PARADISEC:

Building methods for preserving ethnographic fieldwork recordings and providing long term access

Nick Thieberger, Linguistics, University of Melbourne

Nicholas Thieberger is a linguist who works with South Efate, a language from central Vanuatu. In 2003 he helped establish PARADISEC (paradisec.org.au), a digital archive (of which he is now Director). He is a co-director of the Resource Network for Linguistic Diversity (RNLD) and in 2008 he established a linguistic archive at the University of Hawai'i. He is interested in developments in e-humanities methods and their potential to improve research practice and he is now developing methods for creation of reusable data sets from fieldwork on previously unrecorded languages. He is the editor of the journal Language Documentation & Conservation. He taught in the Department of Linguistics at the University of Hawai'i at Mānoa and is now an Australian Research Council QEII Fellow at the University of Melbourne.

Among the many analogue recordings that we know are out there, those that were made in small communities and minority indigenous languages of Australia's region are of particular cultural heritage value. These could have been made by researchers like musicologists, anthropologists or linguists, or by patrol officers, missionaries, or travellers. Finding these tapes, digitising them and making them findable by others is what the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)[1] has been doing since 2003. In this paper I will describe PARADISEC and its processes, and give a summary of the present contents of the collection.

Introduction

PARADISEC began in 2003 by digitising collections of audio recordings made since the early 1960s by Australian National

1 http://paradisec.org.au

University (ANU) researchers. No Australian collecting agency was able to preserve these tapes, and none was digitising them. A group of linguists and musicologists set about learning what were the best ways in which to digitise the audio and what metadata should be used to describe it. We 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 catalogue's metadata) and for the more hands-on requirements of cleaning and repairing mouldy or damaged analogue tapes.

Having designed the system, we applied for Australian Research Council funding to buy a Quadriga workstation and associated equipment (playback machines, low temperature vacuum oven) and to fund staffing to begin digitising the several hundred tapes that were part of their first survey of such material. We built a catalogue, initially written in FileMaker Pro, then, after a couple of years, we moved to an online SQL/PHP system, and, with a further round of funding, we built our own online system (called 'Nabu') that manages the ingestion, description and curation of our collection.

Over the decade during which it has been running, PARADISEC has digitised several thousand hours of analogue recordings in three ingestion units based at each of the participating universities (University of Sydney, University of Melbourne and the Australian National University). As of March 2014 we hold 58,000 files, of which 11,000 are wav audio files.

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 languages represented. In 2011 we initiated an online survey[2] to locate

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

further endangered analogue collections and to work with their custodians to find funds to digitise and curate them before they are lost.

What is in the collection?

The contents of collections range from hundreds of recordings of a particular language made in the course of extensive fieldwork, through to short examples recorded opportunistically in a language. The records 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 catalogue. Occasionally there are handwritten transcripts of these recordings which we have included as scanned TIF or pdf files.

PARADISEC is making information available in an ethically appropriate way and we have established working relationships with agencies in our region like the Vanuatu Cultural Centre, Institute of Papua New Guinea Studies, University of French Polynesia, and the University of New Caledonia, among others. In 2014 we have applied for funds with the Solomon Islands Museum to digitise some hundreds of tapes they hold in Honiara. We have started a crowdfunding campaign to try to raise the funds necessary to do this work[3] and to locate more endangered collections of analogue recordings.

Since building the necessary online tools for entering cataloguing information for this kind of material we have had a number of born-digital collections deposited. It is particularly interesting for scholars to now be able 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

3 http://paradisec.org.au/sponsorship.htm

by an archive. Archiving before the analysis makes the research grounded and replicable and it turns on its head the more traditional approach of archiving primary recordings 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 PhD candidate in Theatre at Cornell University in the USA who was writing a dissertation on Oceanic theatre and drama. She needed access to a play that was listed in our catalogue 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[4]. Dr. Looser transcribed the tapes and prepared the only extant version of the script which she then redeposited in the collection. This re-use of research material in new ways can only be achieved if that material is stored in accessible locations with licences for use in place and with a catalogue that provides sufficient information to allow it to be located.

Technicalities

We began by installing a Quadriga analogue-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 catalogue. This catalogue presents a set of metadata elements to the user with dropdown menus to enforce standard forms, in particular for terms that are exposed to external harvesting tools to allow remote searching of the catalogue. These terms include country names (ISO 3166-1), language names (ISO-639-3), and datatypes, among other elements.

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

The online catalogue (Nabu[5]) 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 5,500 items (out of 9,800) 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 catalogue makes the depositor provide rather basic information that they may not previously have compiled, including, for each item, a title, date of creation, language spoken, and country recorded in. Further information can include: 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 which can be as rich as the depositor wants. All of this can be improved on by subsequent researchers who may use the collection in their own projects (as we saw above with the item from Tom Dutton's collection).

