Barwick, L. (2003, 31 July). *Planning for PARADISEC.* Paper presented at the Oz-eculture
    Conference, Brisbane.

Note: this is a preprint version. The original is published online by the
Department of Communication Information Technology and the Arts at
http://www.cultureandrecreation.gov.au/conference3/barwick/barwick.pdf

Planning for PARADISEC:
The Pacific And Regional Archive for Digital Sources in Endangered Cultures

Presentation to the Ozeculture conference,
Brisbane Powerhouse, 31 July 2003
by Linda Barwick, University of Sydney

Introduction

PARADISEC is a collaborative digital research resource set up by the University of Sydney, the University of Melbourne and the Australian National University in 2003, with funding from the Australia Research Council's Linkage Infrastructure Equipment and Facilities scheme. I am a Senior Research Fellow in the Department of Music, University of Sydney, and Team Leader of the Sydney Unit of PARADISEC, and have been closely involved in planning and setting up the PARADISEC project.

Conceived and created in cyberspace, the project locates its digitisation equipment at the University of Sydney, its website at ANU, and metadata database at the University of Melbourne, with researcher contributions from all three Universities. Current planning issues concern provision of appropriate levels of digital rights management and access for the many stakeholder communities located throughout the Asia-Pacific region.

This presentation will outline the principles that have guided us in planning and implementation of PARADISEC.

1. Rationale for PARADISEC

PARADISEC's establishment has been driven by a consortium of Australian researchers, mainly linguists and musicologists, who have been aware for many years that their original field recordings were of great significance to communities and to the international research community, often constituting the only records of languages or musical genres that were on the verge of going out of use.

1.1 Endangered regional languages

Over 2000 of the world's 6000 languages are spoken in Australia's region (Oceania, E and SE Asia), but the majority of these are endangered as a result of pressure from English or other lingua franca. In a recent UNESCO report, it was estimated that the number of languages spoken in our region is likely to fall to a few hundred by 2100 (Bjeljac-Babic 2000). This situation is not merely of concern to the researchers and

communities involved. The loss of language leads to the loss of cultural knowledge (e.g. ecological knowledge) and expressions (e.g. songs), and hence to a loss of human diversity.

1.2 Endangered recordings
With the development of portable audio recording technology, first with reel-to-reel tapes in the 1950s, and then cassettes in the 1970s, audio recordings became an essential part of the field researcher's kit, and opened up new ways of transcribing and analysing audio data. A huge quantity of research audio recordings has been produced and a crisis now looms as to how to deal with it (Barwick 2003). Research recordings are endangered in a number of ways.

1.2.1 Format obsolescence
Researchers' field recordings are almost always unique recordings of unrepeatable events. The encroaching obsolescence of analogue sound recording formats is thus of even more concern than it is for commercially published recordings.

1.2.2 Physical deterioration
On top of this problem, research recordings made -- and (even worse) stored -- in tropical climates are now rapidly deteriorating due to mould and dust which rapidly advance the inevitable ageing of the media. We have found recordings made as recently as the 1980s are becoming unplayable due to deterioration of the tape medium.

1.2.3 Separation of recordings from metadata
Furthermore, as the field researchers of the 1950s and 60s have retired or passed away, their audio collections become orphaned, with no-one to look after them and to ensure that the metadata needed to make sense of the recordings stays with them. Many University language and music departments have filing cabinets or store-rooms full of mysterious recordings that, according to your point of view, may be jewels beyond price, or may be junk.

1.3 Australian regional research
Australian researchers already hold many unique field recordings of Pacific region cultures.

1.3.1 Cultural heritage value
These recordings and the language documentation based on them have high cultural heritage value for their home communities, and entail ethical responsibilities between the communities and the researchers

involved. In many cases communities have agreed to collaboration with University-based researchers because they are concerned to make sure that their languages, stories and songs are recorded and archived for future generations. Universities have a responsibility to maintain this data in accessible form.

1.3.2 Research value
Field recordings are the primary data for linguistic and musicological documentation and description, which in turn form the basis for theoretical development. These recordings may have the potential to prove or disprove important hypotheses for the research discipline in question.

1.3.3 Providing for future research and collaboration
There is ongoing Australian involvement in cultural research in the Pacific region, and our geographical proximity means that this involvement is likely to continue. We need to train Australian students and future researchers by providing ongoing access to the results of previous Australian research in the area. In the process of digitising and providing access to research recordings, we also open up new opportunities for engagement and collaboration with stakeholder communities, including repatriation of recordings to stakeholder communities and regional archives.

