Unlocking the archives

Linda Barwick

Sydney Conservatorium of Music
The University of Sydney, NSW 2006, Australia
[linda.barwick@sydney.edu.au]

Nick Thieberger

School of Languages and Linguistics, University of Melbourne,
Parkville, Vic 3010, Australia
[thien@unimelb.edu.au]

Abstract

The popular expression 'locked in the archive' suggests that items are impossible to find and access once they are archived. Benefiting from new technologies, digital language and music archives nowadays provide an increasing number of records online in and about the world's small languages. Just six of these archives list between them over 31,000 items, representing something like 2,300 languages. We can certainly do better at making records more widely available—especially records from small, marginalised and sometimes isolated communities—but how do we build pathways for re-use? We discuss the practice of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) through the rubric provided by the FAIR principles. Building resources for learning and teaching language, history and culture, revitalising local performance traditions or reinforcing social identity through broadcasting are all possible pathways for future re-use of archival material. Ultimately, it is up to community members to decide on what they will do with archival materials once they have access; and it is up to language archives to listen and do our best to keep the pathways open to enable that.

PARADISEC

Archival media recordings can be of immense significance, especially for small or marginalised communities interesting in bolstering the future of their language and culture by using old recordings to support and encourage learning by younger generations. In order to provide access to such recordings we have been building the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[20] since 2003. PARADISEC has sought collections of analogue tapes made in the region around Australia, typically by linguists and musicologists, and worked to digitise and make them available. We currently hold some 7,660 hours of digital audio records, most of which are openly available to registered users and many of which represent orphaned collections of analogue tapes, previously unfindable by anyone but the researcher or their descendants. In this paper, we offer a reflection on our experience of providing access to research records. Our ability to create systems that take advantage of new technologies has meant that the archive continues its online presence during periods of little or no funding. We have automated access to the archive's contents, and allow depositors to assign access to specific users.

The popular expression 'locked in the archive' suggests that items are impossible to find and access once they are archived. In fact, digital language and music archives nowadays provide an increasing number of records online in and about the world's small languages. Just six of these archives list between them over 31,000 items, representing something like 2,300 languages. Digital language archives[21] of the world's small[22] languages represent a novel infrastructure that provides citability of primary research materials, thus building a robust methodology that allows for verification of analytical claims made in research. Language documentation (cf Gippert et al 2006, Thieberger 2016), the digital successor to language description, emphasises creating permanent records of linguistic performance that can serve a range of multi-disciplinary uses. At the same time, because ethnographic (linguistic or musicological) recordings are of heritage value to the people recorded or their families, there is great scope for engaging with communities by delivering accessible records (Barwick 2004). For speakers of many languages, there is no reflection of their traditions on the web, so by lodging recordings with an online archive, researchers can not only ensure the longevity of their

[20] http://paradisec.org.au

[21] In 2017 there are 14 archives in the Digital Endangered Languages and Musics Archives Network (DELAMAN), and 60 listed in the Open Language Archives Community, of which perhaps 22 have primary records from small languages.

[22] We use the term 'small languages' (rather than 'endangered languages') following Dorian 2014.

recordings, but also begin to redress the imbalance in representation.

Given the scarce funding available for archiving language records, digital language archives do an admirable job of describing and curating their collections. Most online archival systems have been designed to make records much more generally findable and accessible than in the past—though no doubt there is work still to be done, especially for the benefit of small, marginalised and sometimes isolated communities. Building pathways for interoperability and re-use, however, remains a challenge.

In our experience, community members are appreciative when they find records in their own languages, and we are keen to make that easier for them to do so. Some recent testimonials include the following response from E'ava Geita, a Koitabu speaker from PNG:

If only you witnessed and captured the reaction in me going through the recordings at home! It is quite an amazing experience! From feeling of awe to emotion to deep excitement! The feeling of knowing that your language has been documented or recorded in a structured way, kept safely somewhere in the world, hearing it spoken 50-60 years ago and by some people you haven't seen but whose names you only hear in history, is quite incredible.

It is most heart-warming to know that it is possible to sustain the life of my language. Thank you once again for the opportunity to listen to the records!

As I see it now, it is "US OR NEVER!" In about three or four Koitabu villages, the observation is that people in my peer group can understand in full but speak a little. Generation after us cannot do both. I believe that if we don't teach ourselves now and pass on, the language will be totally gone in the next 10 to 20 years.

