Chapter 3

Language and music recordings and the responsible researcher

Nick Thieberger

“Indigenous knowledges, cultures and languages, and the remnants of indigenous territories, remain as sites of struggle.”

Linda Tuhiwai Smith, *Decolonizing Methodologies* (1999)

Introduction

Archiving of music and language materials has come a long way in the past 20 years. Curated primary data in an archive, and the practices that go into creating that data, have the potential to radically change the way we conduct fieldwork, cite our data and access records made by others. The apparently simple fact of being able to cite primary data to any level (text, sentence, word, phoneme, song, stanza, line) allows verification of analyses that was not previously possible. Archived recordings also ensure that the records are available for re-use in the future, by both researchers and the broader community. The acts of making re-usable primary records in collaboration with the speakers, their consent in how the records can be shared and the creation of suitable repositories all support significant changes in methodology.

While it seems that many academics find it difficult to archive their recordings, Linda Barwick is an exemplar in having done just that, as well as creating an outstanding output of descriptive and theoretical papers (as can be seen in the list of her output provided with this volume). Here, I will take Barwick’s practice as the starting point for a discussion of the central role of archiving in language and music documentation.

Accessible primary data are part of the exchange relationship we (as researchers) enter into in the field. We make recordings initially for the purpose of academic analysis, but have, until recently, ignored their value for the community we work with. This is evident because we have either not made re-usable records, or have kept them in inaccessible locations. Making recordings available via an archive relieves us of the need to keep track of our records into the future and of the need to deal with occasional requests for copies of their records that require finding the records and then sending them to the requester. This is a positive post-colonial aspect to archiving of language materials, especially in contrast to the prevailing past practice in which no material was made publicly available.

Why then is it that records arising from fieldwork can be so hard to find, especially for the speakers and their communities? A certain amount of detective work can be required to even know that these materials exist. When the records were analogue (reels of audiotape, audiocassettes or papers), a concerted effort was required to find them, first, based on an expectation that there were recordings made to support a piece of research reported in a publication, and, second, requiring travel to a single location to consult the materials. While there are individual catalogues of collections in each particular holding institution, there is no unified catalogue of all of these kinds of collections that would provide, for example, a list of all known material in a given language or musical style. To make it even harder, the only copies of analogue tapes were held by the researcher, and access became harder still if they had retired or died. Getting access in these cases required the good graces of the new custodian (the executors of the estate).

Digital records should, in principle, be easier to locate. After all, they can be copied to various locations in a manner similar to the taper light famously observed by Thomas Jefferson: "he who lights his taper at mine, receives light without darkening me".[1] In fact, digital records are at risk, to extend the taper analogy, of being snuffed out so that no further light can be received from them, especially if they are held on a computer or hard disk as inaccessibly as the analogue files discussed earlier.

Tuzin noted:

> If it is true that one of ethnography's distinguishing features is the moral cloak with which it wraps itself, then it is all the more surprising and ironic that the record of ethnographic conduct is abysmal concerning the preservation and dissemination of its findings. (Tuzin 1995, 24)

[1] Letter to Isaac McPherson, 13 August 1813. https://tinyurl.com/376ray2n.

Sadly, despite the passing of nearly 30 years, and the rise of digital preservation strategies, Tuzin's observations largely hold true today. This chapter outlines the gains made by archiving and then assesses the costs of not archiving language records. It charts the growth of archives over the past two decades and argues for the need for new archives to serve the anticipated future uptake of archiving habits by musicologists and documentary linguists.

Grammars but not records

As an example of how much acceptance there is among researchers of archiving outputs of fieldwork, let us take the example of linguistic fieldwork and grammar writing. In earlier work (Thieberger 2017) I noted the large number of languages for which grammars have been written but for which there are no primary records archived. Some 683 grammars are listed in Glottolog[2] as appearing since the beginning of 2000. Of the languages represented by those grammars, 555 have 40 or fewer items in a digital language archive. So, even in the now 26 years since Himmelmann's 1998 framing of the field of language documentation (which focuses on the creation of primary records, re-usable by others, in addition to the analysis traditionally expected of academic research), it seems that the majority of linguists are either not creating records or, if they are, are not prepared to archive them in a relevant digital language archive. We are not alone in this; Piwowar (2011, 5) finds that, even in biological science, only "25% of studies that performed gene expression microarray experiments have deposited their raw research data in a primary public repository". She did also report an improvement in the use of shared datasets over time from "less than 5% in early years, before mature standards and repositories, to 30–35% in 2007–2009". Perhaps we can also look forward to an increase in archival deposits with a new generation of researchers, and my impression is that those working in Australia have learned the importance of archiving from role models like Barwick (e.g., Barwick 2004; Barwick and Thieberger 2006; Barwick, Green and Vaarzon-Morel 2019). Of course, it may be that language records are archived, for example, in a state library or university repository. But because these kinds of archives are not part of linguistic search mechanisms like the Open Language Archives Community (OLAC, discussed further below), the items are difficult to locate and do not feature in the metrics presented here.

