

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

PARADISEEC, 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



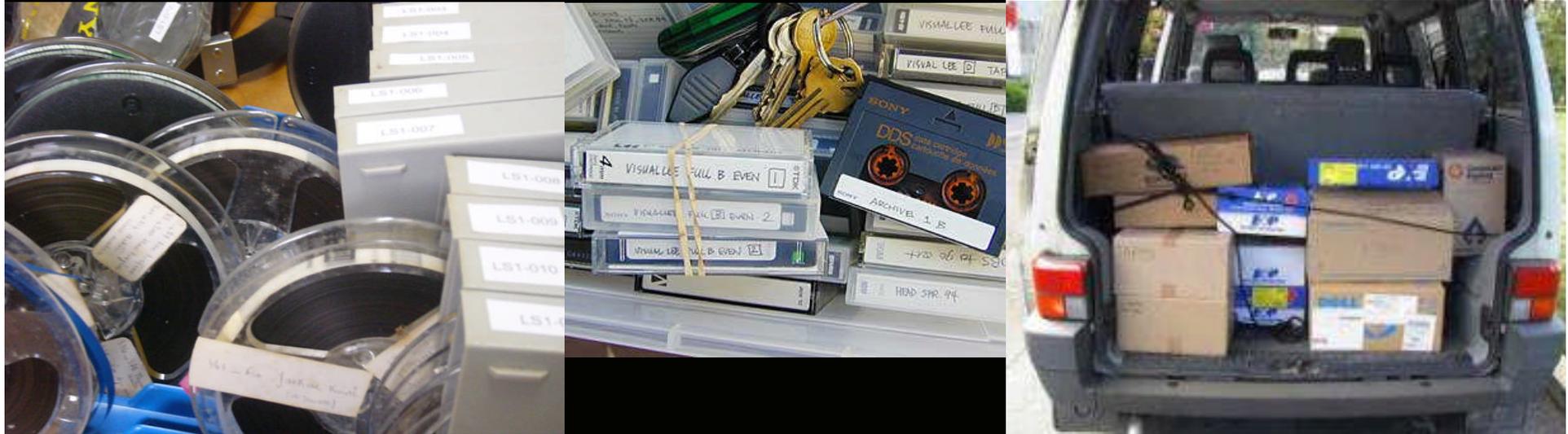
PARADISEEC aims

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

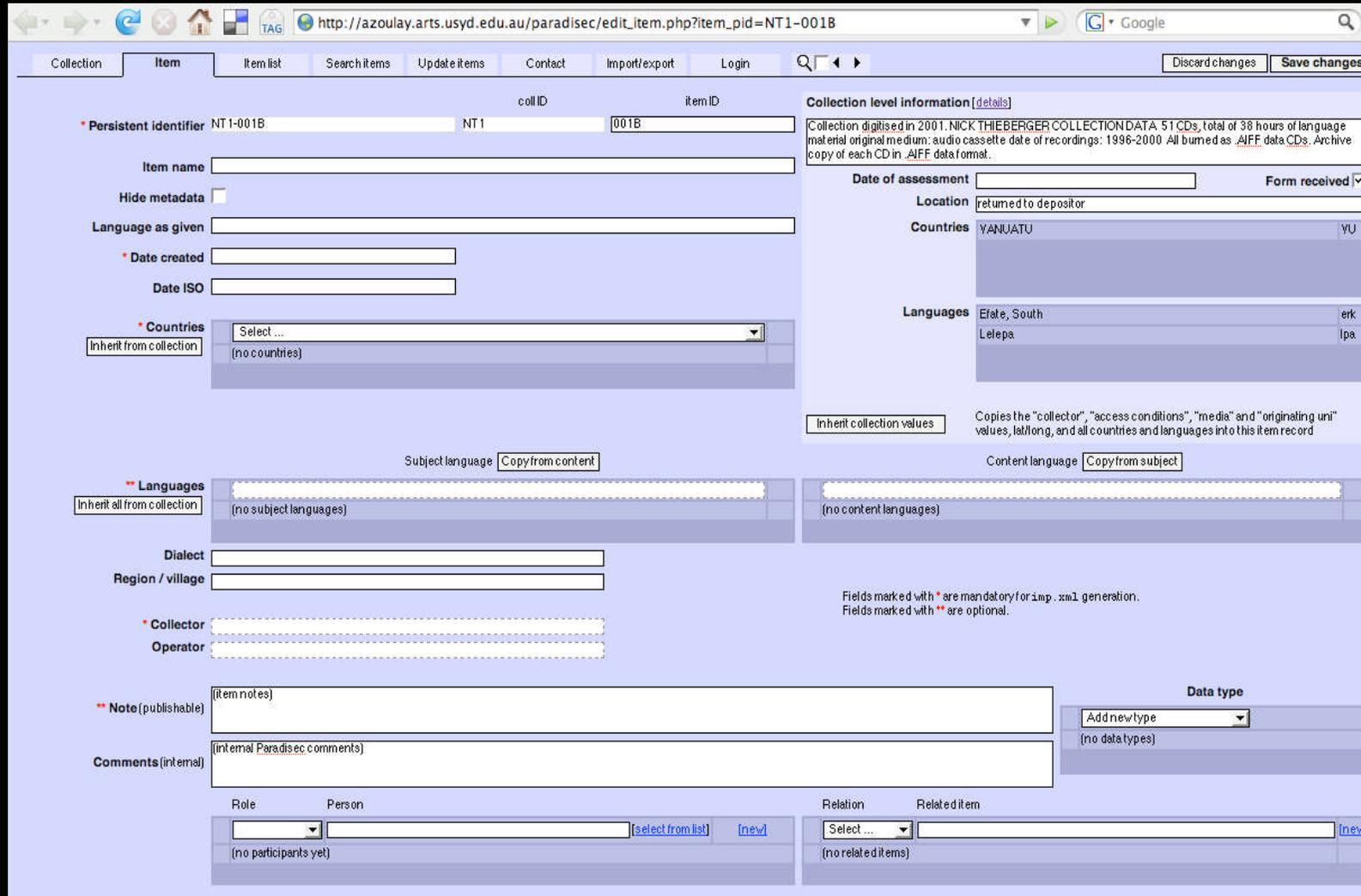
Institute of Papua New Guinea Studies - provision of CD copies of tapes, inclusion of funding for attendance our conferences



