Preview
Volume Number

Doing it for Ourselves: The New Archive Built by and Responsive to the Researcher

Nick Thieberger <thien_at_unimelb_dot_edu_dot_au>, School of Languages and Linguistics, University of Melbourne, Australia

Abstract

In this paper I address the following research questions in the context of having built a research data repository to safeguard cultural research data. How can the PARADISEC team ensure the records we create in the course of our research will exist into the future and remain citable? How can our research data be made available for a wider public, most importantly for the people recorded and their descendants? How can we prepare our students for this new approach to curation of primary research data so that they can build good methodology into their normal research practice, with much more productive outcomes?

Introduction[1]

Imagine a world in which research was valued so that research records were not periodically lost at the end of every research project. To avoid project endings resulting in data loss, researchers need a long-term data service that ensures continued access to their primary records. Storage on hard disks is not curation; it simply amasses files with no metadata, no license or deposit conditions, and no public access. In the absence of institutional or national work to preserve the outputs of research in our disciplines of linguistics and musicology, our project, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), developed exemplary methods and services for curating research data based on accepted standards for metadata and data formats. The data we have focused on is analog tape, recorded in fieldwork, and often the only recording made in a particular indigenous language. Initially recorded for the purposes of research, these recordings have heritage value to the people recorded and their descendants. In addition, we provide citable data for use in building on existing research that was previously inaccessible. We continue to train new generations of researchers to create well-described research materials, recognizing that archival formats should be created in the course of normal research to make the whole process of curation much easier for both the depositor and the archive. Researchers seek advice about how to manage their records and are open to learning how to make better records that they will be able to access in future. We act as a repository that curates and preserves research at all stages of its creation: recordings deposited during fieldwork, or at the end of a research project or a researcher's career (or life). To support longevity we have always written an XML file of the complete catalog entry to the item (the package of files) updated each time the catalog entry is updated, so each item is self-describing. We are now developing an innovative curation method using emerging technology (Oxford Common File Layout and Research Object Crate) to ensure the collection can be decoupled from a catalog and still be self-describing, a first step to ensuring longevity. 1

And, perhaps more important than any of this, digitization has allowed us to engage with the communities in the Pacific, Papua New Guinea (PNG), and South-East Asia that are the source of the recordings, which we can now return, sometimes half a century or more after they were made. 2

We address the following research questions. How can we ensure the records we create in the course of our research will exist into the future and remain citable (see Coble and Karlin in this issue)? How can our research data be made available for a wider public, most importantly for the people recorded and their descendants? How can we prepare our students for this new approach to curation of primary research data 3

so that they can build good methodology into their normal research practice, with much more productive outcomes? Finally, having built a collection of research data from projects that have ended, how can we manage this into the future? PARADISEC could be criticized for creating a collection that has no future, as there is no commitment from a national agency to the long-term carriage of the collection, but we suggest that we are not alone in this and that cases like PARADISEC need to be addressed as part of a strategic approach to research infrastructure. We have shown that it is possible to build the basis of future research by curating existing primary research material, and, in doing that, to provide access to research materials for the broader community. For each depositor their described and structured collection is typically in better shape than it was on their laptops or hard disks.

4 The loss of records when a research project ends is a major tragedy that needs to be avoided (see also Otis in this issue). For digital projects, the evanescence of the digital threatens imminent loss from any number of threats, such as power loss, physical breakdown of equipment, or network failure. For analog records, the senescence of unique media threatens loss by deterioration and is exacerbated by the lack of playback machinery. Much of the primary research data of linguistics, musicology, and anthropology from the latter part of the last century is still located on analog media and there is no more final ending to research data than its decay into unusability. There is an international effort to build language archives that are mainly based in Europe and North America, and linked by a network called the Digital Endangered Languages and Musics Archives Network (DELAMAN).[2] Further, the Open Language Archives Community[3] has developed a standard metadata set that these archives can use that allows for interoperability between language archives.