[2] Glottolog is a service that lists resources available for each of the world's languages (https://glottolog.org).

Language identifiers are an internationally recognised system that avoids problems of variant language name spellings and forms. Each language will have different names in different languages (français, French, Francese, Französisch) so which one do you search for? There are languages whose names are also common words (e.g., Maria, Mono, Mum, Noone, Karen, Kola, Titan). Assigning a code to each language avoids this problem and there are two main systems of codes available: notably ISO-639–3 (the three-letter codes) and Glottolog. In order to take advantage of these codes, there needs to be a system in place that identifies what language an item is in, and publishes that information to make the item findable on the web. This is what language archives do.

As has been observed by Moyse-Faurie:

> the basic description of languages with an oral tradition . . . contributes, with the transition to the written word, to the revitalization of a language, if it is in danger, or to ensuring its documentation in the form of archives which will remain available for any other future use, whether for research in the strict sense or for applied linguistics. (Moyse-Faurie 2014, 140)

This applies equally to musical records and recognises other possible uses of archival records beyond foreseeable issues in research and cultural revitalisation, including cultural heritage and personal histories of the people recorded.

I first met Linda in Perth in the 1980s and discussions with her about fieldwork later inspired my own work on a corpus of Nafsan (South Efate) based in my research in Vanuatu in the 1990s, linked to media using SoundIndex (Thieberger 2004; Thieberger and Jacobson 2010), and the development of the online system EOPAS (Schroeter and Thieberger 2006) for presenting interlinear text and media. All of this work demonstrates the utility of creating time-aligned transcripts of field recordings from the beginning with the primary goal of creating a corpus in which transcripts are always verifiable. As the analysis of the language improves, the ability to re-listen to the source will support re-analysis, especially if it relates to prosodic aspects of the language that are only apparent because of the ability to hear linked media. Examples given as evidence of linguistic phenomena can be checked by the readers, together with the context from which they are extracted.

A major motivation for archiving, from an academic disciplinary perspective, is the ability to cite primary sources (Berez-Kroeker et al. 2018) and so to allow verification of examples and the context in which they occur, potentially leading to new analyses. Barwick's work in establishing PARADISEC helps popularise what Nick Evans (this volume) calls "dialogic repatriation", the ability to take archival records, interpret them, and enrich the archival collection, while also using them to support relearning of traditions that may have been lost. While I am mainly discussing language records here, the same issues are applicable to music records, especially where the two are combined in single archival collections. Linguists have the advantage of shared tools and infrastructure and a consensus about metadata terms, largely provided by two sources, OLAC (mentioned earlier) and The Language Archive (TLA).[3] This is the basis for the kinds of search mechanisms discussed in this chapter.

Archiving language material, the successes

Online digital archives have the potential to be dynamic centres of activity (Holton 2012), and to support revitalisation of heritage knowledge of performance and language by providing a link between generations in places where little else was recorded in the past. In many cases, the archival recording is the only online reflection of performances from a village or family member. Success for an archive can be measured in several ways, but a critical success is the number of times people unexpectedly find recordings related to their families or languages. Another is the re-use of archival recordings for new research purposes. In both cases, access to the primary records can result in enriching existing materials with transcriptions or additional metadata. Of course, another side to the ubiquity of digital material is the need to ensure appropriate access, with licences in place that indicate how material can be used, as specified by the depositor. There is always a risk that material will be misused, but that risk needs to be weighed against the risk that speakers will not find recordings of performances that may inform current practice, or whose records they may be able to enrich with current knowledge.

Digital language archive catalogues can provide a feed to a service provided by OLAC. It aggregates each archive's catalogue and lists all resources available per language with the URL http://www.language-archives.org/language/ followed by a language code.