Nicholas Thieberger
PARADISEC

NSF Documenting Endangered Languages Workshop,
Durham, New Hampshire, October 2007

Online catalogue: paradisec.org.au/catalog



The screenshot shows the 'edit_item.php' page for item NT1-001B. The browser address bar shows the URL: http://azoulay.arts.usyd.edu.au/paradisec/edit_item.php?item_pid=NT1-001B. The page has a navigation menu with options: Collection, Item, Item list, Search items, Update items, Contact, Import/export, Login. There are buttons for 'Discard changes' and 'Save changes'.

Item Identification:

- Persistent identifier:** NT1-001B (collID: NT1, itemID: 001B)
- Item name:** [Text input field]
- Hide metadata:** [checkbox]
- Language as given:** [Text input field]
- Date created:** [Text input field]
- Date ISO:** [Text input field]
- Countries:** [Dropdown menu: Select ... (no countries)]

Collection level information [details]:

Collection digitised in 2001. NICK THIEBERGER COLLECTION DATA 51 CDs, total of 38 hours of language material original medium: audio cassette date of recordings: 1996-2000 All burned as .AIFF data.CDs. Archive copy of each CD in .AIFF data format.

Assessment and Location:

- Date of assessment:** [Text input field]
- Form received:** [checkbox checked]
- Location:** returned to depositor
- Countries:** VANUATU (VU)
- Languages:** Efate, South (erk), Lelepa (lpa)

Inheritance and Copying Options:

- Inherit collection values:** [checkbox]
- Subject language:** [Copy from content]
- Content language:** [Copy from subject]
- Languages:** [Text input field: (no subject languages)]
- Languages:** [Text input field: (no content languages)]

Other Fields:

- Dialect:** [Text input field]
- Region / village:** [Text input field]
- Collector:** [Text input field]
- Operator:** [Text input field]
- Note (publishable):** [Text input field: (item notes)]
- Comments (internal):** [Text input field: (internal Paradisec comments)]
- Data type:** [Dropdown menu: Add new type (no data types)]

Participants and Relations:

- Role:** [Dropdown menu]
- Person:** [Text input field: (no participants yet)]
- Relation:** [Dropdown menu]
- Related item:** [Text input field: (no related items)]

Fields marked with * are mandatory for imp.xml generation. Fields marked with ** are optional.