5 In Australia, hundreds of analog tapes were at risk of loss as no national institution had it in their mandate to collect tapes containing material from outside of Australia. Experts predict [NFSA 2017] that analog tapes will be largely unplayable by the year 2025 due to media breakdown and a lack of playback machinery, and, in many places, the task of preservation and repatriation of the content of these tapes remains to be taken up. Our research group, based at three Australian universities, addressed this challenge in the early 2000s by building a repository and digitization workflow that now includes 210 terabytes of material, representing 1,350 languages in 16,000 hours of audio recordings along with manuscript and video materials. Many of these languages represent small[4] communities of indigenous speakers in the Pacific, South-East Asia, and PNG, and are under-resourced (in number of online resources available), and so the imperative to ensure their records are findable and accessible is all the more urgent [Barwick & Thieberger 2018]. Linguists increasingly require citation of primary records for the purposes of research, and in order to build new kinds of materials on the original records (books, online presentations, dictionaries, and so on), and that is predicated on those records having persistent identification, provided by a suitable repository. Thus, while there is in our case a particular imperative to ensuring these records are properly made, described, and curated, the same principles can be applied to research in many disciplines.

6 In the course of doing this work, we train new researchers in appropriate methods for creating reusable primary records, and we have built a platform for citation of data, allowing for verification and licensed reuse of that data. We have done this training with several 1-year infrastructure grants over the 20-year life of the project. While there was an Australian National Data Service, it paradoxically did not curate data but served metadata and provided data storage with no guarantee of longevity. The current Australian Research Data Commons (ARDC) also has no national service for curating research data (and so, no “Commons,” its focus being on infrastructure for current research). Neither national service had or has within its mandate to implement an ongoing data curation service that would house primary research records into the future. So, despite large funding sources apparently available for research data, they are typically not available for curating the kinds of humanities and social science (HASS) records we are concerned with in PARADISEC, nor to make them FAIR.[5]

7 For our community of linguists, musicologists, and ethnographers, this lack of mandate is a particularly acute problem as our records are the result of fieldwork that has taken some effort to undertake, usually in remote areas in Australia, PNG, or the island nations of the Pacific [Thieberger & Barwick 2012]. We enter into relationships with people whose languages or cultural traditions we record, and we want to act responsibly by making the recordings available to them into the future. These are sometimes the only recordings and

perhaps the only presence of speakers of this language on the web. Furthermore, many of these small languages are at risk of loss, and some are no longer spoken due to many factors. So we have endangered records of endangered languages, compounding the responsibility of researchers to ensure they lodge records in a secure repository, or, if that repository does not exist, to build it, as we did with PARADISEC.

8 For our forebears it was not a simple matter to make recordings available to the communities they worked in. Analog tape could be returned to national cultural centres that had playback equipment, but there was no such capacity outside of capital cities. Clearly a colonial mindset was also at play, in which the records were taken as being the property of the recorder who typically made no provision for their long-term preservation [Thieberger 2020]. In the latter part of the twentieth century there was not the necessary institutional infrastructure to look after these recordings in Australia, with the result that many fieldtapes remained in the possession of retired researchers, or in the hands of their executors once they had died.

9 This situation motivated a collaboration of linguists and musicologists who were successful in obtaining a one-year infrastructure grant from the Australian Research Council in 2003. With the title “Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC),” we set the goal of digitizing 500 hours of recordings in the first year. We established a metadata schema to describe the files we created, conforming to DCMI[6] and the Open Language Archives Community (OLAC)[7]. With advice from the National Library of Australia and the National Film and Sound Archive, we bought reel to reel tape and cassette playback machines and employed an audio-engineer. Over time we built a lab with high quality analog to digital audio converters, a tape cleaning system, and a vacuum oven (required to re-adhere magnetic media to their carrier polyester tape backing) to deal with the range of conditions in which tapes were presented to us. With this foundation, we kept locating more and more tapes, and finding small grants to support continual digitization over the past 20 years. We have now digitized about 7,900 hours of analog tape, and house a further 8,000 hours of born-digital audio. Together with 2,600 hours of video and many text files, there are 420,000 files organized into 680 collections.