As this user suggests, it is important to recognize that recordings of particular performances do not capture 'the language' and cannot claim to do more than provide examples of language in use. Nevertheless, they are priceless reflections of an individual's linguistic abilities, and, in the case of songs or oral poetry, capture some of the local cultural aesthetic. While archives can commit to preserving data and metadata, the only way for languages and other cultural traditions to be preserved is through active use in communities. Re-use of archival materials can support community-led processes of maintenance and revival of language and other cultural traditions. For archives wishing to facilitate re-use of the cultural material they have so carefully preserved and curated, it is essential to: 1. make sure that archive processes support re-use; 2. maximise the chances of the recording being discovered by potential users; and 3. make materials available in suitable formats for re-use.

FAIR language archiving

PARADISEC has been operating for fifteen years, so our practices foreshadowed the FAIR principles recently formalised by Force11.[23] The acronym FAIR stands for the four principles for metadata management – that metadata should be (1) Findable; (2) Accessible; (3) Interoperable, and (4) Re-usable. If applied by researchers and archives to both metadata and data, these principles can go a long way to unlocking the data in an archive. These four principles provide a useful framework for discussing the ways in which we have built the PARADISEC collection and made it accessible. We will now show how each of these is implemented by PARADISEC.

Findable PARADISEC

PARADISEC makes its data findable in several ways. We have an online catalogue that exposes basic metadata and can be searched on our site.[24] In developing our cataloguing terms (metadata) we have balanced making items findable against making our metadata too complex, and we think we have the right mix. Stemming from our initial focus on legacy recordings for which there was little metadata available, our fairly minimalist approach has served us well by making it easier for potential depositors to complete their deposits. The harder it is to fill out the metadata forms, the less likely it is to be done. Of course, too little metadata risks that some resources are not findable via the catalog, so we do have some standard requirements. We provide a simple spreadsheet for depositors to fill out with minimal metadata to describe their items, which can then be uploaded to populate the catalog. Making it as easy as possible to enter information into the catalogue provides an incentive for depositors to complete this otherwise potentially tedious task.

Normal textual searches of language names, regions, countries and individual names are all provided for. One of the important metadata elements that we routinely require is indication of the languages contained in the resource, or that are the target of the documentation. PARADISEC relies on depositors to include in their descriptions and titles key terms in small languages that will make the resources discoverable by speakers. Where a world language other than English may be in widespread use by speakers, some depositors include bilingual descriptions in (for example) Tok Pisin, Bahasa Indonesia or Chinese. Although we recognize the value of such bilingual descriptions to making the resources more findable, PARADISEC doesn't have the resources to routinely implement languages other than English in our metadata or catalogue interface.

[23] https://www.force11.org/group/fairgroup/fairprinciples

[24] http://catalog.paradisec.org.au

We have two feeds[25] of catalogue information that are harvested by a number of search engines, allowing users to find our collections via different types of querying of our catalogue by external services like the Open Language Archives Community.[26] This all makes our catalogue maximally locatable, via Australia's TROVE, [27] international catalogues like the Virtual Language Observatory [28] or WorldCat, [29] and also through google.

To allow for geographic searching of the catalog, a depositor can draw a bounding box on a map to indicate the geographical locus of the research. This information is stored in our catalogue as coordinates (minimum and maximum latitude and longitude) and displayed on the catalogue record for relevant collections and items, as well as feeding to the summary map on our catalog's homepage.

We have been exploring new ways to publicise our collection. In the Glossopticon[30] project we have built a virtual reality system that allows the user to see shards of light emanating from the ground in a 3D map. Each shard represents a language and if PARADISEC has an audio record in that language then we have snipped a 20 second piece from five minutes into that file and made it available to be heard. The effect is of walking through a forest of languages, with the volume of each snippet increasing as you walk towards a language location and decreasing as you walk away. While Glossopticon is currently not showing snippets from the full collection, we are planning that a new version will be a point of entry to the whole of the PARADISEC collection. We have also presented a selection of texts using augmented reality in which a poster presents an image related to the text. Passing an ipad or similar device over that image then triggers a video to play on the device.

Our use of these new technologies has generated considerable press coverage (for example in the PNG Post Courier, in Pursuit, [31] and a number of radio interviews[32]), which has meant that many more people are now aware of the collection. Such publicity, along with mentions in academic publications, presentations at conferences, social media posts and even word-of-mouth, all contribute to making our collections more findable.

We participate in international efforts to encourage the use of 'landing pages' or collection descriptions, which describe the context in which a collection was built and give an overview of its contents. Such overviews can be published as online catalogue pages—as for example in the ELAR archive[33]—in journal articles (e.g., Salffner 2016), or as web pages (e.g., Thieberger n.d.).

Like most online repositories, our archive PARADISEC has often been contacted by community users who have discovered us by a simple web search. But many small or economically disadvantaged communities do not have the resources to find our collections this way. In these cases, personal or place-based networks can be the key allowing people and recordings to be connected. Perhaps the original depositor has ongoing contact with community members? Or can we forge partnerships to reach out to potential users via local cultural centres, schools or broadcasters?

