10

Language Documentation & Conservation Special Publication No. 5 (December 2012)
Melanesian Languages on the Edge of Asia: Challenges for the 21st Century,
ed. by Nicholas Evans and Marian Klamer, pp. 239–253
http://nflrc.hawaii.edu/ldc/sp05/
http://hdl.handle.net/10125/4567

Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

Nicholas Thieberger

University of Melbourne

Linda Barwick

University of Sydney

At the turn of this century, a group of Australian linguistic and musicological researchers recognised that a number of small collections of unique and often irreplaceable field recordings mainly from the Melanesian and broader Pacific regions were not being properly housed and that there was no institution in the region with the capacity to take responsibility for them. The recordings were not held in appropriate conditions and so were deteriorating and in need of digitisation. Further, there was no catalog of their contents or their location so their existence was only known to a few people, typically colleagues of the collector. These practitioners designed the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a digital archive based on internationally accepted standards (Dublin Core/Open Archives Initiative metadata, International Association of Sound Archives audio standards and so on) and obtained funding to build an audio digitisation suite in 2003. This is a new conception of a data repository, built into workflows and research methods of particular disciplines, respecting domain-specific ethical concerns and research priorities, but recognising the need to adhere to broader international standards. This paper outlines the way in which researchers involved in documenting languages of Melanesia can use PARADISEC to make valuable recordings available both to the research community and to the source communities.

1. Introduction. At the turn of this century, a group of Australian linguistic and musicological researchers recognised that a number of small collections of unique and

often irreplaceable field recordings mainly from the Melanesian and Pacific regions were not being properly housed and that there was no institution in Australia which would take responsibility for them. The recordings were not held in appropriate conditions and so were deteriorating and in need of digitisation. Further, there was no catalog of their contents or their location so their existence was only known to a few people, typically colleagues of the collector. These researchers designed the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), a digital archive based on internationally accepted standards and obtained Australian Research Council Infrastructure funding to develop an audio digitisation suite in 2003. This is a new conception of a data repository, built into workflows and research methods of particular disciplines, respecting domain-specific ethical concerns and research priorities, but recognising the need to adhere to broader international standards.

2. Background. Researchers (in particular linguists, musicologists and anthropologists) working with speakers of small languages (those with few speakers) typically conduct fieldwork to learn how aspects of these societies function, how the languages are structured, or how musicological knowledge is constituted, in addition to recording life stories, ethnobiological and other information. Typically these are minority endangered languages for which no prior documentation exists. This is vitally important work which often records language structures and knowledge of the culture and physical environment that would otherwise be lost (see e.g., Evans 2009, Maffi 2001, Harrison 2007). While it is typical for the interpretation and analysis of this data to be published eventually, the raw data is rarely made available. The data—tapes, field notes, photographs, and video—are often not properly described, catalogued, or made accessible, especially in the absence of a dedicated repository. This means that enormous amounts of data, often the only information we have on disappearing languages, remain inaccessible both to the language community itself, and to ongoing linguistic research.

The data that we create as part of our research endeavour should be reusable, both by ourselves and by others, and, in particular by the speakers and the general community with an interest in the nature of linguistic diversity in Melanesia. Beside the imperative to ensure there are good records of these languages this is also because any claims that we make based on that data must themselves be replicable and testable by others, and because the effort of creating the data should not be duplicated later by others, and provide a foundation that can be built on. In order to be made accessible, the data recorded by researchers must be properly collated and indexed for public presentation and archiving (see Himmelmann 1998, Woodbury 1998, 2003). However, until recently there has been no simple means for doing this and access to physical analog records can be difficult, if not virtually impossible, when they are stored in a single location.

This issue is being faced by scholars in many disciplines, and is being addressed under the rubric of cyberinfrastructure (National Science Foundation (U.S.), 2003) or ehumanities—how to build on existing knowledge and how to add new data that is being created in the course of various research projects so that the broader research community can benefit from it. This is all the more important when a linguist makes the only recordings for an endangered language–one that may no longer be spoken in the near future. Australia and its immediate neighbours are home to a third of the world's languages, most of which may never be recorded. Many of these languages could include completely novel structures

or ways of viewing the world, but each of them reflects the history of their speakers and is worthy of detailed recording. Melanesia in particular is among the most linguistically diverse regions (see Hammarström & Nordhoff this isssue), with Vanuatu having the highest density of languages per person of any country.

