When Your Data is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

COLLECTION: MULTIDISCIPLINARY DATA ACTIVITIES BRIDGING THE RESEARCH COMMUNITY AND SOCIETY

PRACTICE PAPER

ubiquity press

NICK THIEBERGER

AMANDA HARRIS

Author affiliations can be found in the back matter of this article

ABSTRACT

In this paper we discuss the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a research repository that explicitly aims to act as a conduit for research outputs to a range of audiences, both within and outside of academia. PARADISEC has been operating for 19 years, and has grown to hold over 390,000 files currently totaling 150 terabytes and representing 1,312 languages, many of them from Papua New Guinea and the Pacific. Our focus is on recordings and transcripts in the many small languages of the world, the songs and stories that are unique cultural expressions. While this research data is created for a particular project, it has huge value beyond academic research as it is typically oral tradition recorded in places where little else has been recorded. There is an increasing focus in academia on reproducible research and research data management, and repositories are the key to successful data management. We discuss the importance for research practice of having discipline-specific repositories. The data in our work is also cultural material that has value to the people recorded and their descendants, it is their grandparents and so we, as outsider researchers, have special responsibilities to treat the materials with respect and to ensure they are accessible to the people we have worked with.

CORRESPONDING AUTHOR:

Assoc/Prof Nick Thieberger

School of Languages and Linguistics, University of Melbourne, Australia

thien@unimelb.edu.au

KEYWORDS:

linguistics archiving; musicological archiving; language documentation; language data management

TO CITE THIS ARTICLE:

Thieberger, N and Harris, A. 2022. When Your Data is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). *Data Science Journal*, 21: 9, pp. 1–7. DOI: *https://doi.org/10.5334/dsj-2022-009*

INTRODUCTION

In this paper we discuss a research repository that explicitly aims to act as a conduit for research outputs to a range of audiences, both within and outside of academia. Our focus is on recordings and transcripts in the many small languages of the world, the songs and stories that are unique cultural expressions. While this data is usually created for a particular research project, it has huge value beyond academic research as it is typically oral tradition recorded in places where little else has been recorded.

Digital language and music archives are a recent phenomenon (Barwick 2004) that provide descriptions of such items in standard terms and then curate these culturally significant materials, often the result of digitisation of analog recordings. These analog recordings are at-risk and can no longer be heard in the source communities as playback equipment is unavailable. There is an international network of language archives, called the Digital Endangered Languages and Musics Archives Network (DELAMAN[1]) which shares information and provides a point of reference for anyone seeking language collections. Our project, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) is a member of DELAMAN, and has been operating for 19 years. It has grown to hold over 390,000 files currently totaling 150 terabytes and representing 1,312 languages, many of them from Papua New Guinea and the Pacific. The main focus of the collection has been audio tapes recorded since the 1950s, and much of our effort is devoted to locating and digitising unique analog tapes.

It is important for research practice to have discipline-specific repositories. Citable data is integral to proper research (Berez-Kroeker etal. 2018), allowing verification of events on which analytic claims are made, and replication of analysis (cf Himmelmann 1998). But the data in our work is also cultural material that has value to the people recorded and their descendants, it is their grandparents and so we, as outsider researchers, have special responsibilities to treat the materials with respect and to ensure they are accessible to the people we have worked with.

There are some 7,000 languages in the world (Eberhard etal. 2022), and few records exist for most of them. Often, the most detailed records are those made by academic researchers – linguists, musicologists, anthropologists – who have spent time studying performance in those languages. But, without a digital repository to store these unique records, they are at risk of being lost. At the same time, there is an increasing focus in academia on reproducible research and research data management (Corti et al. 2014), and repositories are the key to successful data management. In the Humanities, in particular, research records are likely to be of interest to a broader public than is most scientific data, so there is a greater need for infrastructure to describe, curate, and make this material accessible.

