National Science Foundation Workshop for Documenting Endangered Languages

October 14 - 17, 2007
The New England Center,
University of New Hampshire,
Durham, NH

All materials contained herein are for viewing only. Reproduction or dissemination without permission is strictly prohibited. For permission or more information, write to DEL@NSF.GOV
PARADISEC, the Pacific And Regional Archive for Digital Sources in Endangered Cultures

Nick Thieberger
Linguistics Department
University of Melbourne

Documenting Endangered Languages Workshop, November 2007
What is PARADISEC?

Project aiming to preserve and make accessible researchers’ field recordings of cultural materials:

- fieldtapes
- notes,
- dictionaries,
- grammars,
- texts,
- etc.
What is PARADISEC?

Collaborative digital research resource set up by University of Sydney, University of Melbourne & Australian National University, 2003. (UNE joined 2004)

75% initial funding from Australian Research Council LIEF Scheme
PARADISEC aims

Recognition of the responsibility of researchers to preserve outputs of their research

Preservation: to adopt current optimal standards and formats to maximise sustainability and future usability of the collection
Endangered recordings

- Small and endangered languages recorded on analogue formats becoming obsolete
- Recordings physically deteriorating due to poor storage conditions (mould, dust etc)
• Examples:
• Stephen Wurm’s 1970s Solomon Islands tapes (~120 tapes and transcripts/fieldnotes)
• Arthur Capell’s 114 tapes, Pacific and PNG 1950s (and 30 archive boxes of fieldnotes)
• Bert Voorhoeve’s 180 tapes - West Papua
• Tom Dutton’s 295 PNG tapes
Endangered recordings

• Difficult to discover existence and thus plan to preserve such collections
• Virtually impossible for speakers to locate material in their languages
• Loss of research heritage and education sector investment in research
• No current repository to house this material
Regional links

Vanuatu Kaljoral Senta - provision of safe ‘blind’ backup of parts of their collection

University of New Caledonia
Digitisation of mouldy field recordings

Tjibaou Centre - New Caledonia - discussion of metadata and archiving methods

Institute of Papua New Guinea Studies - provision of CD copies of tapes, inclusion of funding for attendance our conferences
Online catalogue: paradisec.org.au/catalog
Online catalogue: paradisec.org.au/catalog
Online catalogue: paradisec.org.au/catalog
Data access (paradisec.org.au/repository)
Data access (paradisec.org.au/repository)

PARADISEC repository

collections » NT1 » 98004

Item NT1/98004

Here's a listing of the files of item NT1/98004 viewable by your username nxt563. Refer to the repository website glossary for an explanation of terms and abbreviations used in the repository listings.

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Duration</th>
<th>Class</th>
<th>Status</th>
<th>Metadata</th>
<th>Catalog Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT1-98004-98004A.mp3</td>
<td>42.28MB</td>
<td>00:46:10:91</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98004-98004A.wav</td>
<td>467.45MB</td>
<td>00:46:17:20</td>
<td>audio</td>
<td>offline</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98004-98004B.mp3</td>
<td>42.57MB</td>
<td>00:46:30:07</td>
<td>audio</td>
<td>online</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98004-98004B.wav</td>
<td>470.68MB</td>
<td>00:46:36:40</td>
<td>audio</td>
<td>offline</td>
<td>metadata</td>
<td>metadata</td>
</tr>
<tr>
<td>NT1-98004-98004a.xml</td>
<td>55.02KB</td>
<td>--</td>
<td>xml</td>
<td>offline</td>
<td>metadata</td>
<td>metadata</td>
</tr>
</tbody>
</table>

5 files 1023.04MB 03:05:34:58 -- -- -- --

Questions? Comments?