Significant resources are now being devoted to recording endangered languages in Europe (the Documentation of Endangered Languages project administered by the Max Planck Institute, Nijmegen) and the UK (Endangered Languages Documentation Project) and in the USA (the joint NSF/NEH program titled Documenting Endangered Languages). Furthermore, there are many local initiatives for recording oral tradition, like the fieldworker programme at the Vanuatu Kaljoral Senta or the collections being made by the Agence de Développement de la Culture Kanak (ADCK) or the Academy for Kanak Languages in New Caledonia. If the data arising from all of this effort is not properly safeguarded in our region it will represent a loss of cultural information, not to mention an enormous waste of effort and money. Many recordings are not described sufficiently to allow their contents to be discovered, and often there is little thought given to the methods involved in managing large multimedia datasets, which are especially vulnerable because they are in digital formats that are at risk (either due to lack of suitable digital data preservation and management infrastructure, or because of format obsolescence in a fast-changing digital media environment). Too much data is stored in ways that make it hard to access for the research community, let alone the broader community. Some research groups develop their own computational solutions which, admittedly, serve their needs well but which renders the group and their data isolated from the rest of the scientific community. The development of a new methodology, which includes the adaptation or development of new tools, must be grounded in application of that methodology to real data (Bird and Simons 2003). There are too many examples of 'proofs of concept' which set out directions for further work but which are not immediately applicable to any real-world problem.

3. Technology gap (the digital divide) and multimedia. It is a concern to some that we use increasingly technological methods for recording traditional practices, while the cultures in which they are embedded and the people who practise them have little access to the benefits offered by these technologies. How appropriate is it to use high technology, such as digital multimedia, with languages from villages that have no electricity? Of course, there is nothing new about the gap between the resources available to the researcher and those available to the researched, this is the colonial essence of any research project run by a first-world linguist. Suggesting that a video recorder is more colonial than handwritten notes (see for example Aikhenvald 2007) ignores the extractive nature of both forms of recording, and, more importantly, ignores the need for researchers to make the richest possible record for reuse by the speaker community. We should think in terms of what technology is appropriate for the task, and, in the case of recording oral tradition as the basis for both linguistic research and for heritage purposes, it is clear that we must use methods based in digital technologies (Bowden and Hajek 2006), because analog recording formats and equipment are all but obsolete (Schüller 2004).

The realisation that we can use multimedia data to enrich our understanding of performance is not something recent, and indeed goes back to the days of phonograph recordings, as this quote from Malinowski about his fieldwork in the Trobriand Islands illustrates:

> If I could, by a good phonographic record, counterfeit the living voice of Tokulubakiki: […] I should certainly be better able to translate the text in the sense of imparting to it its full cultural flavour and significance. Again, if by cinematographic picture I could reproduce the facial expression, the bodily attitude, the significant gestures, this would add another contextual dimension. (Malinowski 1935: 26)

While the technology to record and play back performances has been available since the late 1800s, it was rarely used by linguists until the second half of the twentieth century, and even then, analog recordings were difficult to create in the field, and later, and to access. It is only with the advent of digital media that we see the development of instant access to time points within large media corpora and the associated (but still painfully slow) realization among linguists that they can create reusable corpora in which their analysis can be embedded (Thieberger 2009). It is critical that a distinction is clearly made between archival forms of the media (held in high resolution files, such as 24-bit 96 kilohertz uncompressed audio, which are described in a catalog, and given persistent location and naming) and delivery or access forms of multimedia (which will be of lower resolution and often compressed for delivery via appropriate formats, such as the web or mobile phones). Multimedia presentations are seductive in their ability to relate parts of collections, linking texts to media or images and media to dictionaries. We have, however, seen enough examples of multimedia packages that are costly, contain relatively small amounts of information and become unplayable after a few years.

4. Access to digital data in the region. Williams (2002:15), in a report on the status of digital community services in the Pacific, noted that:

> [i]nformation on hardware resources […] shows that while all libraries, archives and museums that responded have access to at least a computer, the situation is bleak. Except for libraries in the Republic of Palau (and presumably in the Micronesian region) and university libraries and centres in the University of the South Pacific network, Fiji Institute of Technology, Fiji School of Medicine, National University of Samoa and University of Papua New Guinea, the computers are used by staff for work operations. In the Library Service of Fiji, there is no computer for public use, with only one computer in the library. The Suva Public Library is in a better situation. The Niue Public School Library, Tuvalu Culture Office and the Samoa National Archives also do not have computer access for students or members of the community.

It is clear from reports such as this (and from our own observation) that there is still a long way to go in the provision of digital information in small Pacific Island communities. Nevertheless, in the decade since Williams's report there have been unexpected advances in access to digital resources in even quite remote areas of the Pacific. Mobile phone technology has been taken up with enthusiasm, and has coverage in many previously unconnected locations, allowing remote use of both telephony and the 'mobile web' (See

Picture 1). The World Bank 'Rural Communication Project' (World Bank 2010) in PNG aims to significantly increase the number of internet users there, from the estimated current 50,000 mostly based in Port Moresby, and to increase coverage in rural district centers.