As we will show, discipline-specific repositories that conform to proper standards can address the needs of researchers while, at the same time, providing metadata feeds (OAI-PMH, RIF-CS) to national data systems, as, in our case, to the National Library of Australia, or Research Data Australia. And, perhaps most importantly, the records can be returned to the people recorded in them, and this can occur iteratively to any location that has an internet connection. Increasingly, mobile phones provide internet access in even the most remote areas, so, when someone goes looking for information in their own language, we hope they will be able to find it in our collection. We want to emphasise that this work aims to bridge the digital divide in a most practical way, building a repository with a simple ingestion system and an appropriate metadata schema to make it as easy as possible to add new items, and for them to be licensed and made accessible (Barwick and Thieberger 2018).

This kind of archive provides a fixed, citable form of research data, and so is a locus of activity, a place from which research materials can be commented on and re-used in novel ways, with new knowledge reflected back into the collection over time. Far from being the endpoint for

1 Current members of DELAMAN (*https://www.delaman.org/*) are: AILLA (Archive of the Indigenous Languages of Latin America); Alaska Native Language Archive; California Language Archive; ELAR (Endangered Languages Archive at SOAS); Kaipuleohone (University of Hawai'i Digital Ethnographic Archive); Native American Languages Collection at the Sam Noble Museum of Natural History; PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures); RWAAI (Lund University); SIL International Language and Culture Archives; The Language Archive (Max Planck Institute for Psycholinguistics); The Library of the American Philosophical Society (APS).

research, the archive reinserts these materials into an ongoing and dialogic relationship with the people recorded and with future researchers (Barwick 2004). Without the archival effort, these materials would remain inaccessible once the project that created them ended. In our experience, without the archive, the materials would probably also be inaccessible to all but the most organised researchers, if not lost due to hard disk or computer failure.

It is important to distinguish a curated set of archival files from files on a storage system. Requests to our system administrators at a university for longterm archiving are typically met by suggestions to use file storage systems, and, even then, there is usually no guarantee of longevity beyond the current funding cycle. File storage is a necessary first step, and can provide needed backup while a collection is being established. However, file storage alone does not provide the metadata to describe items and it applies no licences to let others know how the materials can be used. An archive also checks files regularly, doing checksums to ensure data integrity. And for all of those reasons, the files in a storage system will not be findable, accessible, interoperable, or re-usable, as they should be according to the FAIR principles.[2] Further, we adhere to the CARE[3] principles as much as possible. We recognise our responsibility is to ensure the safety and longevity of analog recordings that have been kept out of circulation and have been inaccessible to the people whose ancestors are recorded in them. As we deal with a large range of languages (currently 1,312 represented in the collection) we are unable to consult with speakers, who are often in remote locations. We then have a choice about making the records available or waiting until we have permission. We have seen institutions that close off access in this situation, and we feel that this approach compounds the problem. We use a takedown principle, in which we advise users of the catalog to contact us if they feel an item should not be made available, and we encourage speakers with a connection to a recording to contact us.

WHAT DOES PARADISEC DO?

PARADISEC is an archive that explores exciting new possibilities for research materials. While the initial impetus was to digitise analog tapes from an earlier period of research (since the 1950s), once we had built the necessary infrastructure we now also curate born-digital files arising from current fieldwork, and they represent about half of the current collection.

We need to be clear that this project does not 'save languages' and does not 'save musics'. We are saving *records* of performance, and that serves to reflect the diversity of language and performance that exists in the world. It also serves to give presence to voices that are usually marginal and excluded from the internet. In what could be considered applied post-colonial practice, we aim to make these records available again to the people they came from (Thieberger 2020).

We have built a system that automates most of the processes of file ingestion, quality assurance, user management, and access for collections of research materials, especially media recordings, transcripts and material associated with linguistic or musicological fieldwork. We provide advice on our webpage about data management and file naming, and we run regular training sessions to encourage thinking about archiving from the beginning of fieldwork. We run training workshops in the use of appropriate tools whose output can be archived and is not locked into proprietary formats. Once files are in the collection we assign digital object identifiers (doi), and our system enforces access conditions. Each registered user of the catalog accepts a set of conditions[4] and each depositor specifies how their materials can be used. Even if an item is given 'closed' status, meaning no-one can access its files, the depositor can assign individual rights to other registered users to use that item. We also allow for a 'private' status as a collection is being built, that closes even the metadata from public view, and no doi is assigned until that private status is ended. For items listed as 'open', a registered user can download the file.

