Oral Tradition, 28/2 (2013):253-260

Curation of Oral Tradition from Legacy Recordings: An Australian Example

Nick Thieberger

Introduction

Hundreds of hours of ethnographic field recordings and their associated oral tradition were destined to be lost until the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC, http://paradisec.org.au) was established in 2003 to digitize and curate this legacy made by Australian academic researchers since the 1960s (Barwick and Thieberger 2006; Thieberger and Barwick 2012).[1] These recordings in the languages of the region around Australia (broadly speaking, an area that includes Indonesia, Papua New Guinea [PNG], and the Pacific Islands) have high cultural value and are often the only records in these languages. Many languages in this region are spoken by few people and are in danger of being lost because of the pressure from neighboring languages or metropolitan languages such as Indonesian, Tok Pisin, English, or French, and so the records made a generation or more ago become all the more valuable. However, despite their unique heritage value, these recordings were not eligible to be preserved or curated by any existing Australian collecting institution.

A group of linguists and musicologists planned PARADISEC and sought advice from relevant agencies (in particular from the National Library of Australia and the National Film and Sound Archive). This advice was particularly valuable in allowing us to determine appropriate metadata standards (we use Dublin Core and Open Archives Initiative metadata terms as a subset of our catalog's metadata) and to understand the more hands-on requirements of cleaning and repairing moldy or damaged analog tapes. We then applied for and received infrastructure funding from the Australian Research Council. With a grant that was to last for just one year, we had to build a successful archive prototype that could then attract further funds.

Over the decade during which it has been running, PARADISEC has digitized several thousand hours of analog recordings in three ingestion units based at each of the participating universities: the University of Sydney, the University of Melbourne, and the Australian National University. We have also broadened our scope to include any relevant material that needs preservation, regardless of the geographic area it represents or the state of endangerment of the

[1] Thanks to Linda Barwick and Amanda Harris for discussions that have improved this paper. The work described here was partially funded by the Australian Research Council (grant DP0984419). Thanks to the Department of Linguistics at the University of Cologne for hosting me during 2013 and to the Alexander von Humboldt Foundation for awarding me a Ludwig Leichhardt Jubilee Fellowship.

languages involved. In 2011 we initiated an online survey[2] to locate further endangered analog collections and to work with their custodians in order to find funds to digitize and curate them before they are lost.

What Is in the Collection?

The contents of the various collections range from hundreds of recordings on a particular language made in the course of extensive fieldwork all the way through to isolated, short examples recorded opportunistically in a language. The records themselves range from narratives through to sung, chanted, and spoken performances as well as instrumental music. The collections from the 1960s and 1970s typically represent the work of deceased or retired scholars, so there is usually limited contextual information to include in the catalog. Occasionally there are handwritten transcripts of these recordings that we have included as scanned TIF or PDF files. These legacy collections include: Professor Stephen Wurm's several hundred tapes, with 120 Solomon Islands tapes and transcripts/fieldnotes from the 1970s (some of which have been used in later research by Åshild Næss [2006]); the ethnographer Roderic Lacey's collection of 118 tapes from the early 1970s used as the basis for his work on "Oral Traditions as History: An Exploration of Oral Sources among the Enga of the New Guinea Highlands"; James Weiner's collection of some 100 cassettes in the Foi language of Highlands PNG, the basis for his work on poetics in the language; Arthur Capell's 114 tapes from the Pacific and PNG from the 1950s (and 30 archive boxes of fieldnotes of which we have placed 14,000 page images online[3]); Bert Voorhoeve's 180 tapes from West Papua (mainly in Asmat) from the late 1960s; and Tom Dutton's 295 PNG tapes from the 1970s. Currently in our accession queue is a collection of recordings made by the anthropologist Ted Schwartz during his fieldwork with Margaret Mead on Manus island in the 1950s.

PARADISEC is making information available in an ethically appropriate way, and we have established working relationships with agencies in our region such as the Vanuatu Cultural Centre, the Institute of Papua New Guinea Studies, the University of French Polynesia, and the University of New Caledonia, among others. In 2013 we applied for funds with the Solomon Islands Museum to digitize hundreds of tapes they hold in Honiara. We have started a crowdfunding campaign to try to raise the funds necessary to do this work[4] and to locate more endangered collections of analog recordings.