OLAC lists 373,197 items in 60 archives (in 2023) compared to 223,664 items listed in 58 archives in 2008, a 66 per cent increase in fifteen years. But, if we set an arbitrary goal of at least 200 archival items for each of the world's languages (see below for more hypothesising about future records of languages), we would expect there to be in the order of 1,400,000 archival sources.

Table 3.1 shows the number of resources (bundles or items, which contain files) and languages represented by ten selected archives.

[3] https://archive.mpi.nl/tla, formerly funded by the DoBeS program of the Volkswagen Stiftung (*Documentation of Endangered Languages*, https://dobes.mpi.nl).

Table 3.1 Sample of language archive statistics provided by OLAC (July 2022).

Archive	Number of Resources	Distinct Languages
Archive of the Indigenous Languages of Latin America (AILLA) (USA)	289	257
C'ek'aedi Hwnax Ahtna Regional Linguistic and Ethnographic Archive (USA)	1,474	3
California Language Archive (USA)	14,959	295
COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO) (France)	17,495	273
Endangered Languages Archive (UK)	93,687	594
Kaipuleohone (USA)	5,621	226
The Language Archive (Nijmegen, Netherlands)	167,996	460
Living Archive of Aboriginal Languages (Australia)	3,738	41
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (Australia)	27,219	1,360
Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library (USA)	5,292	749

A further important development in the recent past was the formation, in 2003, of the network of archives known as the Digital Endangered Languages and Musics Archives Network (DELAMAN) of which Barwick was the inaugural president. With 14 current member archives, DELAMAN agrees on archival standards and helps with referring users to appropriate archives.

An ongoing challenge for academic repositories is how to make the materials they hold more accessible. Directories like OLAC increase the exposure of archive catalogues and make them available via other online services (e.g. the Virtual Language Observatory, or WorldCat, and also through Google). The webpages they present are, in general, not particularly attractive or intuitive to navigate as their effort is typically directed to building and maintaining collections with little funding. As most archives include information from a number of languages it is not possible to localise their interfaces, with a notable exception being the Archive of the Indigenous Languages of Latin America, which has both an English and Spanish (and soon also Portuguese) version of its catalogue and webpages.

More archives needed!

The extent of the task ahead can be daunting. Consider the many millions of words, sentences, recordings and movies in languages with many speakers (English, Spanish, French, Russian, Hindi, Mandarin to pick just six) that will provide an enduring record. To create comprehensive records of performance in each of the world's languages would mean extending the sort of records linguists and musicologists currently produce. Evans observed that "it is becoming feasible to record around 500 hours of linguistic material in the course of a year or two's fieldwork, thanks to the miniaturisation, fidelity and portability of our recording devices" (Evans 2017a, 41). As new technologies allow more recording, so increases the need for careful management of this wealth of material. As an impressionistic measure of the current scale of such outputs of fieldwork, let us assume there are 100 language documentation projects underway around the world at the moment and that they have a life of around five years each. Ideally, each modern fieldwork-based documentation project would create audio and video recordings, images, transcripts, and a dictionary. Looking at deposits in PARADISEC, an average modern collection would include something like 30 hours of audio, 10 hours of video, with transcripts, and associated files, in 300 files totalling around 100 gigabytes of data (nowhere near the ideal that Evans hoped for above). Of course some collections are much larger than this but for present purposes this will do. Over the next 20 years there would, on this artificial and probably conservative estimate, be four such sets of 100 projects, creating 40 terabytes of material that needs to be described, accessioned and curated over time by language archives. Additionally, if Evans's ambitious forecast eventuates, there will be 500 hours of recordings per language, a mix of video and audio that would amount to about a terabyte per language, and thus a requirement of 400 terabytes of storage. It can be envisaged that, in future, such volumes will not be problematic to obtain or as expensive as they are today.

It is important to distinguish between storage and archiving of these kinds of records. Most academics are provided with storage by their universities, and that is a necessary first step towards ensuring files survive into the future. But storage on its own does not include a catalogue of the contents, with controlled metadata terms, or licences for use of the material. Storage is often restricted to those within the institution and so is not a public-facing resource. Storage does not allow for differential or restricted access based on a user's characteristics. Storage does not enforce file-naming conventions, or convert formats over time as is necessary, or produce compressed formats of files for web delivery.