10 On accessing the collection, delivery versions of high resolution files are automatically created, for example, wav files have mp3 versions and tif files have jpg versions. A digital object identifier (doi) is assigned to all public collections, items, and files. Depositors assign rights, licensing the use of the files, which can include making items closed to all users except those nominated by the depositor (who can alter access conditions over time). They can also make items or collections private, which means no search engine finds them, and no doi is assigned, which is useful when a collection is being constructed. Once users are registered, they are then able to access any ‘open’ item, having first agreed to standard access conditions.

11 To maximize access to the collection, our catalog has several APIs that are picked up by national and international aggregators, including OLAC[8], Research Data Australia[9], and the National Library of Australia’s TROVE[10]. OLAC, in turn, presents its harvested records in single pages per language so that our metadata is then available along with any other material related to any one of the 7,000 or so of the world’s languages. And, of course, Google also picks up our catalog.

12 There are many locations in the world for which internet access is difficult, and often expensive, and so access to heritage materials in or near their source communities remains problematic. A solution we have tested in a few locations (a village in Vanuatu, an Australian Western Desert community, workshops in Honiara and in Papeete) is to load relevant items from the collection onto a small low-powered computer, known as a Raspberry Pi[11], which includes a wifi transmitter. We plug in a usb drive with a catalog of just those files in a simple html form that can be retrieved on a mobile phone, from a local signal transmitted by the Raspberry Pi, independent of any internet access. The catalog is created by a service we wrote, called dataloader[12], that harvests the XML file stored in each item and then creates the catalog of just that small collection. This catalog includes services that allow video files to be viewed, audio files to be heard, and pdf files to be scrolled. All files can then be downloaded to the mobile phone without being impeded by bandwidth considerations. In 2022/23 we plan to provide installations of these units to regional cultural centres (museums and similar agencies) in the Pacific so that files can be accessed locally. Recall that these are recordings made in the past few generations. They were analog tapes of people from villages in many locations, stored inaccessibly, and so they have not been available for the people recorded or their families.

Once these recordings are back in the relevant communities they can inspire revitalization of oral traditions and additional contextual information that can enrich our catalog.

What has been preserved?

13 While the total hours and size of the collection is impressive, it is the content of each of the files that gives a sense of the enormous value held by PARADISEC. For example, a major collection among those that initially motivated our project was that left by Arthur Capell, the professor at the University of Sydney in the 1950s, with tapes in nearly 300 languages[13], and thousands of pages of notes[14]. All of this was left in the house of Capell's executor, who undertook the massive task of managing and describing the tapes and papers. We have digitized this collection as some 1,300 items, and put it online. This includes over 15,000 pages of notes from a large number of Pacific and south-east Asian languages, together with many recordings on open reels and cassettes. In my own experience, in Capell's collection I have found a recording of a speaker of Nafsan, the language that I research in Vanuatu, and took that recording and associated notes back to the village where speakers live. The speakers can now also access all of this information via mobile phones themselves. So, a collection that was previously available only to visitors to the executor's house is now an openly available and licensed set of research and heritage material. The physical copies of these items are held by the National Library of Australia.

14 An unsolicited collection was offered to us by Ruth Roesler, who had no other suitable repository available to her. This work, created with her husband Calvin Roesler in the 1950s[15], was among the Asmat people of the south coast of what is now West Papua. It includes folktales, origin stories, documentations of customs, songs, descriptions of daily life and linguistic analysis. Items 001 to 054 were open reel audio recordings that we digitized and all following items are notes, most of which are transcripts of audio, and we have also put these into an online display[16].