Online catalogue: paradisec.org.au/catalog

**** Languages**

Inherit all from collection Erate, South erk Bislama Erate, South

Dialect

Region / village

*** Collector** Thieberger, N

Operator Bone, Contr

**** Note (publishable)**

Comments (internal)

Role

recorder

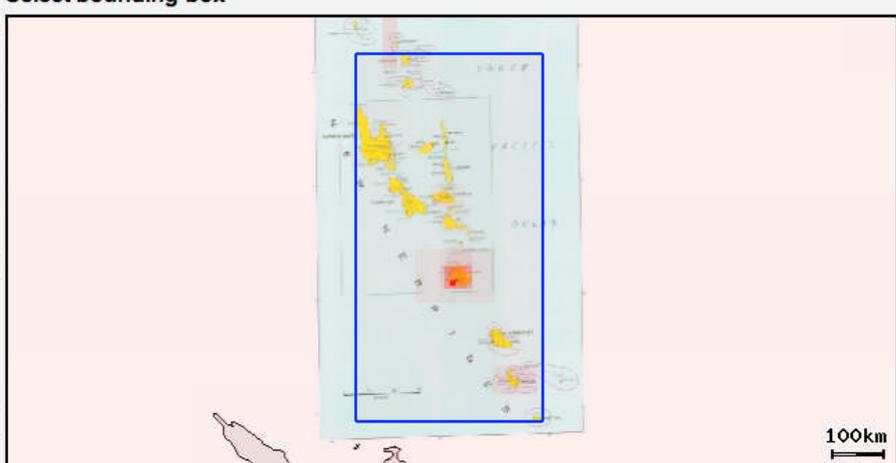
speaker

Filename

Files

- NT1-001-4
- NT1-001-4
- NT1-001-4

Select bounding-box X



100km

Layers

- Honshu
- Java & Bali
- Kai Islands
- Language Groups of The Pacific *
- Micronesia
- New Britain
- New Caledonia *
- New Hebrides ***
- Northeastern Irian Jaya
- Northern Celebes
- Northern Australia
- Northern Borneo
- Northern Mainland Southeast Asia

Show item density

Longitude -

Latitude -

Media

Tracking

Access conditions

Originating uni

	cassettes	r-to-r tapes (metr
number	<input type="text" value="0"/>	<input type="text" value="0"/>
length	<input type="text" value="0"/>	<input type="text" value="0"/>
total	<input type="text" value="0"/>	<input type="text" value="0"/>
search set total	<input type="text" value="0"/>	<input type="text" value="0"/>
	reel-to-reel speed	
	reel-to-reel time	<input type="text" value="0"/>

tape received data received

tape labelled

metadata entered

digitised date digitised

ready for imp.xml imp.xml generated

CD burnt CD id

Online catalogue: paradisec.org.au/catalog

Collection **Item** Item list Search items Update items Contact Import/export Login

coll ID DL1 item ID 018

*** Persistent identifier** DL1-018 DL1 018 [Collection level information \[details\]](#)

Catalogue link <https://store.apac.edu.au/paradisec/repository/DL1/018>

Item name Music of New Guinea. Recorded

Hide metadata

Language as given Buin & others

*** Date created** 1967-03-01

Date ISO

*** Countries** Inherit from collection

FRENCH POLYNESIA

PAPUA NEW GUINEA

**** Languages** Inherit all from collection

Terei

Dialect

Region / village Bangwis village, Waskuk village.

*** Collector** Laycock, Don

Operator Davey, Frank [\[details\]](#)

Select subject language [\[cancel\]](#)

Search languages (showing only languages from countries in Music of New Guinea. Recorded by D.C.Laycock, ANU)

Search Language name contains Ethnologue code is

Clear search

Showing languages 1 to 10 of 841 found [\[next matches\]](#)

Language name	Language code	
[select] Ari	aac	PG
[select] Amal	aad	PG
[select] Arifama-Miniafia	aai	PG
[select] Ankave	aak	PG
[select] Abau	aau	PG
[select] Solong	aaw	PG
[select] Abaga	abg	PG
[select] Ambulas	abt	PG
[select] Pal	abw	PG
[select] Aneme Wake	aby	PG

**** Note (publishable)** NB List is identical to DL1-017 (with extra info for tracks 11 and 12- copy? Compilation of songs recorded from 1959 -1967. list on inner cover. Sepik, Mt Hagen Show, Buin<p>Side 1<p>1. 1959 Flutes from Bangwis village, near Ambunti<p>2. Flutes from Waskuk village, Upper Sepik river<p>3. Crocodile trumpets from Waskuk<p>4. Man-killing singing from Waskuk<p>Side 2<p>1. 1959 Repeat of Waskuk singing<p>2. Spirit-repelling singing from Swagup<p>3. Drums from southern Maprik area<p>4. ...