Accessible PARADISEC

Since its establishment, PARADISEC has always used persistent identifiers for objects in our collections, down to the file-level. In 2016 we added digital object identifiers (DOI)[34] to all collections, items and files. This provides certainty that the objects referenced will remain accessible into the future even if the URLs or services change over time (that is, if the resource moves or the host service restructures). The same DOI will continue to resolve to the correct resources at their new locations.

Accessibility of data also relies on the format of the data provided by the archive being readable, and on the user's access to information on how it is licensed. PARADISEC requires data to be formatted according to normal archival standards, meaning that files we provide can be read by common and widely available software. In some cases, data deposited has to be converted from other formats before it can be accessioned. For example, we will not store compiled databases, but will accept a properly exported textual version of them, and similarly, Microsoft Word (.doc) documents need to be converted to an open format to ensure their legibility over time. For audiovisual media, digitisation to a suitable international standard is the only means of preserving future access to audio or video originally recorded in obsolete formats. Archival preservation master versions need to be high resolution (where possible uncompressed) to maximise their future usability for subsequent editing, repackaging, and/or publication.

[25] http://catalog.paradisec.org.au/apidoc

[26] http://www.language-archives.org/

[27] http://trove.nla.gov.au/

[28] https://vlo.clarin.eu

[29] https://www.worldcat.org/

[30] See a news story and YouTube clip about Glossopticon here: https://rebrand.ly/ABCGlossopticonStory

[31] http://link.coedl.net/PursuitVRarticle

[32] e.g., https://rebrand.ly/PacificBeat-VRstory

[33] https://elar.soas.ac.uk

[34] https://www.doi.org

The ultimate key to accessibility of archival holdings is having robust licences that specify the terms under which items are held and distributed, and systems to support authorised distribution. Depositors assign access conditions to all items in their collections, down to the level of an individual item, and can also assign specific access rights to individual users. The content of the catalogue itself is licensed using a Creative Commons Attribution-ShareAlike 4.0 International License.[35]

We have designed forms and systems to minimize the barriers to access, so that any item with the access condition "Open" can be accessed, listened to, or downloaded by registered users. Registration simply involves sending an email request, then confirming one's email address and accepting the access conditions.[36]

Interoperable PARADISEC

To be interoperable, metadata and data need to use formal, accessible, shared, and broadly applicable languages and standards for knowledge representation. PARADISEC's work is interoperable at two levels: one is the structured metadata stored in our catalog, and the other is the data in the collection.

The terms used in our catalog—such as title, date, description, role, country and so on—conform to international metadata standards (Open Archives Initiative[37] and Dublin Core[38]). This is what makes our metadata feeds interoperable with the harvesting services noted earlier. To ensure consistency when items are entered into the catalog, we use controlled vocabularies in dropdown menus, as well as text completion based on existing records. The language vocabulary we use is based on the International Standards Organisation 639-3[39] codes for representation of the names of the world's languages, thus facilitating interoperation with other language archives world-wide. Each time a catalogue record is saved, the information is written as an xml file into the same directory as the stored data, ensuring that the knowledge held as structured metadata in the catalogue is also kept in the collection itself.

The interoperability of the data in the collection largely depends on data formats and how data is structured—this is the depositor's responsibility. We provide online advice and also run occasional training sessions for depositors so that they know how to create data in interoperable formats. We also work to convert incoming data from proprietary formats to open formats where possible. For example, a dictionary written in a word processor is less useful for other services than is a lexical database created in a tool like Fieldworks Language Explorer.[40] Several dictionaries deposited as Microsoft Word documents have been converted to structured text to make them more archivable and interoperable.

Re-usable PARADISEC

Re-usability requires data to be in standard formats. Archives keep the best possible version of a file, which is often too large to distribute for most purposes. As part of the deposit process we automatically transcode files from high resolution to smaller files for quicker access and store them all together. For example, our preservation audio standard results in 24-bit 96khz stereo broadcast wave files, which come out to about 2 GB per hour or audio. MP3 dissemination audio files of the same duration come out to approximately 56MB, or approximately 1/20th the size of the uncompressed preservation master.

PARADISEC allows authorised online users to view, stream and download materials via our web interface. Various projects have explored innovative ways of enriching the online experience by linking text and media. In the online system EOPAS,[41] we demonstrate the advantage of linking media to interlinear text (which includes word-by-word meanings as well as a free translation), and then presenting the interlinked text and media online. We have also helped set up community repositories in partnership with local cultural centres, which can then assist end users to copy relevant material. In places without easy access to digital infrastructure, books or other printed materials based on archival media may be the best way to provide access.