15 We worked with the anthropologist Ian Frazer who had more than 200 cassettes, and 40 open reels recorded from 1971 to 1985, in fieldwork in North Malaita, Solomon Islands, mainly in To'abaita but with neighbouring groups as well (Lau, Baelelea, Pijin). These are recordings of music (traditional and contemporary), traditional stories, history, life histories, traditional and present day customs/culture, political history, labour history, and much else. This collection was located as a result of PARADISEC's "Lost and Found"[17] project and digitization was funded by an Endangered Languages Documentation Programme Legacy Materials Grant and the ARC Centre of Excellence for the Dynamics of Language.

16 An important aspect of all of this work is taking inaccessible records and preparing them for access, including licensing, and online curation. As noted in [Thieberger 2020], by archiving with PARADISEC, records made with indigenous people are available to those people as soon as possible after they are recorded. As [Jimerson 2007, 256] notes, archives contribute to the public interest "by documenting underrepresented social groups and fostering ethnic and community identities." Similarly, Smith observes that

> imperialism and colonialism brought complete disorder to colonized peoples, disconnecting them from their histories, their landscapes, their languages, their social relations and their own ways of thinking, feeling and interacting with the world. […] To discover how fragmented this process was one needs only to stand in a museum, a library, a bookshop, and ask where indigenous peoples are located. [Smith 1999, 28]

17 I suggest that the process of making records available addresses some of Smith's concerns and works to decolonize academic practice.

18 To represent the beauty of the sounds in many of our collections, we have written a soundscape that takes a curated set of samples of audio and displays it on a map, playing a sample together with the metadata[18]. This kind of view is possible because the collection has a predictable structure, includes geographic metadata, and automatically transcodes files to a deliverable mp3 on ingestion. The soundscape provides another way to discover the material in the collection. We also made a virtual reality exploration of a geographic landscape in which shards of light emanating from the ground represent each language, allowing

users to navigate between them, and hearing, in effect, a forest of languages as they pass through them.[19] In a similar vein, material from our collection has been used by the artist Lena Herzog in her work “Last Whispers.”[20]

19 A collection of research data is of relevance to academic research as it provides a citable basis, for example, for verification of claims made in papers. Humanities research data has an additional appeal in that is it is likely to be comprehensible to the general public, and to be of particular relevance to the people involved in its creation. One way in which we are publicizing the collection is by producing a series of podcasts in which users of items in the collection discuss their reactions to finding the material, and, in some cases, relearn songs or oral tradition from the earlier materials.[21]

How to prevent data endings

20 A repository can provide a solution for project endings for the primary data created in the course of research. Projects of fieldwork and analysis of language materials typically result in grammatical analysis, often a dissertation, or in publications that cite the primary data. Recently our discipline has focused more on citation of primary data to provide necessary context for analytical claims ([Thieberger 2016], see also Holmes and Takeda, this volume). At the same time, we have increasingly recognized our responsibility to prepare records for the speakers we work with. These two desiderata have led to an understanding of the need to create records in ways that allow them to be re-used, which, in turn, requires training for the researcher. We run regular training courses in the tools and methods that will result in well-constructed research corpora, emphasizing the importance of standard metadata descriptions that can then be imported into PARADISEC’s catalog.

21 In our experience, it is often only when arriving at the point of depositing records in an archive that a researcher consciously takes stock of their materials and prepares them in a structured format. At this point they may realize what remains to be done – which media have not been transcribed, which texts have not been annotated, which photographs are not identified, and so on. For most current researchers, this, in itself, is a satisfactory outcome. Regardless of the longevity of the archive, they are now able to find items in their own collections and to continue to build their research based on these well-structured data files. Of course, this management of research data is done more easily at the moment of creation of the records, and to assist with the task of building a well-structured collection from that moment, we have worked with colleagues[22] to develop a metadata entry tool called Lameta[23], now available for general use. Lameta is a standalone app that presents files on your computer for description and organizes structured metadata that can then be exported to formats accepted by a range of existing language archives. For a researcher to end a project gracefully they need to have good guidance in managing their records, and a tool like Lameta is part of that guidance, making it as easy as possible to provide structured metadata. To promote the use of this and other tools we regularly run training workshops, and discuss the workflow [Thieberger & Berez 2012] that takes records through from fieldwork, transcription, annotation, and to archiving.