Comments (internal) (internal Paradisec comments)

Role **Person** **Relation** **Related item**

performer: Malemole [\[select from list\]](#) [\[new\]](#)

performer: Tuura

(no related items)

Data access (paradisec.org.au/repository)



Welcome to the PARADISEC repository
Please log in to identify yourself

Username:

Password:



PARADISEC repository

[logout](#) | [collections](#) | [news](#) | [FAQ](#) | [glossary](#) | [admin](#)

Repository collections

Here's a listing of the collections viewable by your username *rxz563*. Refer to the [repository website glossary](#) for an explanation of terms and abbreviations used in the repository listings.

Collection	Items	Files	Size	Duration	Catalog Metadata
AB1	11 items	44 files	29.38GB	14:47:59.60	metadata
AC1	58 items	307 files	64.81GB	33:15:06.00	metadata
AC2	1,025 items	12,567 files	18.06GB	00:00:00.00	metadata
AM2	14 items	56 files	36.65GB	19:08:27.40	metadata
AM3	13 items	46 files	32.41GB	16:19:28.90	metadata
AM4	1 items	2 files	1.50GB	00:45:17.99	metadata
AP2	4 items	14 files	6.00GB	03:01:18.90	metadata
AR1	13 items	42 files	23.00GB	14:04:24.40	metadata
AR2	3 items	8 files	6.09GB	05:58:15.00	metadata
AS1	6 items	20 files	10.81GB	05:26:47.90	metadata
BE1	1 items	14 files	2.15MB	00:00:00.00	metadata
BH1	21 items	42 files	3.75GB	03:40:37.50	metadata
BM1	2 items	28 files	3.02GB	04:40:35.40	metadata
BP1	2 items	12 files	6.01GB	03:01:29.80	metadata
BP2	9 items	32 files	15.73GB	07:55:18.80	metadata
BP3	3 items	6 files	1.65GB	00:49:47.11	metadata
CB3	5 items	16 files	5.68GB	02:51:47.30	metadata
CH1	42 items	202 files	23.28GB	14:54:39.10	metadata
CH2	18 items	370 files	175.16GB	00:00:00.00	metadata

Data access (paradisec.org.au/repository)



PARADISEC

PARADISEC repository
[logout](#) | [collections](#) | [news](#) | [FAQ](#) | [glossary](#) | [admin](#)

[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.

File	Size	Duration	Class	Status	Metadata	Catalog Metadata
NT1-98004-98004A.mp3	42.28MB	00:46:10.91	audio	online	metadata	metadata
NT1-98004-98004A.wav	467.45MB	00:46:17.20	audio	offline	metadata	metadata
NT1-98004-98004B.mp3	42.57MB	00:46:30.07	audio	online	metadata	metadata
NT1-98004-98004B.wav	470.68MB	00:46:36.40	audio	offline	metadata	metadata
NT1-98004-98004a.xml	55.02KB	--	xml	offline	metadata	metadata
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

<http://maenad.itee.uq.edu.au/exist/exist/eopas3/transcript/13009745>

Catalog metadata: SAW2/009/SAW2-009-A.mp3

PARADISEC project repository

Located at the APAC National Facility, Canberra

Catalog metadata: SAW2/009/SAW2-009-A.mp3
[SAW2-009-excerpt.mp3/](#)

Name	Value
Identifier	SAW2-009
Title	Näkenaa ä Nuū Nubulaa
Date	1979-07-19
Media	NSW Gov C-60 cassette
Country	SOLOMON ISLANDS
Type	lexicon
Description	Words and Explanations. Image files of accompanying documentation.
Contributor	Wurm, S.A. (researcher)
Contributor	Moia, Martin (recorder)

Metadata notes
 The "Name" column in the table above shows the name of the metadata element stored in the catalog and the "Value" column the value of that element. Note that not all files have values for all catalog metadata elements.

[\[logout\]](#) [\[catalog\]](#) [\[help\]](#) [\[FAQ\]](#)

Wurm, Stephen Adolphe (1922 - 2001) Guide to Records Image Viewer

Wurm, Stephen Adolphe (1922 - 2001) Guide to Records

Read Enlarge Image 1 of 17 Back to Guide

you may need to select landscape paper orientation for better results.

NÄKENAA Ä NUU NUBULAA SAW2-009

ANOTHER CUSTOM STORY DONE BY FR. MARTIN FOR PRO. STEPHEN A. WURM.