If you have questions or comments about this web page or the repository website the repository website FAQ will provide some answers and tell you the best way to get in touch.

nxt563 logged in

Site copyright © 2006 APAC National Facility.
Rights

- Depositor and user agreement forms online

- Rights information embedded in the processing system for eventual automated access or restriction of access

- Password access currently implemented on shared database and store files
Access

• Currently only depositor access
• Download whole files from data store (e.g. for authorised community use)
• CD audio/data copies provided to depositors and to relevant cultural centre if appropriate
Access

- Streaming media (browsing, using Annodex)
- Audition section of file (planned)
- Sample stories with time-aligned transcripts (EOPAS)
- Building on LACITO’s work

Nicholas Thieberger
PARADISEC


SAW2-009-excerpt.mp3/
Access

- Images of fieldnotes
  - Wurm notes (initially 120 items)
  - Capell notes (30 boxes, 14,000 images)
  - Roesler notes (600 images)
Training

We have run training sessions in the use of linguistic software (in particular Shoebox, Toolbox, Transcriber, Elan and regular expressions) at the following locations during 2004-2007:

- Melbourne University (4 x)
- Sydney University (3 x)
- University of Queensland
- Kalgoorlie Language Centre
- Muurrbay Many Rivers Language Centre (Nambucca Heads)
- New South Wales Aboriginal Languages Research and Resource Centre (Sydney)
- Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS)
- Victorian Aboriginal Corporation for Languages (Melbourne)
- Australian Linguistic Society conferences
- University of Hawai'i at Manoa (3 x)
- LSA Summer Institute, July 2007
PARADISEC Progress report

As at September 17th 2007 - 4,219 items in the catalog; 26,543 files totaling 3.34 TB, with 1854 hours of audio

Data from 599 languages from 55 countries

PARADISEC one of 36 participating OLAC archives - OLAC is a sub-community of the Open Archives Initiative
Linkages

Importance of relationships with regional cultural organisations, including repatriation of copies of tapes

- Vanuatu Kaljoral Senta - provision of safe ‘blind’ backup of their digitised sound collection
- University of New Caledonia - Digitisation of mouldy field recordings
- Institute of PNG studies
- Need more such links
Working well?

- Relationships with regional agencies
- Workflow for digitisation, metadata entry etc
- Training of new researchers
- Developing trust of depositors
- Extent of data converted from analog
Critical issues not covered?

• Outreach to our region
  • Location of endangered collections in the region
  • Preservation of these collections

• Funding!
  • Loss of expertise during funding hiatus

• Real need to establish methods for data curation, metadata etc that are easy to use
Cooperation between similar programs?

- OLAC
- DELAMAN

More efficient use of existing resources:

- Provision of templates and cataloging software
Contacts

http://paradisec.org.au

Director (Sydney)
linda.barwick@paradisec.org.au

Project manager (Melbourne)
nicholas.thieberger@paradisec.org.au
Preservation - principles

- Conform to international standards
- Use standard digital archival formats
- Open source software (reusability of components) where possible
- Plan for user communities (speakers and their descendants)
Workflow

• To build good data while doing normal work:
  • Fieldwork
  • Transcription
  • Interlinearisation
  • Lexicography
  • Grammatical analysis
Typical workflow resulting in well-formed data

Recording - named → analogue digitised/digital captured → archival digital file → archived with PARADISEC

- descriptive metadata added
- concordance of texts, navigation tool
- Media corpus instantiates links to media (e.g. Audiamus)
- output to e.g. Shoebox for interlinearising
- Texts, dictionary etc

transcribed and linked (using e.g. Transcriber or Elan)

archived with PARADISEC
Linkages

Testbed for the Australian Partnership for Sustainable Repositories project
Support from the Australian Partnership for Advanced Computing (APAC)
Participant in the Australian GrangeNet highspeed network
ANU Internet Futures Project (programming for web interface to the APAC account)
Australian Academy of the Social Sciences (French cooperation)
Sydney Uni International Development fund (U Texas visit)
EMELD, (airfares, accommodation and registration at the EMELD conference in Michigan, USA).
School of Society Culture and Performance, University of Sydney (RIBG funding support)
Faculty of Arts, University of Sydney (refurbishment of rooms and infrastructural and training support)
Test project for EthnoER media annotation grant
More ...