We can expect to see mobile phones taking over functions of portable computers in remote locations and so should also plan on building access to cultural collections using these technologies. The development of mobile phone dictionaries of small languages based on common formats of lexical databases (see, for example, the PARADISEC project Wunderkammer) can already provide online or local access to electronic dictionaries with sound and images. Similarly, new methods of streaming digital media allow for efficient delivery of ethnographic recordings over low bandwidth, including mobile phones. The PARADISEC project EOPAS streams audio or video recordings of stories over the internet together with text (see the discussion below) using HTML5 and open-source media. HTML5 is an emerging web standard that allows streaming of multimedia within the stamdard web page, thus obviating the need for users to install additional software or plugins (Pfeiffer 2010). All of this indicates that creating proper forms of recordings, images and so on that conform to accepted archival standards will allow them to be transformed into delivery formats appropriate to the context in which they are to be used.

Figure 1. Publicity billboard for internet access via mobile phones (Port Vila, June 2011). Photo by Nick Thieberger

5. Ethics of information provision. In addition to the question of equitable access to the kind of cultural information that is now becoming commonplace on the internet, there is the more complex issue of the sensitivity of archival records being reintroduced in new contexts. Recordings made in the 1950s may take on a considerably different meaning when used today, especially if there are land disputes that otherwise rely on oral accounts remembered by the current generation. The archival record can assume an authority (whether justified or not) that may be advantageous to some in the present dispute, but detrimental to others. While those running an archive can be aware that such problems may arise, it is impossible for them to know such details for all of the locations from which the archive stores material.

In most societies there is some kind of protocol in place for access to certain kinds of information. Not everyone can read the records of company meetings, for example, or of secret government business. In smaller societies, such protocols may include access to songs or stories that relate to the first creation of the land or to the travels of ancestral beings: see for example Lindstrom (1990) on what he terms 'the economy of knowledge' in Tanna, southern Vanuatu. The provision of such information from an archive may subvert the very power structures that promote the ongoing use of traditional languages and clearly this is a potentially difficult situation for an archivist to find themselves in. The Endangered Languages Archive at SOAS has been working on a system for allowing more fine grained access conditions to be specified, including, for example, the ability for people other than depositors to determine who can access the recordings of themselves speaking. However, our present focus has been on preservation of the records we have located and we consider it more important that the material be stored for later reuse than that the safer option (that there be no archival record) be adopted.

6. Implementation of PARADISEC. In the initial phase of the PARADISEC project (2003) we established a steering committee with representatives of each of the partner universities (initially Sydney University, the University of Melbourne, ANU, and later UNE). The director of the project is Linda Barwick at the University of Sydney.

With invaluable technical support from both the National Library of Australia and the National Film and Sound Archive and with funds from the Australian Research Council we bought a Quadriga digitisation suite and employed an audio engineer and administrative assistant, based at the University of Sydney. We also built a vacuum chamber and low-temperature oven to allow us to treat mouldy tapes that required special care before being playable. Tapes stored at the ANU were identified and located and then permission was sought from the collectors or their agents to digitise and accession them into the collection.

In the first year of funding we had to come up with outcomes that would justify further funding grants and we aimed for 500 hours of digitized tapes in that first year (we achieved this goal in ten months). We wrote a catalog database in Filemaker Pro, aware that it would provide us with an immediately usable tool that would ultimately have to be converted to an online database. This database allowed us to refine data entry forms and controlled vocabularies without relying on a programmer. This first catalog worked well and exported to the XML files required for inclusion as headers in Broadcast Wave Format (BWF) files, and also exported to a static repository for Open Archives Initiative harvesting via the Open Language Archives Community harvester.

Files generated by this system (at 96khz/24 bit) are large, around 1.5 Gb per 45-minute

side of a cassette, and so require dedicated storage facilities. We established a tape backup system which ran periodically to copy files from the hard disk to storage tapes, but were fortunate when the Australian Partnership for Advanced Computing (APAC) designated PARADISEC a 'Project of National Significance', allowing us to use their mass data storage system, with considerable storage space provided to support our work. They further provided programming support by writing specialized software (called 'Babble') which provides weekly, monthly and quarterly reports on the state of the collection, as well as nightly querying the server in Sydney and copying files that are ready for archiving.