2 *www.go-fair.org*.

3 *https://www.gida-global.org/care*.

4 *https://www.paradisec.org.au/deposit/access-conditions/*.

Each file entering the collection is processed according to its type, with illegal filetypes immediately rejected and an email report generated. Our filename format reflects the hierarchy of collection id-item id-filename.extension (e.g., NT1-98007-01.wav). For permitted filetypes, a processing pipeline is opened, so, for example, a wav file is wrapped in a set of metadata terms copied from the catalog in an xml file to produce a broadcast wav format (BWF) file and an mp3 version is created for online delivery. A tif file is converted to jpg for delivery. We are processing video files offline as they take a great deal of processor power, and we produce an archival mxf file and mp4 for delivery. The archival form of the file and its compressed derivative are sent to the appropriate directory in the archive based on their filename.

We run a daily report on the checksum status of files in case any changes have occurred. Our catalog is backed up daily to another location, as is the collection, and we periodically do a disaster recovery exercise in which we retrieve files from the backup to ensure its integrity.

In 2019 we received the World Data System3[5] data seal, signifying we conform to all required standards. In 2013 PARADISEC's collections were inscribed in the UNESCO Australian Memory of the World.[6]

AN URGENT TASK

The unique recordings we have focused on in PARADISEC were made on analog tape up until the 1990s when digital recordings became the norm. Analog tape, on reels or cassettes, has a major problem in that it is likely to become unplayable within the next few years.[7] The lack of playback equipment for these tapes is one factor in their inaccessibility, in particular for open reels, but also for cassettes. More critically, the tapes themselves will begin falling apart, having reached the end of their lives. Paradoxically, while analog tape is fragile, we know that digital records are even more fragile, yet digitisation is currently the recommended means of preservation of analog audio.[8] Analog tapes made in the 1950s are still playable, but we have probably all had the experience of being unable to open digital files made even ten years ago, due to changes in formats and software. A partial solution is to ensure that all files are converted to a format we know has more chance of surviving, and here we simply follow established standards. Thus we archive wav, txt, xml, and tif, but also store lower resolution copies as mp3, pdf, or jpg for delivery. And we make daily backup copies in different physical locations.

THE COST OF NOT ARCHIVING RESEARCH MATERIALS

The value of research data can be calculated either as the cost of creating it, or its value to subsequent generations as an accessible set of material. For the kinds of materials discussed here, value is a product of the cost of mounting research projects, and the amount of effort put into the recordings and their annotation and transcription. For fieldwork in places far from a researcher's base, there are travel costs, time involved in building relationships with performers, speakers, and the communities in which they live. There is a cultural value to the only recordings in a particular language that must also be taken into account when considering how to attribute costs. The recordings made are the result of collaborative and trusting relationships developed over time. Ensuring that records made in this context can be accessed in future is part of the research process and is increasingly being recognised as integral to good practice by funding agencies, who expect data management plans to be followed by grantees. All of that is at risk if provision is not made to look after the records in the long term.

5 *https://www.icsu-wds.org*.

6 *http://www.amw.org.au/register/listings/pacific-and-regional-archive-digital-sources-endangered-cultures-paradisec*.

7 NFSA (National Film and Sound Archive of Australia), 2014. *Deadline 2025: collections at risk*. *https://www.nfsa.gov.au/collection/curated/deadline-2025-0*.

8 International Association of Sound and Audio-visual Archives, Guidelines on the Production and Preservation of Digital Audio Objects, *https://www.iasa-web.org/tc04/audio-preservation*.