1.4 Lack of suitable current repository
No current central facility exists for deposit of Pacific region sound recordings or other digital data, and PARADISEC has been created to fill this significant gap. Unlike researchers whose primary interests are in Australian field research, researchers working in the Pacific region have had no suitable place of preservation of their primary research data.

1.4.1 The geographical area of interest falls outside the collection policy of Australian national institutions. The Australian Institute of Aboriginal and Torres Strait Islander Studies, the National Library of Australia, and Screensound Australia all focus their collections on Australian materials.

1.4.2 Cultural centres exist in some, but not all Pacific regional countries (e.g. the Vanuatu Kaljoral Senta, the Institute of Papua New Guinea Studies), but such centres are generally under-resourced (for a summary of computer infrastructure in the Pacific region, see Batiri Williams 2002). Australian researchers continue to deposit their field materials in local repositories where these are available, but older collections created before the establishment of such centres represent a source of highly-

valued intangible cultural heritage in need of a coordinated approach for digitisation and repatriation.

1.4.3 Some, but by no means all, Australian research institutions provide facilities and support for archiving of research materials. Developments in human ethics and intellectual property policies and regulations adopted by Australian Universities and required by funding bodies may entail institutions taking greater responsibility to look after these valuable research resources. Again, a coordinated approach is desirable in order to provide economies of scale and interoperability in creation and management of the resources (Bird and Simons 2002).

2. PARADISEC Project Goals

Digital archiving of endangered recorded field material from the region around Australia is our primary goal. At present we are focusing on audio material because there are established archival formats. Preservation of video presents a challenge that we have deferred until clear agreed archival formats emerge. Existing digital material, such as transcripts, theses, dictionaries and so on are also within the purview of the project.

Originally we planned simply to digitise field recordings as a service to researchers, with a central database to keep track of the materials, but exciting recent developments in Australian research infrastructure have encouraged us to extend our vision: to create an entirely electronic digital archive that could not only store and manage the digitised recordings, but also give networked electronic access to them. We thus have three additional short-term goals related to implementing PARADISEC:

- To exploit the potential of digital systems, in order to build a functional cross-institutional collaborative research resource;
- To develop and implement electronic management of our digital research archive; and
- To create linkages (electronic and otherwise) between Australian research institutions, national archival institutions, stakeholders in the Asia-Pacific region, and international bodies.

We see all three steps as necessary to ensure the future viability of the resource.

3. Standards and technicalities

3.1 Preservation

Because we are a new facility and have not inherited a large collection, we have been able to adopt current and proven digital technology. We aim to conform to international best practice by creating and structuring

the digital data using the best current tools, and by using internationally accepted archival data standards and formats to facilitate management & future migration.

Our audio standard is the Broadcast Wave Format (BWF), ingested and managed using the Quadriga audio system.
- BWF is an international audio standard developed by the European Broadcast Union, and now adopted by key Australian national institutions (Screensound, NLA, AIATSIS, the War Memorial, Australian Archives, amongst others). BWF files consist of uncompressed pulse code modulation (PCM) audio, based on the WAV format, with an additional encapsulated metadata chunk. BWF files have the .wav file extension and can be read by standard audio software.
- We have chosen to use the 24-bit, 96khz data rate for audio because of the unique heritage values of the recordings in our collection; and to facilitate any future noise reduction. We aim to produce a good quality flat transfer from analogue to digital, without any further digital signal processing at the time of ingestion.
- Encapsulated metadata includes unique permanent identifiers, coding history, and content descriptors harvested from our metadata catalogue. Both our metadata catalogue and the Quadriga system import and export data in XML format.
- Digital 'sealing' for data authentication is provided by the Quadriga system.

PARADISEC is a digital archive only. It provides temporary secure storage while analogue tapes are being digitised, but physical archiving of the originals remains the responsibility of the originating institution or depositor. When the originals have been digitised, they are returned to the depositor with advice, if requested, on suitable options for physical archiving. With the current adoption of digital technologies that record directly to hard disk, in future no original carrier may exist, only digital clones. PARADISEC plans to develop and provide advice as to proper handling of such recordings.