Data is organized by collector, but also by the internal logic of the collections (the same collector working on two different languages will have two collections, or a collection of video may be distinct from a collection of still images). The collection-level also speeds up a user's typing into the catalog as common fields from the collection level can be inherited down to the item level. Our naming convention is rather simple ('CollectionID'-'ItemID'-'FileID'.'extension') and it also provides the hierarchical file structure into which files are placed and stored on the server (with directories corresponding to the collection level and subdirectories corresponding to the item level).

Subsequently and with funds from the Australian Research Council Linkage Infrastructure Equipment and Facilities (LIEF) programme, we built digitisation suites in Melbourne and Canberra, allowing us to preserve important heritage tape collections such as those shown in table 2, by no means an exhaustive list. Without a dedicated infrastructure to describe, manage and store this material it would simply be lost.

Mark Durie (Acehnese, Indonesia)	Cindy Schneider (Apma, Vanuatu)
Barry Alpher (Cape York, Australia)	Sébastien Lacrampe (Lelepa, Vanuatu)
Sander Adelaar (Selako, Indonesia)	Stephen Morey (Assam, India)
Sebastian Fedden (Mianmin, PNG)	Robyn Loughnane (Oksapmin, PNG),
Amanda Brotchie (Tirax, Vanuatu)	Nick Thieberger (South Efate, Vanuatu)

TABLE 1. Examples of collections from Australia and its the region that have either been digitised by PARADISEC or accessioned as digital data by PARADISEC

Now that many researchers are recording directly to digital formats, we provide advice and guidance on suitable formats and workflows to facilitate ingestion into the repository. On return from fieldwork, depositing in PARADISEC provides a means of secure backup of researchers' otherwise vulnerable digital media files. We still have a need for digitization of older analog collections, a much slower process to produce a high quality digital preservation master file for archiving (International Association of Sound and Audiovisual Archives (IASA), 2004).

7. LICENSING USE OF ITEMS IN THE COLLECTION. The primary aim of the project to date has been on preservation of unique cultural records. Including a licence, or information about how each item can be used, is critical to the establishment of a properly curated collection because without it there is no way of providing access. Each depositor must fill

out a deposit form specifying any conditions that may apply to the material. We provide a default set of access conditions which any user must agree to prior to being given access to data, and depositors can choose to allow this set of conditions to govern their collection, or to determine their own conditions. We are presently investigating the use of Creative Commons licences as a less restrictive and more standardised form of agreement (Newman 2007, Seeger 2005).

8. Delivery of archival material, page images and dynamic media. We provide material from the collection to those authorized to receive it, typically in the form of downloadable files, however we have also worked on specific methods for the online delivery of two kinds of material – page images and time-coded media. We made available images of 14,000 pages of fieldnotes (see figure 2) from three deceased researchers using the Heritage Document Management System with a digital camera rig that we took to the home of the estate's executor, or to the office in which the papers were stored. These notes from deceased researchers would otherwise have only been available in a single physical location. As we do not have the resources to keyboard all of these manuscripts the images are stored in the collection with sufficient contextual metadata to make them discoverable on the web. As noted earlier, the archival version of each image is stored separately from the representational version.