PLANNING FOR PROJECT ENDING

The practice of preparing a collection for archiving with PARADISEC is already a great advance on the usual assemblage of materials typically held by a researcher. It is only when faced with having to describe files, no matter how minimally, for deposit into the collection, that researchers do a stocktake and identify what they have, what they have processed (e.g. transcribed or annotated), and what work remains. The archive caters to the ending of a research project, but also provides a basis for ongoing interaction with the data. The archive creates persistent identifiers for these items, so they can be cited in examples in publications, and other people can access them to verify claims made based on these examples. In addition, the materials now become available for others to use, as the deposit process includes specifying what kind of access conditions are to be applied. In this way, a more robust research methodology is established, and, as the records we are concerned with have heritage value, an additional benefit is the ability of speakers of these languages to access the recordings.

PARADISEC has had many accolades, but we are aware that it exists in a research infrastructure environment, both at universities and at a national level, that does not provide for longterm curation of research outputs. We have pioneered a system for describing and curating particular kinds of research data, and know that our methods can be extended to other types of data. We had seen examples of systems in which data and its metadata descriptions were separated, the metadata in a database and the files in the collection stored elsewhere, and we did not want to use that approach in our collections.

We do have a database that controls metadata entry, and provides checking of incoming materials for conformance to our standards (e.g., file naming, metadata terms, file types, unique identifiers, assignment of digital object identifiers). We have always written a complete description of each item into the directory that holds the data, so it is a self-describing collection, independent of cataloging software, and the whole collection's catalog can be reconstituted from the collection itself. Each item stores collection-level and item-level metadata that is updated every time the catalog entry is edited and saved.

Since each directory in the collection is self-describing, we are able to create arbitrary subsets of the collection and not lose the metadata (contextual information). One of our common tasks is loading files onto a hard disk to make sets of records available to cultural centres or museums in the Pacific. Simply putting files on a hard disk does make them available, but, without a catalog, they are difficult to navigate. We have written an app[9] that looks into this set of files on the hard disk and writes an html catalog of just those files to create a local viewer. The same services that we provide in the online catalog (media players, image viewers, and so on) are also available in this local viewer. We write these bundles of sub-collections and a catalog to raspberry pi units which serve a small wifi network, allowing the catalog and files to be read on a mobile device.

As our current system is now ageing, we are moving to use the Oxford Common File Layout (OCFL[10]) to store the files and Research Object Crate (RO-Crate[11]) to provide the same kind of description we had in our non-standard XML files, but now in a standards-compliant format. Our previous system required having specialist knowledge of Ruby on Rails and we found it difficult to engage people with the skills required to maintain that system. OCFL/RO-Crate are written in json which is a commonly used technology and so should be more robust for our next phase of development. We have built a demonstrator using these standards.[12] Given that we hold 150 terabytes, it is not feasible to have a sandbox version of the entire collection, so the demonstrator has the entire catalog and about three-quarters of the files. The new version should be ready in early 2023.

CONCLUSION

We have presented an ongoing project that builds a work cycle from creation of primary data, through its primary use, file naming and data format selection (for optimal longevity

9 *https://language-archives.services/about/data-loader*.

10 *https://ocfl.io*.

11 *http://www.researchobject.org/ro-crate*.

12 *https://mod.paradisec.org.au*.

of the files), to deposit in a repository. Deposit conditions are applied by the user, and access is governed by another set of conditions, agreed to by registered users. Together with the other data-checking systems discussed here, PARADISEC provides a model for the curation of research data in a responsible way that is also responsive to the broader community's interest in these materials, and acknowledging that what is song data for research is also a recording of someone's grandparents singing.

While this example is specific to our context and experience, we offer it here as a model of what can be done, recognising the need for many more language archives around the world, able to address local language conditions and being closer to the people most interested in the collections, the speakers and their descendants.

FUNDING INFORMATION

Australian Research Council LIEF grants LE110100142, LE0560711, LE0453247, ARC DP0450342, DP0984419, & FT140100214, ARC CE14010004, Australian Research Data Commons (2019).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Each author has participated equally in the work described in this article and in writing the article.