Research uses of the collection

An example of the research use that a citable collection like PARADISEC offers is the work done by Åshild Næss (2006) on the nature of the Reefs-Santa Cruz (RSC) (Solomon Islands) languages. The late Professor Stephen Wurm (of the ANU) had a considerable number of recordings from these languages in his house and office when he died. Næss was based in

5 http://catalog.paradisec.org.au and the open source of the catalog software is available at https://github.com/nabu-catalog/nabu

Norway and unable to get copies of the recordings, most of which were un-catalogued and known to her only by oblique references in Wurm's work. As she notes, "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[6] which takes the media outputs of linguistic fieldwork together with texts[7] 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 (parts smaller than a word), all resolving via a mouse-click to the context of the morpheme.

Community access to the collection

A key aspect of the creation of digital repositories like PARADISEC is that they can provide access to primary material to any authorised user. Authorisation for most material is simply obtained by supplying a valid email address. Delivery of media

6 EthnoER Online Presentation and Annotation System http://www.eopas.org
7 Actually interlinear text, that is, text with translations at the word or smaller level.

allows for web or mobile phone access, and, in cases where there is not yet easy internet or mobile coverage, we have also trialled simpler solutions, like making CDs or creating ITunes installations for school computer systems[8]. With properly constructed field recordings and time-aligned transcripts it is a relatively simple matter to locate the parts needed to create a set of stories told by elders which become tracks in ITunes. Users can then establish favourites and burn their selection to a CD for their use at home.

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 the text (utterances or words) directly to the position where they occur in the media. Current field methods include the use of tools like Elan[9] for creating such transcripts, but emerging methods for automated alignment of a transcript and media (e.g., WebMAUS[10]) 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 towards creating more detailed descriptions. In this way it may be possible to automatically identify different speakers, varying performance 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

8 PARADISEC, Endangered Languages and cultures, http://www.paradisec.org.au/blog/2010/03/how-can-we-get-the-material-we-have-used-in-our-research-back-to-the-people-we-recorded/

9 Max Planck Institute for Psycholinguistics, The Language Archive, http://tla.mpi.nl/tools/tla-tools/elan/

10 http://phonetik.uni-muenchen.de/BASWebServices/

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 catalogue and files via mobile devices.

Citing primary research records

We are particularly interested to provide advice and training for researchers so that their records (be they recordings, photographs, transcripts or more analytical work like corpora, dictionaries or grammars) will be archive-able and reusable by others in future, emphasising the importance of linguistic data management (Thieberger & Berez 2012), and based on the principles established by Bird and Simons (2003) for the portability of research material. It is obvious from this training that the more that 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[11] 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[12] which has a mailing list and FAQ page on relevant topics aimed at supporting many aspects of language documentation and language revitalisation.

11 http://paradisec.org.au/blog
12 Resource Network for Linguistic Diversity http://rnld.ord

Recognition

We have created some nine terabytes of curated records that, without our work, would otherwise be only un-catalogued analogue material, and, as a result PARADISEC was cited as an exemplary system for audiovisual archiving using digital mass storage systems by the International Association of Sound and Audiovisual Archives[13] and was also included as an exemplary case study in the Australian Government's *Strategic Roadmap for Australian Research Infrastructure*[14]. In 2008 we won the Victorian eResearch Strategic Initiative (VeRSI) eResearch Prize (HASS category). In the words of the judges: "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[15] for the quality of our metadata. In 2012 our collection was awarded a European Data Seal of Approval[16] and, in 2013, PARADISEC's collection was inscribed in the UNESCO Australian Memory of the World programme.

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

13 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 (IASA), p. 51.
14 2008 Strategic Roadmap for Australian Research Infrastructure, http://www.nectar.org.au/sites/default/files/Strategic_Roadmap_Aug_2008.pdf, p.42 (viewed 26/3/2014)
15 http://www.language-archives.org/metrics/paradisec.org.au
16 https://assessment.datasealofapproval.org/assessment_75/seal/html/

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 digitised set of material we have produced over the past decade.

***

References

Barwick, Linda and Nicholas Thieberger. 2006. "Cybraries in paradise: New technologies and ethnographic repositories". In Libr@ries: Changing information space and practice Ed. by Cushla Kapitzke & Bruce C. Bruce. Mahwah, NJ: Lawrence Erlbaum. pp.133-149. (http://repository.unimelb.edu.au/10187/1672)

Bird, Steven, and Gary Simons. 2003. "Seven dimensions of portability for language documentation and description". Language 79:557-582. (http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf)

Næss Aashild. 2006. '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.157-162. (http://hdl.handle.net/2123/1299)

Thieberger, Nick. (forthcoming) "Walking to Erro: Stories of travel, origins, or affection". In Alexandre François, Sebastien Lacrampe, Stefan Schnell, and Mike Franjieh. (eds), The Languages of Vanuatu: Unity and Diversity. Studies in the Languages of Island Melanesia. Canberra: Asia-Pacific Linguistics.

Thieberger, Nicholas and Andrea Berez. 2012. "Linguistic data management". In The Oxford Handbook of Linguistic Fieldwork Ed. by Nicholas Thieberger. Oxford: Oxford University Press. pp.90-118. (http://bit.ly/lingdatamanagement)

Thieberger, Nicholas and Linda Barwick. 2012. "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. [LD&C Special Publication No. 5].

Honolulu: University of Hawai'i Press. pp.239-253. (http://hdl.handle.net/10125/4567)