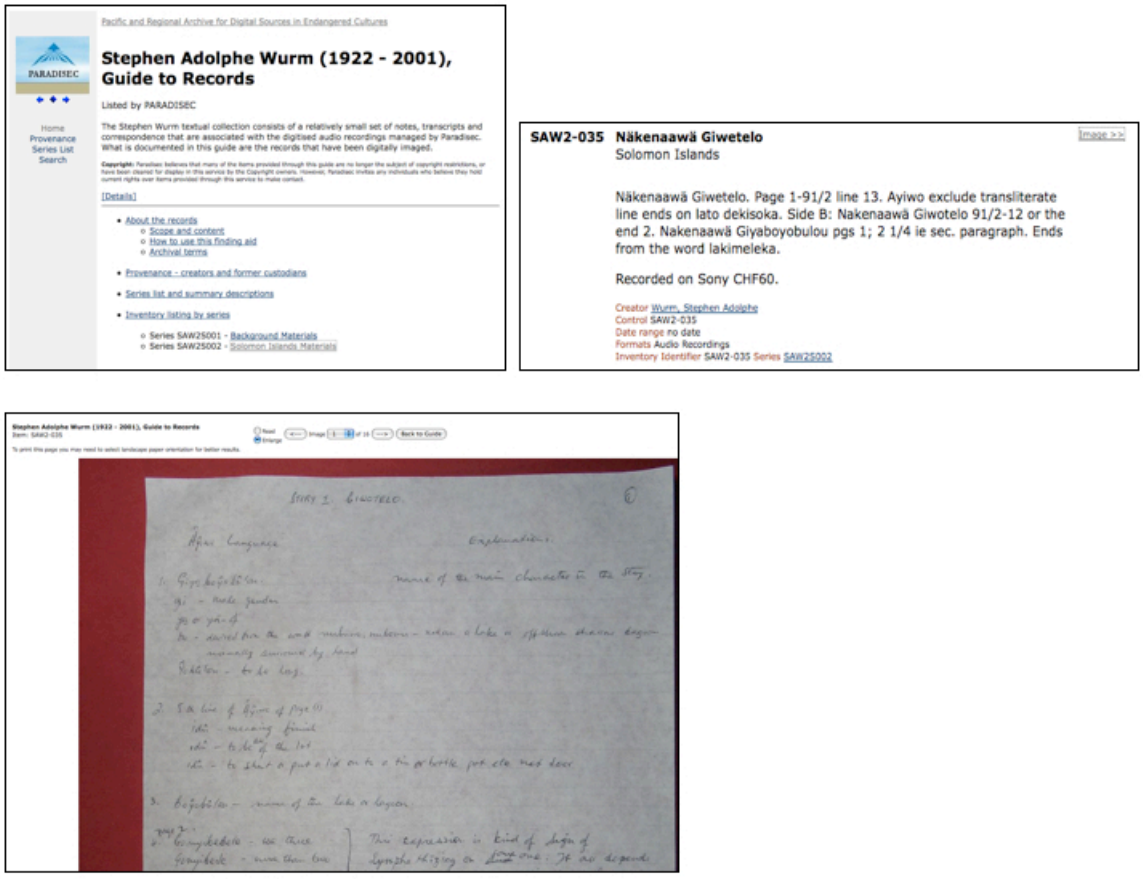

FIGURE 2. Page images from the Wurm collection of online manuscripts showing finding aids from the highest level (top left), to the item level (top right) and finally the image itself (bottom) (http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm)

9. THE ETHNOER ONLINE PRESENTATION AND ANNOTATION SYSTEM (EOPAS). While building a method for working with our own data we consider it important to create generalisable models and structures for others to use, and to engage in discussions and training sessions both in order to refine our methodologies and to impart new ideas. An example of such development is our work on the online presentation of interlinear glossed text together with recorded media (EOPAS), allowing material from any language to be heard in concert with its transcript and translation (Schroeter and Thieberger 2006). A number of tools for annotating language data have been produced recently and it is clear that more are envisaged now that several large projects are engaging with these issues in the USA, UK, Germany and the Netherlands. Annotation is a basic task that is undertaken following recording, and now it is typically carried out with time-alignment, meaning

that the text has references to timepoints within the media file (using software such as Elan or Transcriber) and can take several forms, the most common of which, for linguists, is interlinear text. These texts are analysed and parsed by a glossing tool that produces parallel lines of text, word translation and grammatical information, together with a free translation. These texts are then input into EOPAS, a schema-based XML system for making explicit the relationship between parts of interlinear texts together with links to the source media (see figure 3) which allows searching and concordancing linked directly to the media. EOPAS is portable (the source code is freely available), allowing other initiatives to capitalise on the work and potentially develop it in different directions. The ultimate aim of this approach is to allow new perspectives on the data itself, provided by contextualised access to primary data, and then to allow new research questions to be asked, and richer answers to be provided, all in a fraction of the time that it would have taken with analog data.