22 An example of the new life of research data is my own collection of recordings made in Efate (Vanuatu) since the mid-1990s in the local language Nafsan. I prepared a set of data using the tools expected of my discipline, with outputs conforming to the standards required by PARADISEC (high resolution media, textual transcripts, structured lexicon, and so on). A guide to this collection is available as a webpage.[24] Some of this material was then prepared as a book of 70 bilingual stories, a dictionary produced both as a book and a phone app, and the corpus of audio and transcripts and texts has been used in four international projects examining particular linguistic features and each requiring a structured media and text corpus. A PhD student began work on analysing tense and aspect in the language by working through the texts, before going on to conduct her own fieldwork, and a postdoc worked in detail on aspects of the language that I had only briefly addressed. Several videos from my collection have been put onto Facebook pages by speakers of the language. None of this would have been possible if I had kept the records on my laptop, or lost them when the hard disk failed.

PARADISEC's endings

We have ensured the longevity of many research collections in PARADISEC, all the while expecting that there would be a national digital repository that would take responsibility for collections like ours. There was in 2003 no national Australian repository for digital HASS research of this kind, motivating the work that we undertook, and, unfortunately, this continues to be the case. The timescale that projects like ours operate under is governed by funding cycles, and national research infrastructure for HASS in Australia is typically no different, usually with three-year horizons of funding and no commitment to long-term curation of the research data that has been invested in so heavily by the taxpayer. While repositories for genomics, bio-informatics, or astro-physical data are designated as national research capabilities[25] whose future is assured by governments, not a single humanities data repository is granted that status. This is despite our collection, for example, being just a fraction of the size of any of these STEM repositories. There is an Australian national server program, and data storage is provided on a merit-based system which we have benefited from, to house both the online collection and a mirrored offsite copy. However, storage on its own is not enough to ensure the safety of a collection like ours: we provide a rich catalog and set of services to allow interaction with the data, including discipline-specific viewers for media and transcripts, citation forms for data, and feeds to international aggregators for this kind of material, using standard language identifiers to build resource guides for each language in the world. All of this contributes to the value of a curated repository, and is much more than simply storage on disk. Because of the lack of a national data service, we have continued to operate PARADISEC for 20 years, despite funding hiatuses. We have built an automated system that allows ingestion of new items with minimal handling, and continues to provide access as long as the servers are running. This means that we can operate with a skeleton staff when needed, and, when funds are available, we can continue the more labour-intensive work of processing analog tapes. Clearly, this requires a commitment by our personnel that goes beyond a normal employment contract, a commitment for which we are all very grateful, and which has allowed the collection to endure over two decades so far. 23

We have learned, in our efforts to uncover unique materials on closed-down servers or hard disks, that files on a disk are only part of the preservation story, and the very simple solution that this suggests is one that keeps data and metadata together. For the past decade each time we save a catalog item it writes an XML file to the collection, and any subsequent edits to that catalog item are saved to the same location, so that files and metadata are co-located. As a result, we were able to quickly take advantage of a small grant in 2019 to convert the catalog and collection to a new format, using Research Object Crate (RO-Crate) and inspired by a platform we called Arkisto[26]. This creates bundles of data that can be stored on disk with no external catalog and that still maintain their contextual information, metadata, licence information, and access conditions, and so stand a far better chance of surviving into the future than files in proprietary systems, or those using monolithic databases for their descriptions. As the metadata is written in json, a commonly used format, it is not difficult to re-create a catalog of any of these bundles, independent of whatever software was used to create them originally. 24