There is a backlog of legacy records yet to be accessioned and in some cases yet to be found. Legacy records include analogue tape collections made by linguists, musicologists and anthropologists. PARADISEC conducts an ongoing survey[4] to find these collections and has so far digitised most of those found. A recent example is the collection of 240 tapes made by Ian Frazer in the To'aba'ita language in northern Malaita (Solomon Islands) between 1971 and 1985. These tapes, together with a large set of notes, are kept in his house in Dunedin and appear to be the only known recordings in this language. In 2018 PARADISEC was granted funds by the ELDP's Legacy Materials Grants to digitise and accession this collection. There are many more similar collections yet to be located.

Cultural agencies involved in collecting oral tradition, like local cultural centres in Pacific nations, often have a backlog of recordings that need to be digitised, catalogued and archived. They need advice and help with choosing software to keep track of their collections, and ultimately they need an archive that they can trust to look after this material. PARADISEC, under Barwick's leadership, has worked to obtain the necessary funds and then to digitise tape collection from agencies including the Solomon Islands National Museum and the Vanuatu Cultural Centre, and, in 2022, the Yap National Archives.

The use of mobile phones as portable recorders is also increasing the amount of documentation that can be made by speakers themselves. Websites, online video, Facebook pages and so on all present a dynamic and often heterogeneous set of linguistic performances, with variation in spelling and mixing of language varieties. The challenge is to capture this record ethically and efficiently so that it is part of the long-term record for all small languages.

There is therefore a great need for more archives to deal with this foreseeable proliferation of language records. As was noted above, there have only been two new archives added to OLAC since 2008. While digital archives do not, in theory, need to be located in any particular place, it still makes sense for archives to be close to the speakers of the languages they represent. So, for example there is no general language archive in Canada but there have been discussions about setting one up there. In the United States of America, there are several archives that each focus on a particular region (Alaska, Hawaii, Latin America, California, Oklahoma) and recently the American Philosophical Society has emerged as a more general archive, but still only dealing with American languages. For Americans conducting research in other parts of the world there is no American archive. In the Francophone Pacific, the LACITO archive (Pangloss/CoCoON) in Paris serves local needs, and

[4] "Project Lost and Found", *DELAMAN*, https://www.delaman.org/project-lost-found.

there is a new archive based in French Polynesia (Anavevo[5]). In India, there is the CORSAL[6] repository (based in Texas but focused on north-east India). Japan has the resources and the research and recording tradition to build an archive but has yet to do so. An archive devoted to a single linguist has been established in the Philippines (Or and Estrellado 2023). Elsewhere in Asia there could be several digital archives, potentially in Singapore and Thailand. Similarly, in Cameroon, there was an archive operating with systems supplied by The Language Archive (TLA) in Nijmegen, but that no longer seems to be the case (the website of the Archive of Languages and Oral Resources of Africa – ALORA – is not accessible at the time of writing[7]). There does not seem to be another language archive specifically serving African languages.

The cost of not archiving

There is a clear social benefit to making cultural recordings safe and available over time. As noted elsewhere in this chapter, it should be a normal part of research practice to create records of the language we are working on in ways that can be accessed and used by the speakers and by their descendants. To fail to do this risks the trust we have established, which, besides being an act of bad faith in itself, can then make it more difficult for future researchers to work in the same area. In a science discipline it may be possible to re-create lost data, but, even then, notorious examples can be found of unique data being put at risk. In a (possibly apocryphal) story that resonates with my experience in searching for audiotape collections, experimental data from the moon landings was stored on data tapes that had apparently been overwritten to save costs, and no copies could be located in the archives.[8] After 35 years, determined researchers hunted down a copy in a basement storeroom, and the data was eventually copied for future use. Without this dedicated sleuthing, it would have been lost.

There is a real financial cost in not securing outputs of linguistic research. Typical fieldwork is an expensive project, requiring years of researcher training, then sometimes considerable travel to meet the speakers. Great care is taken to have good recording equipment and to make backups of recordings for use in research. There is the labour of the researcher analysing the recordings, organising them and

[5] "About us", *Anavevo*, https://v-anavevo.upf.pf/apropos.

[6] "CoRSAL: the computational resource for South Asian Languages", *University of North Texas*, https://corsal.unt.edu.

[7] https://www.delaman.org/members/alora/ [ALORA's link was not working when tested in September 2023].

[8] https://en.wikipedia.org/wiki/Apollo_11_missing_tapes.

making them accessible. There is an imposition on the language speaker's time and a possible expectation on their part that they are contributing to a long-term record of their culture and language. A major part of the investment, both financial and emotional, in the creation of fieldwork records will be lost if there is no service that takes the output of research and safeguards it for future use.