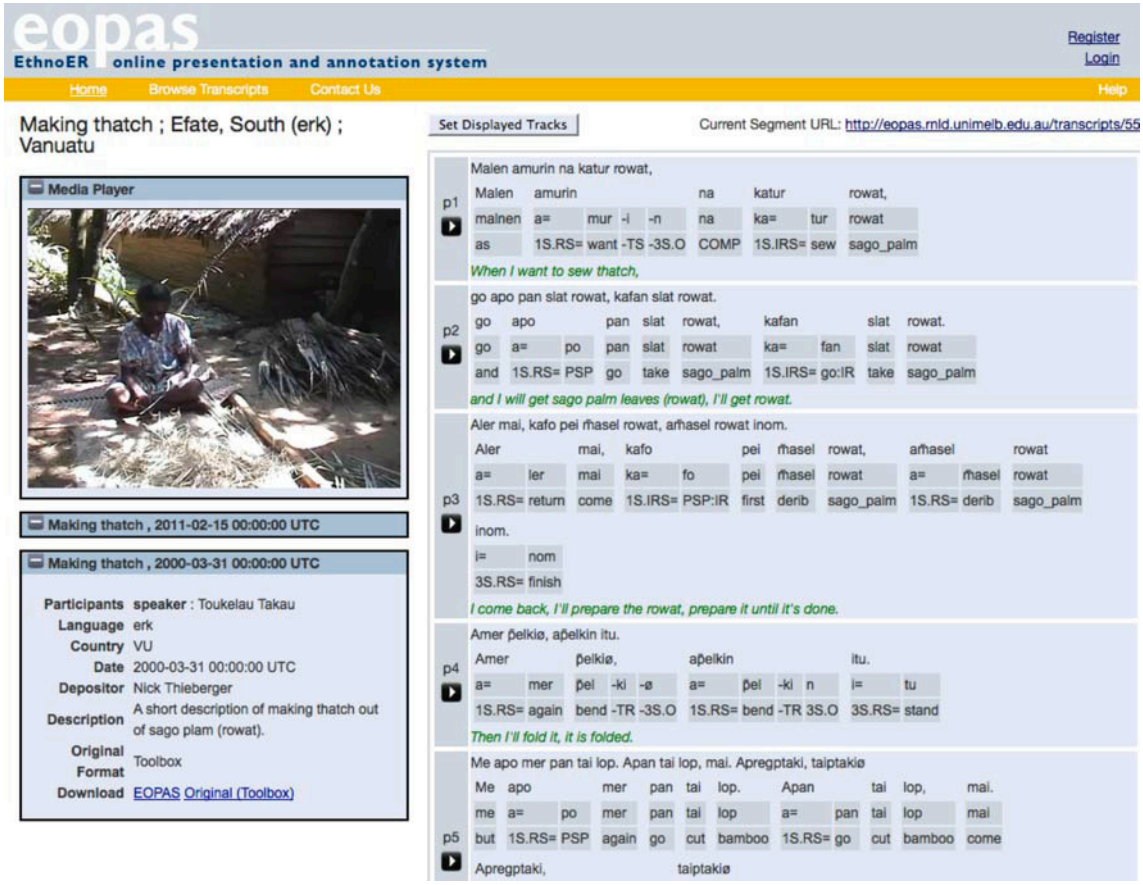

FIGURE 3. Example of a video clip with time-aligned text as presented in EOPAS.

10. Current status of the PARADISEC collection. Currently (late 2011) PARADISEC contains 7,226 items made up of 48,606 files totaling 5.2 TB, with just over 3,046 hours of audio data. Digital video already makes up an increasingly significant part of the collection. We hold data representing 650 languages from 60 countries (see examples of the kinds of collections in table 1) which is organized into 163 collections, some 85 of which represent new fieldworkers who have deposited material on their return from fieldwork (and one during the course of her fieldwork), thus providing a citable form of their data for their own research. This means that in their dissertations and publications they can refer precisely to the relevant linguistic data through citing the timecodes associated with the persistent identifier (web location) of their recordings in the PARADISEC collection. Citation of primary data is a critical step in conducting new research based on that data. The remaining collections are digitised from recordings made since the 1950s. The provision of this service requires ongoing support and negotiation with depositors and we have found that a key to establishing the collection has been the depositors' perception of the benefit accruing to them and to their data in having it well described. In addition, there are collections we know about and would dearly love to digitise but we do not have the resources to do this work. These include large audiotape collections at radio stations around the Pacific, many in local languages, and collections in regional cultural centres that do not have any local equipment to digitize their collections. Further, we are regularly approached by former colonial patrol officers or missionaries who have recordings, notes or photographs that they want to preserve.

Arthur Capell	1950s Pacific and PNG (114 tapes and 30 archive boxes of fieldnotes)
Tom Dutton	1960s onwards, PNG, 295 tapes
William Foley	1970s, PNG, 34 tapes
John Harris	1960s, Kiwai, PNG, 75 tapes
Don Laycock	1960s, PNG, 98 tapes
Al Schütz	1960s onwards, Vanuatu, six tapes
Stephen Wurm	1970s Solomon Islands tapes (~120 tapes and transcripts/ fieldnotes)
Bert Voorhoeve	West Papua, 180 tapes

Table 2. Example collections that have been digitized, described or curated by the PARADISEC project.

We have published on our website a detailed description of our workflow, developed over seven years of operation, that describes the various processes involved in locating tapes and then assessing, accessioning, digitising and describing them, managing the resulting data and metadata, and the return of original tapes. PARADISEC has been cited as an exemplary system for audiovisual archiving using digital mass storage systems by the International Association of Sound and Audiovisual Archives and, in 2008, won the Victorian Eresearch Strategic Initiative prize for humanities eresearch.