Future work on this promising development will include changing the user's interaction with the collection to a single page application that is built from the collection itself, with regular indexing creating the page. For the time being our catalog will continue to create the metadata that is then stored in the RO-Crate file with the collection object. The new catalog viewer already has improved ways of interacting with the collection. For example, any media that has a time-aligned transcript is playable together with that transcript, and the text of the transcript is searchable. This means that the text in the entire collection can be searched and the selected chunk of media can be heard or seen for each search result. This viewer can be deployed to other ways of delivering the collection, for example, on a Raspberry Pi for local wifi access. Such microservices can be added to the system as required, with the underlying structure of the collection in the standard RO-Crate format. This makes the catalog available for delivery in various forms, allowing for tailoring to different user groups if required (for example, localizing the interface or changing search terms depending on who the target users are). 25

Conclusion

There is an urgent need for national repositories to take up the challenge presented by the increase in the number of orphaned and fragile digital projects. When those projects have created some of the few records in small languages there is a greater imperative to curate those records and ensure their longevity. The great risk to a collection of this kind is a lack of perennity common to all “ended projects,” and we advocate for national data services that will capitalize on the investments already made in each research grant, each research project, and each researcher’s effort. 26

PARADISEC demonstrates a functioning repository that arises out of a disciplinary base. It models data management and provides training in data creation that benefits both the researcher and the repository, and as a result, the broader community can access curated and licensed items in the collection. PARADISEC has provided structure for collections that were previously disparate and undescribed. It has digitized analog materials and so made them more accessible. Copies of these collections are held by the depositors and, if permitted, by cultural or language centres relevant to the content of a particular collection. This, in itself, has been a valuable service provided by our project. 27

Ended projects have found a home in PARADISEC in a format that will endure as properly structured and described files on disk, even if the worst case eventuates and the current repository has no further funding. Our challenge as researchers, and the challenge for national research infrastructure planners, is how to make collections like PARADISEC’s continue to provide access into the future. 28

Notes

[1] I acknowledge that I work on the unceded lands of the Gadigal people of the Eora Nation, the Ngunawal, and the Woiwurrung. I thank two anonymous reviewers for comments that have improved this article, and thank the organizers of the Project Endings for their work in bringing such an interesting group together to discuss the critical issue of longevity of access to research data. I particularly want to thank all members of the PARADISEC team, in particular my co-founders of the project, Linda Barwick, with Amanda Harris, and all who have contributed in various capacities: Sander Adelaar, I Wayan Arka, Peter Austin, Corinne Bannister, Grace Barr, Stas Belkov, Rosey Billington, Steven Bird, Lauren Booker, John Bowden, Kevin Bradley, Georgie Burke, Brighde Collins, Linda Connor, Aaron Corn, Miriam Corris, Ashisha Cunningham, Emma Cupitt, Frank Davey, Hugh de Ferranti, Mark Ellison, Nick Enfield, Nick Evans, Bethwyn Evans, Cathy Falk, Haofei Feng, John Ferlito, Janet Fletcher, William Foley, Nick Fowler-Gilmore, Steven Gagau, Lauren Gawne, Amit German, Cliff Goddard, Geneva Goldenberg, Tina Gregor, John Hajek, Jeremy Hammond, Michael Homsey, Tom Honeyman, Stuart Hungerford, Kari James, Katie Jepson, Jodie Kell, Sam King, Prash Krishnan, Marco La Rosa, Vi King Lim, Ewan Maidment, Allan Marett, David Marett, Julia Miller, Liana Molina, Kylie Moloney, Diego Mora, Mark Mosko, Aashild Naess, David Nathan, Peter Newton, Rachel Nordlinger, Carmel O'Shannessy, Zephyr Pavey, Andrew Pawley, Murray-Luke Peard, Silvia Pfeiffer, Prashad Rajendra, Melody Ann Ross, Malcolm Ross, Alan Rumsey, Ely Ruttico, Ryan Schram, Jane Simpson, Robyn Sloggett, Juanita Sumner, Andrew Tanner, Nick Thieberger, Paul Trilsbeek, Jill Vaughan, Jacques Vernaudon, Michael Walsh, Nick Ward, Gillian Wigglesworth, and Aidan Wilson.