Building resources for learning and teaching language, history and culture, revitalising local performance traditions or reinforcing social identity through broadcasting are all possible pathways for future re-use of archival material. Ultimately, it is up to community members to decide on what they will do with archival materials once they have access; and it is up to language archives to listen and do our best to keep the pathways open to enable that.

Of course, re-usability is also dependent on the other FAIR features already discussed. Any future **re-use** depends on being able to **find** materials, which need to be in formats that are **accessible** and **interoperable** and they need to be of sufficient quality (in content and in format) that they can be used.

[35] http://creativecommons.org/licenses/by-sa/4.0/

[36] Our access conditions are set out here: http://www.paradisec.org.au/deposit/access-conditions/

[37] https://www.openarchives.org/

[38] http://dublincore.org/documents/

[39] https://www.iso.org/standard/39534.html

[40] https://software.sil.org/fieldworks/

[41] http://eopas.org/

Beyond FAIR

Up to this point, the FAIR principles have provided a useful framework for discussing what we do in PARADISEC. However, we also engage in other activity that helps unlock the archive.

In 2015 we built a viewer[42] for items in the collection that identifies what kind of data type an item contains and presents a viewer appropriate for that item. For example, we pre-generate and store thumbnails as a navigation aid for an item containing images. Audio files in a streaming format can be played from any point in the file, and if they have time-aligned transcripts then the transcript will scroll along with the media. Interlinear text transcripts are also provided for in this viewer. We have plans to include a spectrograph image above the audio slider to allow navigation via the visual cues in the spectrograph to points in the audio file that it presents.

In a number of projects, we have digitised collections for museums or cultural centres in the Pacific. We then store the files and deliver them via our catalogue (if we are permitted to by those agencies). We also return hard drives of all the files to the relevant agency for loading on a suitable local computer. It would be ideal if we had a way of presenting the catalogue for these items as a subset of our online catalog, but at present we lack that capability. One solution that we have tried for local access points is to build an iTunes installation[43] for which we embed metadata from our catalogue in mp3 files, together with a photo of the performer recorded on each file. This has the advantage of using an existing software platform that is widely available, operates on multiple platforms, and with which many computer users are already familiar. Since the installation is only available on a single computer it avoids internet authentication issues.

We are routinely asked for hard disk copies of multiple files related to a particular language and we happily provide this service when we can. In some cases feedback from users leads to an enrichment of our catalogue description. In a recent case, the library at the Divine Word University in Madang arranged for recordings to be played at the local market in order to see if anyone recognised speakers or could add more information to the catalog.

We are also exploring the possibilities offered by librarybox,[44] which establishes a local wifi network independent of the internet. This means that any device that can receive wifi can be used to access records transmitted from the librarybox device.

[42] This is an open source javascript system available here: https://github.com/MLR-au/pdsc-collection-viewer/

[43] Discussed here: http://rebrand.ly/ITunesRepatriation

[44] http://librarybox.us/

Conclusion

PARADISEC is an archive established and built by researchers. It has taken advantage of technological advances that, in some cases, traditional archives are not yet able to adopt and is part of an international network of similarly innovative archives. We have been involved in discussions with several groups who would like to establish regional archives and hope that more such projects will emerge in the near future, especially in our region (perhaps in Singapore or Japan for example). While our collection is already accessible to online users, we continue to publicise the collection in order to make it as findable and accessible as possible. Archives like PARADISEC, in making previously locked-up analogue recordings available for re-use, are giving new life to old records.

Acknowledgments

Work described here was carried out with the help of the following grants: Endangered Archives Programme grant 693 (Preservation of Solomon Islands analogue recordings), Australian Research Council LIEF program (2003, 2004, 2006, 2011), ELDP LMG0009 Vanuatu Cultural Centre tape digitisation, ARC Centre of Excellence for the Dynamics of Language, ARC Future Fellowship FT140100214.

References

Barwick, L. (2004). Turning it all upside down . . . Imagining a distributed digital audiovisual archive. *Literary and Linguistic Computing,* 19(3):253-263. http://hdl.handle.net/2123/13114

Dorian, N. C. (2014). *Small-Language Fates and Prospects: Lessons of Persistence and Change from Endangered Languages: Collected Essays.* Brill's Studies in Language, Cognition and Culture, volume 6. Leiden: Brill.

Gippert, J., N. P. Himmelmann, and U. Mosel (eds). (2006). *Essentials of Language Documentation.* Berlin: Mouton de Gruyter.

Salffner, S. (2015). A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9:237–267.

Thieberger, N. (2016). Documentary Linguistics: Methodological Challenges and Innovatory Responses. *Applied Linguistics.* 37.1:88-99. doi: 10.1093/applin/amv076

Thieberger, N. (n.d.) Guide to the Nafsan, South Efate, collection. http://www.nthieberger.net/sefate.html