Once we built the infrastructure for a research repository, including the catalog, file system and naming conventions, it has been taken up by those researchers who are aware of the need to describe and preserve their research material. Often it is only in the process of depositing with PARADISEC that a collection is first described in a systematic way – one that then allows the description to be searched by Open Archives Initiative search engines (and also google). Every eight hours the PARADISEC catalog is queried by a service run by the Open Language Archives Community (OLAC) and any new or edited catalog entries are copied and made available to their aggregated search mechanism. Similarly, because the catalog complies with relevant standards, the Australian National Data Service (ANDS) has been able to incorporate our 163 collections into its national search mechanism. The quality of the metadata we provide ensures that targeted searches by language name can be resolved without locating similar but irrelevant forms.

11. Regional links and training. While the initial focus for our collection was the region around Australia (as suggested by the name we chose at the outset of the project), it has become clear that we need to accept material that has no other place to be archived. Typically, this means supporting Australian researchers whose research is outside of Australia, with the geographic spread of material we house now extending from India, into China, and across to Rapanui (Easter Island). With limited resources PARADISEC has nevertheless established working relationships with cultural centres in the Pacific region (e.g., the Vanuatu Kaljoral Senta, or the Institute of PNG Studies) which have involved providing CD copies of relevant material and, in the case of the University of New Caledonia, cleaning and digitising old reel-to-reel tapes in Drehu. A serious concern for many such agencies in the region (as observed in Williams' report, above) is the lack of continuity in funding and in staffing, with the potential result that collections established and curated over time may be at risk. We would like to be able to digitize the many hours of tapes held, often in less than ideal conditions, in countries of the region. We have begun an occasional mass backup of significant collections of digital material from the Vanuatu Kaljoral Senta and would like to extend this as a service to other agencies.

We regularly offer training workshops in linguistic research methods, including the use of appropriate tools and recording methods and in data management for ethnographic field material. This is extremely important, as the more informed the research community can become about the need for reuse of primary data, the more likely they are to be creating well-formed data that needs no extra handling by PARADISEC to be accessioned into the collection. Such training has been offered at community Indigenous language centres as well as in academic settings.

We cooperate in two further initiatives for disseminating information. The first is a blog (Endangered Languages and Cultures) and the second a resource website with FAQs and a mailing list (the Resource Network for Linguistic Diversity). Because of the rapid changes in methods for recording, transcribing, and analysing human performance no one can keep completely up to date, so these web-based resources are widely quoted and appreciated by the community of researchers.

12. The future of the collection. As the value of data curation becomes clearer and the use of the collection increases, we will see more theoretical work based on properly curated archival material. We have already seen linguists retrieving what are now historical language records for use in comparison with current usage and for analysis of language change. Serendipitous discoveries in the collections have included the drama specialist Diana Looser finding a performance of Albert Toro's 1977 radio serial, Sugar Cane Days, a historical drama about the 'blackbirding' days of indentured Kanak labour in the Queensland canefields. While discrete sections of Toro's play had been published in local literary anthologies and magazines in the early 1980s, no complete script of the play was available. Tom Dutton had recorded the complete five-part performance taken in Port Moresby in the 1970s, as well as an interview with Toro about the inspiration for, and genesis of, the play. These unique sound files allowed Looser not only to listen to the original radio play in performance, but to create a verbatim transcript from the recording.

PARADISEC is a project ahead of its time and so suffers from a lack of vision among funding agencies. It is truly collaborative, multi-institutional and cross-disciplinary which, despite frequent funding-agency rhetoric to the contrary, weighs against it being supported through normal research funding sources.

We would like to extend the streaming server we have established to allow delivery of any accessible material in the collection. We are also in the process of developing an access system with authentication and authorization of users.

PARADISEC is part of several international networks of similar projects (DELAMAN or OLAC, cited above), but is a leading exponent of linguistic data curation even among that field. Australian government moves to establish a national digital data service (a system of repositories hosting digital data in the way that PARADISEC has done) are still in their early stages, but we are confident that PARADISEC will become part of such a service within the next decade. Our unique collection needs to be safely shepherded through the intervening period, identifying more collections in need of digitisation, accessioning them, and providing the infrastructure for current researchers and postgraduate students to describe and preserve their field recordings. We need to continually provide training and advice for researchers in order that their outputs can be accessioned with minimal extra handling. Research that is conducted without an awareness of appropriate data structures and formats will result in poor outputs that need to be converted, often with considerable effort, to make them archivable. It is unlikely that this arduous conversion effort will be resourced and so we risk losing primary research data.