In a study of the cost of data centres or archives, Beagrie and Houghton (2014) found that their use resulted in very significant increases in research, teaching and studying efficiency. They noted that the "value to users exceeds the investment made in data sharing and curation via the centres".[9] Critically, they found that it is the ability to re-use existing data that increases "measurable returns on investment" (Beagrie and Houghton 2014, 16). In the specific case of the Archaeology Data Service, they calculated a benefit to users that is "5 times the costs of operation, data deposit and use" (Beagrie and Houghton 2013, 7). They also "identified a potential increase in return on investment in data creation/collection resulting from the additional use that was facilitated by ADS that may be worth between £2.4 million and £9.7 million over thirty years in net present value from one-year's investment – a 2-fold to 8-fold return on investment" (Beagrie and Houghton 2013, 7). The benefit of data curation, abstruse though it may appear, should therefore be apparent to even the most neoliberal and output-oriented administrator, of the kind who are making many decisions about research funding today.

There is also a great personal relief in archiving one's research materials. Many researchers only organise their materials when they come to archive their collection, so there is a benefit for them in then being able to retrieve their own materials over time, providing the crucially needed backup of a file long since lost from their laptop. Archiving is not so onerous if researchers are trained in data management methods (and practise what they have learned!). As Corti and colleagues noted, data management "reduces time and financial costs and greatly enhances the quality of the data you use" (Corti, Van den Eynden et al. 2014, 10). It also relieves the stress of knowing that the records in your care deserve to be made available but, at the same time, not knowing what to do with them.

From the perspective of funding agencies, the creation of primary research data is one of the outcomes they have supported. Some agencies already ask that primary data be made available and this will increasingly become the case (Tenopir et al. 2011). As an example of the scope of work that needs to be done, PARADISEC, in 2023 in its 20th year of operation, has archived 16,500 hours of audio recordings, representing over 1,360 languages, about half of it digitised from analogue tapes

[9] Beagrie and Houghton 2014, 4.

recorded in the mid-20th century. Typically, these older collections come from retired or deceased academic linguists and musicologists and are limited to audio and some paper notes. New digital collections are large and can contain many video files, in addition to transcripts and image files, all of which require more management than the earlier analogue collections.

The post-colonial archive

While archiving is often and easily considered to be an extension of the colonial enterprise, a major motivation for digital language archives is making records available to the people recorded and their families. I suggest this is a post-colonial activity.[10] In the past, for example, linguists were criticised for taking materials from communities with nothing being returned. For example, Paulina Yourupi, from Chuuk in Micronesia, asked:

> Whatever happens to the previous research? What benefits were they to our community? Were researchers who seek to get their PhDs merely exploiting us or was it for a greater good? When do we see products and results and not more study?[11]

Similarly, Smith (1999) critiques the role of academia in dividing indigenous cultures into discipline areas, "disconnecting them from their histories, their landscapes, their languages, their social relations and their own ways of thinking, feeling and interacting with the world" (Smith 1999, 28, and see also the opening quote above). Errington's (2001, 34) critique of a "colonial linguistics" is at a more abstract level than the issues discussed here, but his observation resonates with the argument put in this chapter that linguistic texts can be "more meaningful than their authors knew, moving beyond while also incorporating knowledge they provide – in some case, the only knowledge available".

What is to be done?

It should be clear that documentation created in the course of research requires archives, and it requires records to be deposited in archives. Keeping records in one place is no longer tenable – recall the disastrous fire that destroyed the national museum of Brazil in September 2018 (Solly 2019), taking with it countless unique language records. We can no longer do nothing and hope it will all work:

[10] See also Thieberger 2020.

[11] Paulina Yourupi, 2008, class paper, U. Hawai'i Mānoa Linguistics Department.

> [doing nothing] is actually a choice too . . . Thoughtful repatriation of ethnographic materials can assist not only in the decolonisation of anthropology, but in empowering both communities and the people who comprise them by allowing easier access to a greater range of ethnographic information. (Chambers et al. 2002, 213–214)

As an established researcher who has embraced the possibilities offered by new technologies, Barwick offers a fine example to the next generation of researchers of the decolonising possibilities offered by repatriation of musicological and linguistic records and the possibilities of building networks and infrastructure within academia that also serve the needs of speakers, performers and the general community.

Acknowledgements

Thanks to two anonymous reviewers for constructive comments that have improved this chapter.