplayable due to the failure of the tapes themselves or due to the lack of playback equipment. It is now a race against time to find tapes that need to be preserved, tapes that contain unique cultural information. We set up an online questionnaire (DELMAN n.d.) that we publicise every year asking anthropologists, linguists, and musicologists about tapes they own or know about, that need to be digitised. As a result, we have been able to digitise significant collections. For example, we recently arranged for a collection of several hundred tapes held at the Basel Museum, Switzerland, to be sent to our colleagues at the language archive in Nijmegen, Netherlands, for digitisation. A small collection of eight tapes in Yonggom, Papua New Guinea, were sent to our colleagues in Texas, USA, who digitised the tapes and sent the files to us to accession into the collection. Similarly, we arranged for a collection of 44 tapes from Papua New Guinea in the Wampar language, recorded in the years between 1958 and 1972 and held on cassettes in Switzerland, to be digitised by colleagues in London, UK, who then sent a hard disk to us for accession. This demonstrates the kind of productive international collaboration we have developed between language archives. The need for this work has also been recognised by funding agencies like the Endangered Archives Programme (British Library n.d.) and the Endangered Languages Documentation Programme’s Legacy Materials Grants (ELDP 2020).

As many of the tapes we locate have little metadata beyond what is written on their cases, we enter what we can into the catalogue (typically country, collector, date, and maybe language or village) and invite anyone with more information to send it to us. In 2016 we were funded to enrich our Papua New Guinea collection metadata and employed Steven Gagau, a PNG national to work on the collection. He was able to add to the descriptions by listening to the recordings (not something we generally have the time to do) and by asking his networks for more information. In 2017 we worked with the Divine Word University in Madang, Papua New Guinea, who had a project of playing audio from our collection at a local market stall and asking people there to add what they could about the recordings. They sent us a spreadsheet of new metadata that we added to the catalogue. We hope to increase this kind of activity in future.

In order to allow feedback from the level of a person in a small village or town about a language in our collection, the recording needs to be delivered in a way that can be interpreted locally. The recording is downsampled (a digital audio signal made smaller) as part of our ingestion process (wav files become mp3, and video is transcoded to low-resolution mp4), so that it can be downloaded over low-bandwidth connections. For places with limited or no internet connection we are exploring methods for creating sub-collections that are self-describing sets of files. Each time a PARADISEC catalogue entry is saved, it writes an XML file to the item’s directory, thus keeping all contextual metadata together with the primary records. These XML files can then be collected together to write a catalogue of all items in the selected sub-collection. The catalogue can be supplied together with a hard disk of files, or, where a hard disk is not appropriate, for example where computers are not commonly available, on a local wifi transmitter. This means that any device that can receive wifi can be used to access records transmitted from the wifi transmitter.

Implications for Current Practice

Because we have seen so many recordings in deceased estates, we have developed suggestions for current researchers in ways to create their documentation. In particular, we encourage all researchers to make a will and to appoint a literary executor, so that it is as clear as possible
what is to be done with the legacy of records created during their career. Depositing research records as soon as they are created provides a citable form with persistent identification and location, and can be added to by transcripts that can be improved over time. It is now not uncommon to have recordings deposited by fieldworkers from their field location, or as soon as they return from the field, contradicting the previous notion of archiving at the end of a research career (see Barwick 2004).

Citability, reproducibility, and verifiability of claims made in research papers have long been critical issues in scientific research and are becoming increasingly important in the humanities (see Berez-Kroeker et al. 2018). A repository provides a citable form of primary records, which is a critical contribution to research practice (see Thieberger 2016). If a researcher makes a claim about the occurrence of a sentence or construction, performance style, intonation, and so on, they must be able to point to where in their corpus of recorded material it occurred. It is not sufficient to say that some phenomenon has simply been observed, as this leaves open the possibility of examples being constructed to suit the analysis.

A Museum of Language?

In 2016, we were approached by the Canberra Museum and Gallery to prepare a display about PARADISEC for an exhibition on UNESCO’s Memory of the World Programme (CMAG 2016, p. 3). We built an augmented reality poster that allows users to hear stories in 14 different languages, by holding a mobile device over a poster. We also built a virtual reality (VR) display based on PARADISEC’s metadata. We ran an algorithm that took twenty second snippets of mp3 files in the collection and these files were then put into a map of the Pacific, where users could interact via goggles with the data (we call this ‘Glossopticon’) and see locations of each of the languages of the region. Further development has led to a web VR Google cardboard version that can be displayed on any mobile phone. These have had a great impact and have been reported on in the public media, both in Australia and in the Pacific (including the Fiji magazine ‘Turaga’, Islands Business Magazine, and the Air PNG Inflight magazine). They also serve to illustrate that properly constructed data and metadata can have multiple uses beyond that envisaged by the original research project.

Another example of presenting archival material in novel ways is the online exhibition of a collection that represents the estate of Arthur Capell, who was a professor of linguistics at the University of Sydney, with images of many of his records from all over the Pacific (PARADISEC 2014). The exhibition displays images with enough metadata to make them findable, but with no transcriptions, and so is a relatively easy way to make this information available.

Conclusion

Novel means of presentation of recorded cultural heritage need to be based in a collection that provides long-term curation, and long-term identification of the objects it holds. In this paper I have described the ways in which a group of linguists and musicologists have built the infrastructure necessary for us to carry out our work, bearing in mind our dual responsibilities, first to the people we work with to make records of their communities available to them in the long term, and, second, to the research we are creating, providing citable and verifiable primary records for others to build on.

References


Thieberger, N 2016, ‘What remains to be done—Exposing invisible collections in the other 7000 languages and why it is a DH enterprise’, *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 423-434.

Endnote

1 AARNet: Australia’s Academic and Research Network; ANDS: Australian National Data Service; NECTAR: National eResearch Collaboration Tools and Resources; RDSI: Research Data Storage Initiative