Ibe nyiji Nubulaa la imo mo säpelivanona säpelivano
 old man one of village lived with wife wife

la imebetao lamole lamole ~~la~~ nwoa' ıla betoa itukā
 got pregnant lived on and one belly was ready gave birth to

dyca figilai nāyāna Gipoulo. La kumwaleitoa imwaleito
 baby boy named Gipoulo. They nursed him until

elo iwāmoto kwopolautā. Dānyidabudā la kwopolautōa
 a group up he started to sail by canoe One day he was sailing

<http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm>

Access

- Images of fieldnotes
 - Wurm notes (initially 120 items)
<http://paradisec.org.au/fieldnotes/SAW2.htm>
 - Capell notes (30 boxes, 14,000 images)
<http://paradisec.org.au/fieldnotes/AC2.htm>
 - Roesler notes (600 images)
<http://paradisec.org.au/fieldnotes/ROES/web/roes.htm>

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
- Murrumbidgee 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

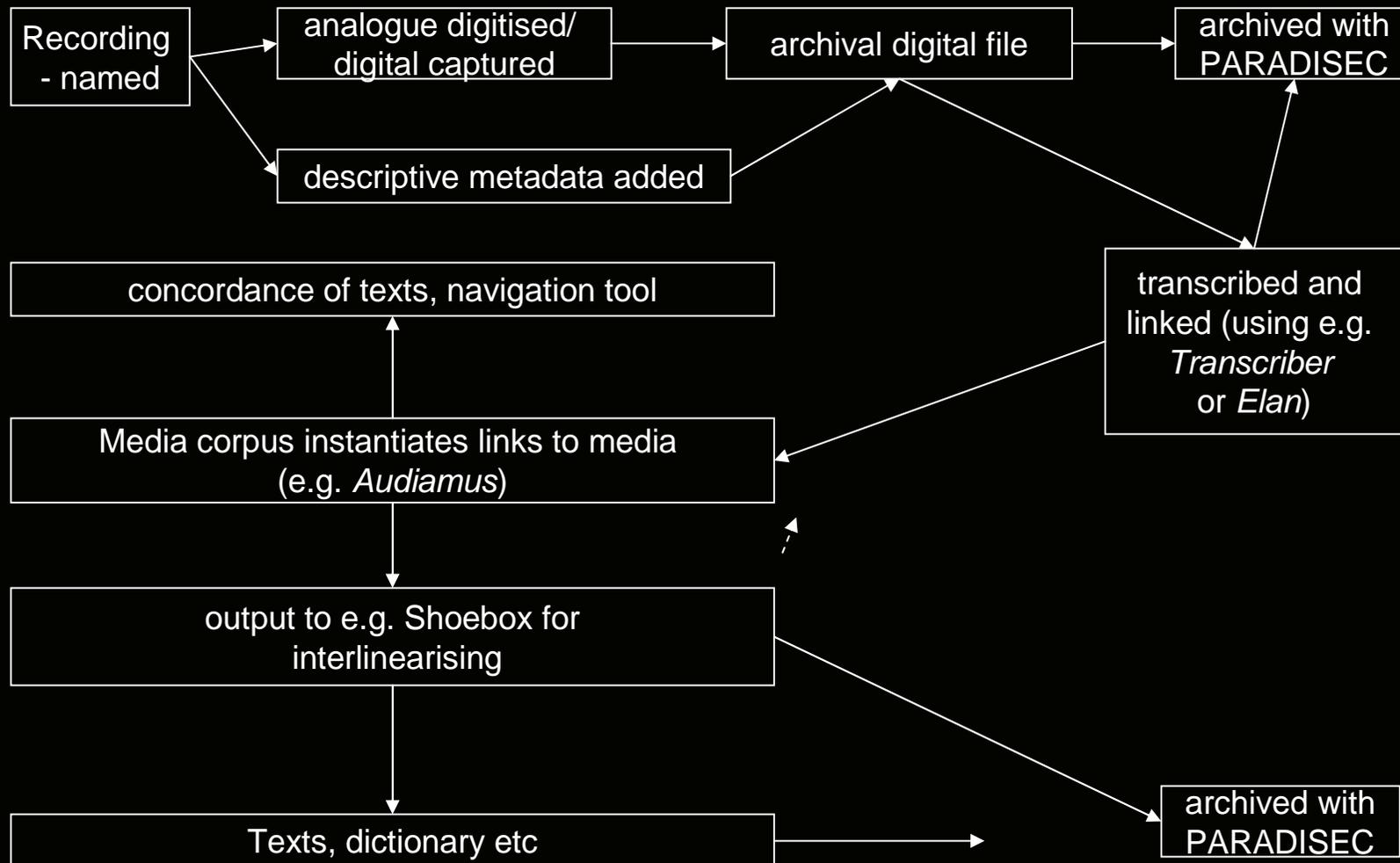
Preservation - principles

- **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



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 ...