[2] https://delaman.org

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

[4] These could also be called endangered languages, but that would limit the focus of our work. While “endangered” is part of the name we established at the outset, it is probably not the most appropriate term to use to cover the many cultures and languages that continue to be spoken, even with few speakers, that are targeted by our work, hence my use of the term “small” language, following [Dorian 2014].

[5] ARDC and similar services all subscribe to the FAIR principles (https://www.force11.org/group/fairgroup/fairprinciples) of Findable, Accessible, Interoperable, and Reusable data, but, without a long-term repository with licenses for re-use, data can not actually be FAIR beyond the current funding cycle.

[6] https://dublincore.org.

[7] www.language-archives.org

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

[9] https://researchdata.edu.au

[10] https://trove.nla.gov.au/

[11] https://language-archives.services/about/pi/

[12] https://language-archives.services/about/data-loader/

[13] https://dx.doi.org/10.4225/72/56E7A74251D08 You must be logged in as a registered user to access the files in this collection.

[14] https://paradisec.org.au/fieldnotes/AC2.htm

[15] https://dx.doi.org/10.4225/72/56E824684C625

[16] https://paradisec.org.au/fieldnotes/ROES/web/serieslist.htm

[17] https://www.delaman.org/project-lost-found

[18] https://www.paradisec.org.au/Soundscape/index.html

[19] https://glossopticon.com

[20] https://www.lenaherzog.com/last-whispers

[21] https://www.paradisec.org.au/toksave-podcast

[22] At the Endangered Languages Archive (London) and at the University of Hawa'i' at Manoa, funded by the National Science Foundation, the ARC Centre of Excellence for the Dynamics of Language, and the Endangered Languages Documentation Programme.

[23] https://lameta.org, written by John Hatton

[24] https://www.nthieberger.net/sefate.html

[25] https://www.dese.gov.au/national-research-infrastructure/funded-research-infrastructure-projects

[26] https://arkisto-platform.github.io/

Works Cited

Barwick & Thieberger 2018 Barwick, L., and Thieberger, N. (2018) "Unlocking the archives", in V. Ferreira and N. Ostler (eds.) *Communities in Control: Learning tools and strategies for multilingual endangered language communities*. *Proceedings of the 2017 XXI FEL conference*. Hungerford: FEL. pp. 135–139.

Dorian 2014 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*, 6. Leiden: Brill.

Jimerson 2007 Jimerson, R. C. (2007) "Archives for All: Professional Responsibility and Social Justice", *The American Archivist*, 70(2), pp. 252–281.

NFSA 2017 NFSA. (2017) *DEADLINE 2025 Collections at risk*. Canberra: National and Film and Sound Archive. Available at: https://www.nfsa.gov.au/corporate-information/publications/deadline-2025.

Smith 1999 Smith, L. T. (1999) *Decolonizing Methodologies: Research and Indigenous Peoples*. London; New York: Zed Books; Dunedin: University of Otago Press.

Thieberger & Barwick 2012 Thieberger, N. and Barwick, L. (2012) "Keeping records of language diversity in Melanesia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)" in Evans, N. and Klamer, M. (eds.) *Melanesian languages on the edge of Asia: Challenges for the 21st Century*. LD&C Special Publication No. 5. Honolulu: University of Hawai'i Press. pp. 239–253. Available at: http://scholarspace.manoa.hawaii.edu/handle/10125/4567.

Thieberger & Berez 2012 Thieberger, N. and Berez, A. (2012) "Linguistic data management" in Thieberger, N. (ed.) *The Oxford Handbook of Linguistic Fieldwork*. Oxford: OUP, pp. 90–118.

Thieberger 2016 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* 32(2), 1 pp. 423–434. Available at: http://dx.doi.org/10.1093/llc/fqw006.

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