Since building the necessary online tools for entering cataloging information for this kind of material, we have had a number of born-digital collections deposited. It is particularly interesting for scholars to be able now to deposit their records directly from the field or soon after their return from fieldwork. In this way they have a safe copy of their primary records and are able to cite those records with the persistent identification provided by an archive. Archiving

---

[2] http://www.paradisec.org.au/PDSCSurvey.html

[3] http://paradisec.org.au/fieldnotes/AC2.htm

[4] http://paradisec.org.au/sponsorship.htm

before the analysis makes the research grounded and replicable, and it turns on its head the more traditional approach of archiving primary recordings only at the end of one's research career.

The value of making the collection as discoverable as possible was made clear when we had a request from Diana Looser, then a Ph.D. candidate in Theatre at Cornell University in the United States who was writing a dissertation on Oceanic theater and drama. She needed access to a play that was listed in our catalog but existed nowhere else that she could find: in his collection, the linguist Tom Dutton had included a tape of playwright Albert Toro's *Sugarcane Days* recorded from ABC radio Port Moresby.[5] Looser transcribed the tapes and prepared the only extant version of the script that she then redeposited in the collection, a sample of which has been reproduced below:

> [RECORDING TD1-PO2179-A 00:00–25:45]
> *THE SUGAR CANE DAYS*, EPISODE I: THE MASSACRE
>
> *[Theme music up and under.]*
>
> ANNOUNCER: The National Broadcasting Commission presents *The Sugar Cane Days* by Albert Toro. This is Episode One, "The Massacre." *The Sugar Cane Days* is set in the period known throughout the Pacific as the "blackbirding days" or the "kanaka trade." That was the period when the cruel practice of forced labor was a near relation to slavery, but handsomely disguised under the polite name, "labor trade." The period is between 1863 and 1907, when the human being market was at its height. This story is based on facts that have almost become legends. Molen, a victim of those days, died in 1976. Here is Molen's story.
>
> MOLEN *[remembering]*: I am a very old man now, you see. I have lived a long life, and it will soon be time for me to die. From those who returned after the end of our contract in the plantation, I am the last man alive. Mally Bulla was everyone's favorite plantation; it is this story I want to tell you, and later, when you are a father of many children, you can tell it to them; and their children will tell their children's children. You are a lucky man today. I hear and see cars, trucks, and ships driven by engines *[…]*
>
> *[Theme music fades; cross-fade into sound of waves on the shore, and roosters crowing, chickens clucking.]*

This re-use of research material in new ways can only be achieved if that material is stored in accessible locations with licenses for use in place and with a catalog that provides sufficient information to allow it to be located.

---

[5] Registered users can hear the first of the audio files of this performance at http://catalog.paradisec.org.au/collections/TD1/items/P02179/essences/1019890.

Technical Features

We began by installing a Quadriga analog-to-digital workstation and developing a system architecture that included data storage and backup, naming conventions, a metadata schema, a workflow for identifying eligible recordings (assessing their physical state and contents), deposit and access conditions, and a catalog. This catalog presents a set of metadata elements to the user with drop-down menus to enforce standard forms, in particular for terms that are exposed to external harvesting tools to allow remote searching of the catalog. These terms include country names (ISO 3166-1), language names (ISO-639-3), and datatypes, among other elements.

The online catalog has been redeveloped over time in response to users' comments. It currently exports a feed that is harvested by the Open Archives Initiative, the Open Language Archives Community, and the Australian National Data Service, all of which helps make items in the collection more discoverable. Each item in the collection has its own deposit conditions, but some 5,000 items (out of 8,100) can be seen or listened to online by registered users—those who have agreed to the conditions of use and registered their email addresses. The remaining items require some kind of permission from the depositor, but we are working with depositors to reduce the number of items in that category.

The structured metadata required by our catalog makes the depositor provide rather basic information that may not previously have been compiled, including for each item a title, date of creation, language spoken, and country in which it was recorded. Further information includes: the role of participants, the language name as it is known locally (which may vary from the standard form), the type of information (lexicon, song, narrative, and so on), geographic location (given by a bounding box on a map), and a free text description of the item that can be as rich as the depositor wants. All of this information can be improved on by subsequent users who use the collection in their own research projects (as we saw above with the item from Tom Dutton's collection).

Transcription

A media recording with a transcript is more useful than a recording on its own, and a transcript that is time-aligned to the media it transcribes is more useful again, providing the possibility for linking units of text (that is, utterances or words) directly to the position that they occur in the media. Current field methods include the use of tools like Elan[6] for creating such transcripts, but emerging methods for automated alignment of a transcript and media (for instance, WebMAUS[7]) promise to speed up this otherwise time-consuming process and can, as a first step, identify segments in the recording according to acoustic characteristics. Many legacy items in the collection have little metadata and no transcripts and would benefit from having a simple description of their content as a first step toward creating more detailed descriptions. In this way it may be possible automatically to identify different speakers, varying performance

[6] http://tla.mpi.nl/tools/tla-tools/elan/.