AUTHOR INFORMATION

Nick Thieberger is concerned to make language materials secure and re-usable both for speakers of the language and for current and future researchers. He worked at the Australian Institute of Aboriginal and Torres Strait Islander Studies to create the *Aboriginal Studies Electronic Data Archive* in 1991.

He is part of the team that established the *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (*paradisec.org.au*). He set up the archive *Kaipuleohone* at the University of Hawai'i in 2008.

He is interested in digital research methods and their potential to improve research practice and is developing methods for creation of reusable data sets from fieldwork on previously unrecorded languages. He is the editor of the journal *Language Documentation & Conservation*. He taught in the Department of Linguistics at the University of Hawai'i at Mānoa and is now an Associate Professor in the School of Languages and Linguistics, University of Melbourne. *http://nthieberger.net*

Amanda Harris's research combines methods from historical studies, musicology and digital humanities. She is interested in collaborative research that engages speaker communities with archival materials and that mentors emerging Indigenous scholars in developing collaborative methodologies. As Partner Investigator in the Leverhulme funded project True Echoes: reconnecting cultures with recordings from the beginning of sound (2019–21), she collaborates with cultural heritage institutions in PNG, Solomon Islands, New Caledonia and the UK. At the Sydney Conservatorium of Music, she is a Senior Research Fellow and works as part of an interdisciplinary Australian research team on a project funded by the Australian Research Council entitled 'Hearing the music of early NSW 1788–1860. Her book *Representing Australian Aboriginal Music and Dance 1930–70* was published by Bloomsbury Publishing in 2020.

AUTHOR AFFILIATIONS

Nick Thieberger *orcid.org/0000-0001-8797-1018*
School of Languages and Linguistics, University of Melbourne, Australia
Amanda Harris *orcid.org/0000-0002-9858-2568*
Sydney Conservatorium of Music, The University of Sydney, Australia

REFERENCES

Barwick, L. 2004. Turning it all upside down... Imagining a distributed digital audiovisual archive. *Literary and Linguistic Computing*, 19(3): 253–263. DOI: *https://doi.org/10.1093/llc/19.3.253*

Barwick, L and **Thieberger, N.** 2018. Unlocking the archives. In Ferreira, V and Ostler, N (eds.), *Communities in Control: Learning tools and strategies for multilingual endangered language communities. Proceedings of the 2017 XXI FEL conference*, 135–139. *http://hdl.handle.net/11343/220007.*

Berez-Kroeker, AL, Gawne, L, Kelly, BF, Heston, T, Kung, S, Holton, G, Pulsifer, P, Beaver, D, Chelliah, S, Dubinsky, S, Meier, R, Thieberger, N, Rice, K and **Woodbury, A.** 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1): 1–18. DOI: *https://doi.org/10.1515/ling-2017-0032*

Corti, L, Van den Eynden, V, Bishop, L and **Woollard, M.** 2014. *Managing and Sharing Research Data: A Guide to Good Practice (Paperback)*. London: Sage Publications.

Eberhard, DM, Simons, GF and **Fennig, CD.** (eds.) 2022. *Ethnologue: Languages of the World.* Twenty-fifth edition. Dallas, Texas: SIL International. Online version: *http://www.ethnologue.com*.

Himmelmann, N. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1): 161–195. DOI: *https://doi.org/10.1515/ling.1998.36.1.161*

Thieberger, N. 2020. Technology in support of languages of the Pacific: neo-colonial or post-colonial? *Asian-European Music Research Journal*, 5–3: 17–24. DOI: *https://doi.org/10.30819/aemr.5-3*

TO CITE THIS ARTICLE:
Thieberger, N and Harris, A. 2022. When Your Data is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). *Data Science Journal*, 21: 9, pp. 1–7. DOI: *https://doi.org/10.5334/dsj-2022-009*

Submitted: 24 March 2021
Accepted: 28 March 2022
Published: 04 April 2022

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.