13. Conclusion. PARADISEC is a practice-based archive, arising from a community of practice who recognised that it was part of our professional responsibility to ensure that the records we create are properly curated into the future. This is a new conception of a data repository, accessioning primary research in the course of fieldwork or shortly after, and building methods and tools to facilitate its deposit and curaton. It is unique in its links on the one hand to fieldworkers and to speakers of Indigenous languages and on the other hand to the cutting-edge technologies of Web 2.0 and HTML5.

PARADISEC has been active in locating records of small languages and making them available for longterm access. We have been particularly aware of the needs of small language communities, especially those in PNG and island Melanesia. In 2012 we

have collaborated with the Solomon Islands Museum and Archives to apply for funding to digitise their audio collections. Similar collections of audio, film and video exist in agencies across the Pacific and are in need of urgent attention. Our new catalog will make streaming media available for viewing on a variety of platforms, including mobile phones, and this should allow delivery of these unique resources to their source communities. PARADISC is keen to attract more funding, so as to locate and digitise more material, and provide training to speakers to create their own records now. We could also increase the representation of languages in our EOPAS system to provide online samples of as many languages of the region as possible. There is much more to be done, but the work done by PARADISEC will allow future work to grow on good foundations.

References

Aikhenvald, Alexandra. 2007. Linguistic fieldwork: setting the scene. *Sprachtypologie und Universalienforschung* 60(1). 3-11.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557-582.

Bowden, John & John Hajek. 2006. When best practice isn't necessarily the best thing to do: dealing with capacity limits in a developing country. In Barwick, Linda & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 45-55. Sydney: Sydney University Press.

Evans, Nicholas. 2009. *Dying words: Endangered languages and what they have to tell us*. Maldon, MA: Wiley-Blackwell.

Harrison, K. David. 2007. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. New York: Oxford University Press.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161-195.

International Association of Sound and Audiovisual Archives (IASA). 2004. *Guidelines on the production and preservation of digital audio objects (IASA-TC04)*. Aarhus, Denmark: International Association of Sound and Audiovisual Archives.

Lindstrom, Lamont. 1990. *Knowledge and power in a South Pacific society.* Washington: Smithsonian Institution Press.

Maffi, Luisa (ed.). 2001. *On biocultural diversity: Linking language, knowledge and the environment*. Smithsonian Institution, Washington, D.C.

Malinowski, Bronislaw. 1935. *Coral gardens and their magic*. Vol 2. London: Allen and Unwin.

National Science Foundation (U.S.). 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: Office of Cyberinfrastructure, National Science Foundation.

Newman, Paul. 2007. Copyright essentials for linguists. *Language Documentation & Conservation,* 1(1). 28–43.

Pfeiffer, Silvia. 2010. *The definitive guide to HTML5 video*. New York: Apress.

Schroeter, Ronald and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Barwick, Linda and Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 99-124. Sydney: Sydney University Press.

Schüller, Dietrich. 2004. Safeguarding the documentary heritage of cultural and linguistic diversity. *Language Archive Newsletter* 1(3). 9-10.

Seeger, Anthony. 2005. New technology requires new collaborations: Changing ourselves to better shape the future. *Musicology Australia 27*(2005-6). 94-111.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Epps, Patricia & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389-408. Berlin & New York: Mouton de Gruyter. http://repository.unimelb.edu.au/10187/4864.

Watson, Amanda H. A. 2010. Communication and culture: mobile telephony in PNG villages. Paper presented at the Asian Media Information and Communication Centre: Technology and Culture: Communication Connectors and Dividers, 21-23 June 2010, Suntec City, Singapore (AMIC). http://eprints.qut.edu.au/32787/. (28 July, 2010.)

Whimp, Kathy & Mark Busse (eds.). 2000. *Protection of intellectual, biological and cultural property in Papua New Guinea*. Canberra: Asia Pacific Press.

Williams, Esther. 2002. *Digital community services: Pacific libraries and archives.* UNESCO.

Woodbury, Anthony. 1998. Documenting rhetorical, aesthetic, and expressive loss in language shift. In Grenoble, L.A. and L. J. Whaley (eds.), *Endangered languages: language loss and community response*, 234-258. Cambridge: Cambridge University Press.

Woodbury, Anthony. 2003. Defining Documentary Linguistics, address given at the Annual Meeting of the Linguistic Society of America, Atlanta, Georgia, on January 3, 2003.

World Bank. 2010. Remote Rural Communities in Papua New Guinea to Benefit from Improved Access to Telecommunications. http://go.worldbank.org/DHJ6XJO2O0 (28 July, 2010.)

Nicholas Thieberger
thien@unimelb.edu.au

Linda Barwick
linda.barwick@gmail.com