[7] http://phonetik.uni-muenchen.de/BASWebServices/.

types, and spoken tape identification at the beginning of the recording, all in order to improve the description of their contents.

Some collections, on the other hand, are heavily annotated and will allow re-use and reanalysis in future research projects, and can also be presented in online services representing languages of the world. There is a range of over 700 languages represented in the collection with a variety of styles, including songs, narratives, and elicitation. Given this rich source of material, there are great possibilities for re-use of the collections (subject, of course, to deposit conditions). It will be possible, for example, to establish crowdsourcing annotation of legacy material, either at the level of simply identifying parts of a recording or, where suitably skilled transcribers are available, to provide transcripts. We are also developing methods for delivery of the catalog and files via mobile devices.

Citing Primary Research Records

An example of the research use that a citable collection such as PARADISEC offers is the work done by Åshild Næss (2006) on the nature of the Reefs-Santa Cruz (RSC) (Solomon Islands) languages. Professor Stephen Wurm (mentioned earlier) had a considerable number of recordings from these languages in his house and office when he died. Næss was based in Norway and unable to get copies of the recordings, most of which were uncataloged and known to her only by oblique references in Wurm's work. As she notes (2006:159),

> Although Wurm published a number of papers on RSC, the actual data cited in these publications is limited to word lists and a few handfuls of frequently repeated example sentences. This makes it difficult to determine to what extent the structural claims, in particular, are actually supported by the data. Being able to evaluate and analyse Wurm's primary data will be of invaluable help in the effort to resolve the question of the origins of the Reefs-Santa Cruz languages.

Such recordings are invaluable to researchers, and we present them as playable objects in our collection for users to access. Furthermore, to make it easier to present interlinked text and media corpora, we have built an online system called EOPAS[8] that takes the media outputs of linguistic fieldwork together with texts[9] that are time-aligned to the source media and presents them online. EOPAS provides information about a text that satisfies several different needs at the same time. It gives the casual web user information about a text, showing grammatical and morphological complexity, but also allowing that complexity to be hidden via a toggle switch if desired. It allows a corpus of any number of texts in a language to be presented and searched, with a keyword-in-context view of any given word or morpheme—all resolving via a mouseclick to the context of the morpheme.

In my own research on the language of South Efate (Vanuatu) I have recently (Thieberger forthcoming) written on the relationship between two islands (Efate and Erromango) for which

[8] http://www.eopas.org.

[9] Actually *interlinear text*, that is, text with translations at the level of words or even smaller units.

there is linguistic evidence suggestive of contact. In the oral accounts that I had recorded on Efate I found a number of references to Erromango, so I was able to include both the text of the stories and a link to a playable version of them in the article. The archival form of this media is available for serious researchers, but a more casual observer can read the text and hear the media via the EOPAS version. In the story titled "Angels and Erromango"[10] a group of young Efate women used to fly to Erromango to wash in a particular river. A local man watched them there and hid the wings of one of the women, forcing her to stay and become his wife. She stayed and bore two children who then find her wings, and she is able to fly back to Efate. Another example is the story titled "Asaraf"[11] that is concerned with the theme of the closeness of the two islands before the giant Asaraf walked between them with the sea not reaching even to his knees, moving the islands apart and then making the sea rise. Ultimately, we hope to build access to the archival form of the media with an EOPAS-style front end. These stories were also published in a volume that can be downloaded from an open-access repository[12] or printed via Amazon's CreateSpace.[13]

Training

We are particularly interested in providing advice and training for researchers so that their records (be they recordings, photographs, transcripts, or more analytical work such as corpora, dictionaries, or grammars) will be archivable and reusable by others in the future, and we therefore emphasize the importance of linguistic data management (Thieberger and Berez 2012) and the principles established by Bird and Simons (2003) for the portability of research material. It is obvious from this training that the more a researcher knows about methods for creating good archival forms of their data and adopts those methods, the easier it is to accession that material into an archive. Another consequence is that their own research materials are also easier for them to access themselves over time.

PARADISEC has a blog (http://paradisec.org.au/blog) that often provides examples of new methods or summaries of projects using innovative approaches. We also helped to establish the Resource Network for Linguistic Diversity[14] that has a mailing list and FAQ page on relevant topics aimed at supporting many aspects of language documentation and language revitalization.