3.2 Description (metadata)
Metadata is handled via an online entry mechanism, and is intended for discovery, assessment, rights management and eventually as point of entry to the collection. PARADISEC metadata conforms to Open Languages Archives Community (OLAC) metadata, which itself conforms to Dublin Core. We have designed our database to be multifunctional, with automated export of information needed for

incorporation into the BWF files, and with the potential to map to the requirements of other search engines or gateways, such as the MusicAustralia gateway.

3.3 Rights

As a publicly funded collaborative cross-institutional research resource, we have a memorandum of understanding between the participant institutions to cover intellectual property issues for the collection as a whole. Depositor and user agreement forms cover specific access requirements for particular recordings, and this information is embedded into the processing system for eventual automated access or restriction of access. Password access is currently implemented on remote access to the shared database and Store files.

3.4 Access

Because PARADISEC is a digital archive, we are keen to exploit the potential of the system to provide remote access to audio (and later, other media). This is turning the traditional model of the archive on its head, to emphasise the distributive capacities of a well-structured digital archive. With specialised material such as ours it is essential to engage with the user communities who will most value the collection. At present users can:
• Download whole files from data store (e.g. for authorised community use)
• Listen to streaming MP3 (e.g. for browsing)
In 2004 we plan to add the following features to enhance the useability of the archive:
• Audition time-coded sections of an audio file
• Use transcripts, dictionaries, and images as point of entry to the audio collection.

3.5 Training & Resources

PARADISEC has already seen significant demand for practical workshops on digital field recording and archiving for researchers and communities. We are keen to encourage researchers to consider archiving as an integral part of their current field recording methodology, not just as something to think about at the end of one's career.

The preliminary PARADISEC website already functions as a gateway for related online resources, and as the project continues, the coordinating function of the website will ensure that additional resources are progressively created and updated. We hope that by publishing the results of our learning process in setting up the resource we may help others

planning similar projects and assist individual researchers with best-practice recommendations and the like.

4. Implementation of PARADISEC

Our current funding is from the Australian Research Council's Linkage Infrastructure Equipment and Facilities programme (1 year) and from participating institutions (University of Sydney, University of Melbourne, Australian National University). The project is governed by a Steering Committee made up of representatives of University of Sydney, ANU and University of Melbourne, and employs three staff: a project administration officer and audio preservation officer (based at the University of Sydney), and a project manager (based at the University of Melbourne).

4.1 Progress report July 2003
- Institutional memorandum of understanding signed.
- Sydney Unit operational from April 2003.
- Our preliminary website is online (http://www.paradisec.org.au).
- PARADISEC Metadata set, revision #3, has been published for comment, and is available via the website.
- Our metadata catalogue has been developed and is in use online for daily administration of the project from both Sydney and Melbourne.
- We have over 1200 assessed records in the database, covering recordings made in Australia, Burma, Fiji, Indonesia, Japan, Laos, Malaysia, Micronesia, New Zealand, Papua New Guinea, Singapore, Taiwan, Vanuatu and Vietnam, with over 150 regional languages represented.
- A trial data set of approximately 200 hours of audio has been ingested and is currently available online to researchers and administration via our APAC store account, with password access..
- PARADISEC is registered with the Open Language Archive Community, an international metadata gateway for linguistic material (http://www.language-archives.org). A selection of our database is now available to an international audience.

4.4 Plans for 2004
Assuming that our LIEF application for 2004 is successful, our most pressing task will be to secure ongoing programme funding for PARADISEC. In 2004 the original consortium will be joined by the University of New England. The high level of interest from researchers in many other research institutions across Australia has persuaded us that in the long run PARADISEC should be established as a national resource.

To cover present and future field research formats, PARADISEC needs to ingest current digital audio recording formats into the same access system as older analogue recordings, so we will field-test and develop archival workflow for ingestion of born-digital recording formats (DAT, CD, minidisc, Flash-RAM and so on). PARADISEC holdings will grow significantly faster because of the faster-than-real-time transfer capacity of many digital media.

We also plan:
- to investigate means for researchers and community users to submit their own information directly into the database via a web-based data entry mechanism for metadata;
- to develop a web-based audio delivery tool linking metadata and the audio material, including access to timecoded sections of audio, and a system for linking images of fieldnotes and relevant audio;
- to develop a policy and pilot for digitisation and management of video materials.

5.1 Linkages
PARADISEC has been able to progress with the aid of goodwill, support and advice from the participating Universities and also from national institutions, including:
- ANU Internet Futures, the Australian Partnership for Advanced Computing, and Grangenet;
- ScreenSound Australia;
- National Library of Australia;
- Australian Institute of Aboriginal and Torres Strait Islander Studies.

