The Pacific and Regional Archive for Digital Sources in Endangered Cultures
Background statement for the APAC Data Collections workshop, 18 October 2005

Linda Barwick, University of Sydney

PARADISEC (http://www.paradisec.org.au) is a cross-institutional research initiative established in 2003 by the Universities of Sydney, Melbourne and the Australian National University, joined in 2004 by the University of New England. Funded by the Australian Research Council's Linkage Infrastructure Equipment and Facilities Programme, participant institutions and Grangenet, PARADISEC offers a web-enabled facility for collaborative digitisation, management and access to Australian researchers' ethnographic recordings of endangered languages and musics of the Asia Pacific region. The collection is housed in APAC's store facility, where it is managed by Stuart Hungerford.

Since portable field recording equipment became readily available in the 1950s, many thousands of hours of ethnographic recordings have been made by Australian researchers. These unique and irreplaceable records are now in danger of being lost to future generations because of the impending obsolescence of analogue recording formats, deterioration of the original tapes, and orphaning of the collections as their creators retire or die. Before PARADISEC, there was no Australian repository available to salvage recordings made in the Asia-Pacific region. Indexing orphaned collections preparatory to digitisation was an important step in itself: our catalogue of 2400 records currently includes data on 390 languages from 50 countries in our region, previously inaccessible information that is now accessible worldwide via our web catalogue. As well as salvaging old recordings, we provide a facility for deposit and management of current research collections and advice on data creation and management for researchers planning future field trips. Our data collection hosts material recorded as long ago as the early 1950s and as recently as 2005.

1) The nature of the data collection

The vast majority of our current Store collection is made up of sound files, digitised for preservation purposes from analogue originals (cassette or reel-to-reel audio tapes) to international best practice standards for digital audio using the Quadriga audio archiving system. Our archival sound format is 24-bit 96khz Broadcast Wave Format (BWF), an extension of the uncompressed sound format .WAV with encapsulated metadata. For stereo recordings this yields about 2GB per hour. Our 2000 .WAV files archived in the APAC store facility represent 1000 hours of original recordings and occupy 1.61TB. For web access purposes we also archive an MP3 derivative of each WAV file (2001 files occupying 53.9GB).[1] We also host a variety of contextual data on these sound files, including texts (e.g. time-coded transcripts and translations of the recording contents) and images (e.g. relevant photographs, or page images of fieldnotes). We are currently in the process of adding to Store 10,000 page images of field notes related to the Capell and Wurm collections.

We are also beginning to grapple with the problem of video recordings. There is a large demand by researchers and communities for facilities to digitise, archive and annotate ethnographic video, but up to now this has been problematic because of the lack of international consensus on digital archival formats and standards as well as the high digital storage requirements (up to 20Mb per second for broadcast quality uncompressed video – 1.2GB per minute, or 720GB per hour). PARADISEC project manager Nick Thieberger leads the 2005-6 E-research initiative 'Sharing access and analytical tools for ethnographic digital media using high speed networks' (ARC SR0566965), in which humanities researchers from the Universities of Sydney, Melbourne, ANU, Macquarie and Newcastle are collaborating with computer scientists from CSIRO, University of Queensland and ANU to establish tools for online annotation and delivery of audiovisual research recordings.

[1] CD-audio quality derivatives are also produced in the process of digitization and returned to depositors on CD, but not archived.

2) The research group that provides and manages the collection

PARADISEC's chief investigators include linguists, musicologists and anthropologists from the Universities of Sydney, Melbourne, ANU and the University of New England.

PARADISEC is founded on a Memorandum of Understanding specifying that IP arrangements for items in the collection do not change by virtue of digitisation (in other words, the originating researchers, institutions and communities maintain their interests and rights in the content of the collection), and that new materials created in the course of the project belong to all participating institutions as tenants-in-common under NSW law. We are governed by a steering committee that includes representatives from each participating institution. Our Director, Linda Barwick, is based at the University of Sydney and our Project Manager, Nick Thieberger, at the University of Melbourne. We give priority to material from the participating institutions, but accept relevant material from researchers at other Australian and international institutions.

3) The users who access the data (and their access methods)

Users include researchers and postgraduate students (both Australian and international) and cultural organisations in the country of origin of the data.

Discovery of the data is made possible via our web-accessible catalogue.[2] Our metadata is also discoverable via various international web portals including the Open Language Archive Community, a subcommunity of the Open Archive Initiative, which harvests a periodic XML dump we provide of core metadata from our catalogue.[3]

The data can be accessed over the internet by depositors and those authorised by them via password authentication. The primary focus of the Store collection is currently archival. Most access is via CD-audio quality derivatives[4] of the archival files, burned to CD and provided to depositors as part of PARADISEC's ingestion workflow. CD-audio WAV is not only the most practical format for researchers to use for transcription and analysis, but is also the most widely used consumer audio format, and thus suitable for return to communities. Once we have sorted out current technical issues (see further below), we anticipate that a considerable proportion of future access to the collection will be via streamed mp3.

4) The tasks (problems, issues) with making the data accessible by the users.

a) Streamlining provision of access to non-networked users (e.g. small Pacific cultural organisations)
b) Providing a geographical search interface to the collection (in development)
c) Ensuring that we protect the intellectual property rights of the originating individuals, communities and institutions. We need secure reliable finegrained and auditable access control that integrates with existing regimes
d) Establishing and maintaining sustainable workflows for submission, authentication and access in a distributed system
e) Establishing appropriate secure cross-platform means for delivery of excerpts of streaming audio and video
f) Developing a suitable format for presentation of streamed media in the context of related information (transcripts, images, archival information)
g) Assuming d) and e), matching delivery formats and resolution to capability of user's network access and machine

[2] http://paradisec.org.au/catalog/
[3] http://www.language-archives.org/tools/search/.
[4] 16-bit 44.1khz stereo WAV files.