---

[10] http://www.eopas.org/transcripts/128

[11] http://www.eopas.org/transcripts/69

[12] http://repository.unimelb.edu.au/10187/9734

[13] The process is discussed in this blog item: http://www.paradisec.org.au/blog/2013/05/print-on-demand-again.

[14] http://rnld.org

Recognition

We have now created some nine terabytes of curated records that without our work would otherwise be only uncataloged analog material, and as a result we have been recognized in various ways. PARADISEC was cited as an exemplary system for audiovisual archiving using digital mass storage systems by the International Association of Sound and Audiovisual Archives[15] and was also included as an exemplary case study in the Australian Government's *Strategic Roadmap for Australian Research Infrastructure.*[16] In 2008 we won the Victorian eResearch Strategic Initiative (VeRSI) eResearch Prize (HASS category). In the words of the judges:[17]

> PARADISEC is an outstanding application of ICT tools in the humanities and social sciences domain that harnesses the work of scholars to store and preserve endangered language and music materials from the Asia-Pacific region and creates an online resource to make these available.

We are rated at five stars (the maximum rating) in the Open Language Archives Community[18] for the quality of our metadata. In 2012 our collection was awarded a European Data Seal of Approval,[19] and in 2013 PARADISEC's collection was inscribed in the UNESCO Australian Memory of the World programme.

Conclusion

Archiving of research outputs is central to language documentation and to the preservation of recorded oral tradition. Researchers have to ensure that speakers are able to locate records made with them or with their ancestors, and properly constructed repositories can provide that function. From a research perspective, the provision of properly curated scholarly material provides the basis for further research and for validation of the research that motivated the collection of the material in the first place. PARADISEC aims to be as responsive as possible (given our shoestring budget) to the individual needs of researchers, in particular those located in isolated and far-away communities who will be the main beneficiaries of the digitized set of material we have produced since we started work.

University of Melbourne

---

[15] Bradley 2004:51.

[16] http://www.nectar.org.au/sites/default/files/Strategic_Roadmap_Aug_2008.pdf

[17] This quotation is from our (unpublished) letter of award of the prize from the Victorian eResearch Strategic Initiative (http://www.versi.edu.au/).

[18] http://www.language-archives.org/metrics/paradisec.org.au

[19] https://assessment.datasealofapproval.org/assessment_75/seal/html/

References

Barwick and Thieberger 2006 — Linda Barwick and Nicholas Thieberger. "Cybraries in Paradise: New Technologies and Ethnographic Repositories." In *Libr@ries: Changing Information Space and Practice*. Ed. by Cushla Kapitzke and B. C. Bruce. Mahwah, NJ: Lawrence Erlbaum. pp. 133-49. Available at http://repository.unimelb.edu.au/10187/1672.

Bird and Simons 2003 — Steven Bird and Gary Simons. "Seven Dimensions of Portability for Language Documentation and Description." *Language*, 79:557-82. Available at http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf.

Bradley 2004 — Kevin Bradley, ed. *Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC04)*. Aarhus: International Association of Sound and Audiovisual Archives.

Lacey 1975 — Roderic Lacey. "Oral Traditions as History: An Exploration of Oral Sources among the Enga of the New Guinea Highlands." Unpub. Ph.D. diss.: University of Wisconsin-Madison.

Næss 2006 — Åshild Næss. "Past, Present and Future in Reefs-Santa Cruz Research." In *Sustainable Data from Digital Fieldwork: From Creation to Archive and Back*. Ed. by Linda Barwick and Nicholas Thieberger. Sydney: Sydney University Press. pp. 157-62. Available at http://hdl.handle.net/2123/1299.

Thieberger forthcoming — Nick Thieberger. "Walking to Erro: Stories of Travel, Origins, or Affection." In *The Languages of Vanuatu: Unity and Diversity*. Ed. by Alexandre François, Sebastien Lacrampe, Stefan Schnell, and Mike Franjieh. Studies in the Languages of Island Melanesia. Canberra: Asia-Pacific Linguistics.

Thieberger and Barwick 2012 — Nicholas Thieberger and Linda Barwick. "Keeping Records of Language Diversity in Melanesia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)." In *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century.* Ed. by Nicholas Evans and Marian Klamer. Honolulu: University of Hawai'i Press. pp. 239-53. Available at http://hdl.handle.net/10125/4567.

Thieberger and Berez 2012 — Nicholas Thieberger and Andrea Berez. "Linguistic Data Management." In *The Oxford Handbook of Linguistic Fieldwork.* Ed. by Nicholas Thieberger. Oxford: Oxford University Press. pp. 90-118.