We are actively pursuing links with other national and international bodies with similar aims including:
- PAMBU, (Pacific Manuscripts Bureau, ANU);
- EMELD (Electronic Metastructures for Endangered Languages Data);
- DELAN (Digital Endangered Languages Archives Network);
- Regional cultural organisations and archives.

5.2 PARADISEC Steering Committee 2003
- Linda Barwick (Team Leader), Music, Sydney
- Jane Simpson, Linguistics, Sydney
- Allan Marett, Music, Sydney
- Nick Evans, Linguistics, Melbourne
- Steven Bird, Computer Science, Melbourne
- Stuart Hungerford, ANU Internet Futures

- John Bowden, RSPAS Linguistics, ANU
- Ewan Maidment, Pacific Manuscripts Bureau, ANU

5.3 PARADISEC Staff 2003
- Project Manager (Nick Thieberger, Melbourne)
- Audio Preservation (Frank Davey, Sydney)
- Project Administration (Amanda Harris, Sydney)

5.4 Contacts
Team leader (Sydney unit)
- lb@paradisec.org.au

Project manager (Melbourne)
- nickt@paradisec.org.au

CREDITS
This paper draws on research and data provided by Nick Thieberger and Amanda Harris.

REFERENCES AND FURTHER READING
Visit PARADISEC's website at www.paradisec.org.au for a selection of online resources relevant to endangered languages and digitisation.

Barwick, Linda 2003 'The Endangered Cultures Research Group's Digitisation Project: Using Digital Audio for Musicological Research,' in C. Cole and H. Craig (eds), *Computing Arts: Digital Resources for Research in the Humanities Proceedings*. Canberra: Australian Academy of the Humanities, in press.

Batiri Williams, Esther 2002 *Digital Community Services: Pacific Libraries and Archives*: *Future Prospects and responsibilities*. New Delhi, UNESCO, 2002.

Bird, Steven and Gary Simons, 2002 'Seven Dimensions of Portability for Language Documentation and Description.' *Proceedings of the Workshop on Portability in Human Language Technology*. Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, May 2002.

Bjeljac-Babic, Ranka 2000 '6,000 Languages: An Embattled Heritage,' *UNESCO International Courier*, April 2000. Online at http://www.unesco.org/courier/2000_04/uk/doss03.htm (accessed 23/1/2003).

Council for Library and Information Resources 2001 *Folk Heritage Collections in Crisis*. Washington: Council for Library and Information Resources. Online version at

www.clir.org/pubs/reports/pub96/rights.html (accessed 27/7/2002).

Grenoble, L. A. and L. J. Whaley 1998 *Endangered Languages: Language Loss and Community Response.* Cambridge UK, Cambridge University Press.

Krauss, Michael 1992 'The World's Languages in Crisis.' *Language* 68.1 (1992): 4-10.

Nathan, David 1997-2003 *Aboriginal Languages*. Website http://www.dnathan.com/VL/austLang.htm (accessed 1/1/2003).

Nelson, Melissa and Philip M. Klasky 2001 'Storyscape: The Power of Song in the Protection of Native Lands,' *Orion Afield* (Autumn 2001), published online at http://www.orionafield.org/pages/oa/01-4oa/01-4oa_Storyscape.html (accessed 17/1/2003).

Nettle, Daniel and Suzanne Romaine 2001 *Vanishing Voices. The Extinction of the World's Languages*. Oxford: Oxford University Press, 2000.

National Library of Australia 2000 *National Library of Australia Digitisation Policy 2000-2004*, published online at http://www.nla.gov.au/policy/digitisation.html (accessed 14/1/2003).

Quinn, E. M. 2001 'Endangered Languages, Endangered Lives.' *Cultural Survival Quarterly* 25(2).

Schmidt, Annette 1990 *The Loss of Australia's Aboriginal Language Heritage*. Canberra: Aboriginal Studies Press.

UNESCO 2002 'Linguistic Diversity: 3,000 Languages in Danger.' UNESCO press release 21 February 2002, online at http://www.unesco.org/education/imld_2002/press.shtml (accessed 17/1/2003).

Walsh, Michael 2001 'Why Language Revitalization Sometimes Works.' Paper delivered to the Endangered Languages Colloquium, Department of Linguistics, University of Arizona, 30 March, 2001, and Endangered Languages Colloquium, AIATSIS, Canberra 18 September, 2001.