



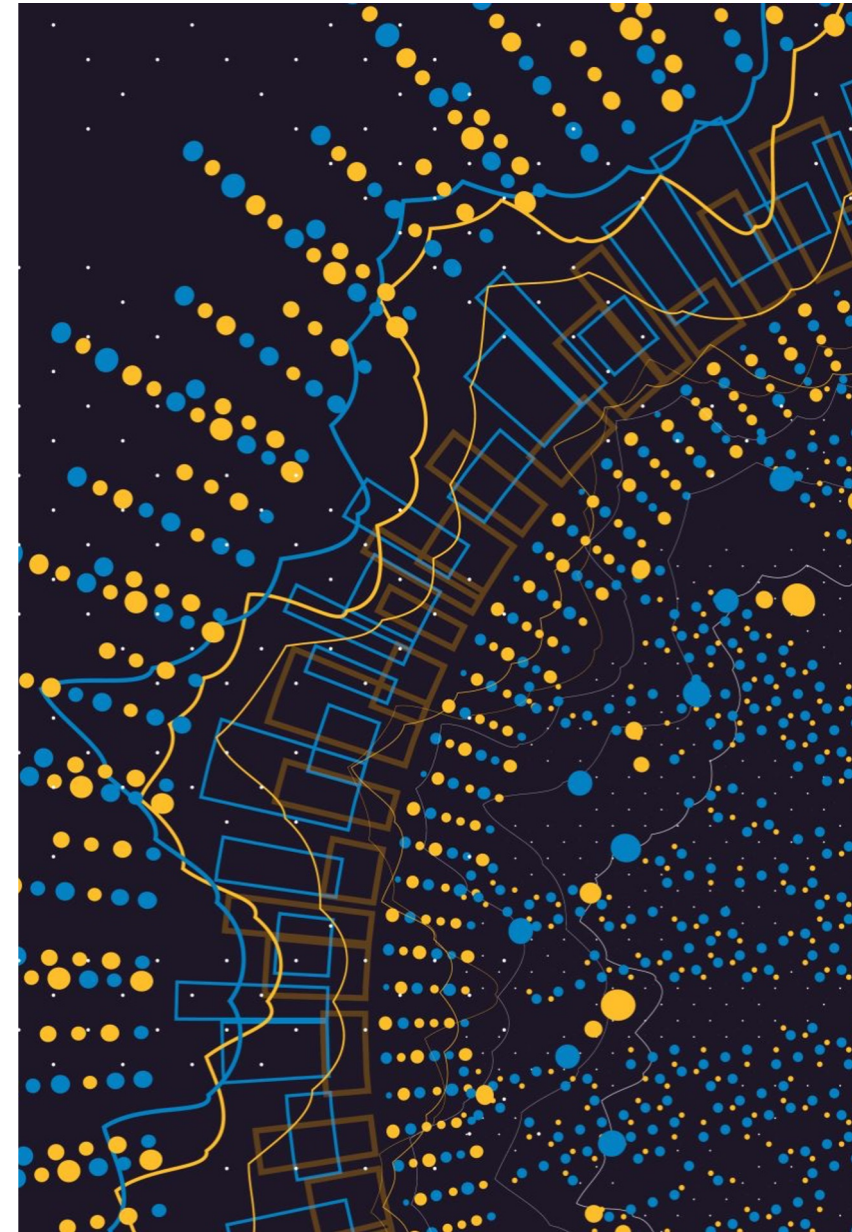
PARADISEC

THE PACIFIC AND REGIONAL ARCHIVE FOR
DIGITAL SOURCES IN ENDANGERED CULTURES
(PARADISEC)

Nick Thieberger,
School of Languages and Linguistics,
University of Melbourne, Australia

<https://paradisec.org.au>

I acknowledge and pay respect to the original owners of the land I
work on, the Woiwurrung people of the Kulin Nation



What we need

A nationally funded and guaranteed data service that will curate research data with appropriate licences so that it is available in future

In the meantime, we have to shepherd our data through various platforms, ensuring it is able to migrate from one to the next

Improving access to and sustainability of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

Access

- Online catalogue
- API feed to a number of services
- Clear licence conditions
- Ability to deliver arbitrary subsets of items with tailored catalogs

Sustainability

- Bitstream preservation
- Multiple backups
- Research-Object Crate (metadata stored with files)

Orphaned tapes:
senescence of the analog
evanescence of the digital



**THERE IS NOW
CONSENSUS AMONG
AUDIOVISUAL ARCHIVES
INTERNATIONALLY
THAT WE WILL NOT
BE ABLE TO SUPPORT
LARGE-SCALE
DIGITISATION OF
MAGNETIC MEDIA
IN THE VERY NEAR
FUTURE. TAPE THAT IS
NOT DIGITISED BY 2025
WILL IN MOST CASES
BE LOST FOREVER.**



No tape left behind, the deadline of 2025: Collections at risk

National and Film and Sound
Archive (Australia) DEADLINE 2025

Urgent need to digitise, and then
to provide a repository for these
files



BAKING TAPES AT
LOW TEMPERATURE
TO EXTRACT ALL
MOISTURE

<https://www.paradisec.org.au/blog/2007/05/baking-tapes-or-analogue-audio-restoration/>



CASSETTE LUBRICATION MACHINE

BUILT BY SAM KING(2025)

<https://www.paradisec.org.au/blog/2025/01/the-tape-restorator/>



PACIFIC AND REGIONAL ARCHIVE FOR DIGITAL SOURCES IN ENDANGERED CULTURES (PARADISEC)

- Established 2003
- Researchers concerned to digitise, preserve, and make accessible recordings in the many languages of the region around Australia
- Initially designed to digitise analog research recordings, now also a major community-facing resource
- No other agency taking responsibility for these recordings so they were at risk of loss



PACIFIC AND REGIONAL ARCHIVE FOR DIGITAL SOURCES IN ENDANGERED CULTURES (PARADISEEC)

- No core funding, relies on grants (Australian Research Council Future Fellowship, ARC Centre of Excellence, ARDC Language Data Commons of Australia funding, ARC Linkage Infrastructure and Equipment and Facilities)
- Catalog exposes the existence of these recordings
- Currently represent
 - **1,380** languages, many of which have no other presence on the web
 - **260** terabytes, with over **17,900** hours of audio recordings, plus **3,500** hrs video, text (incl 14,000 Elan transcript files), and images
- Moved to S3 storage in 2023, managed by the University of Sydney

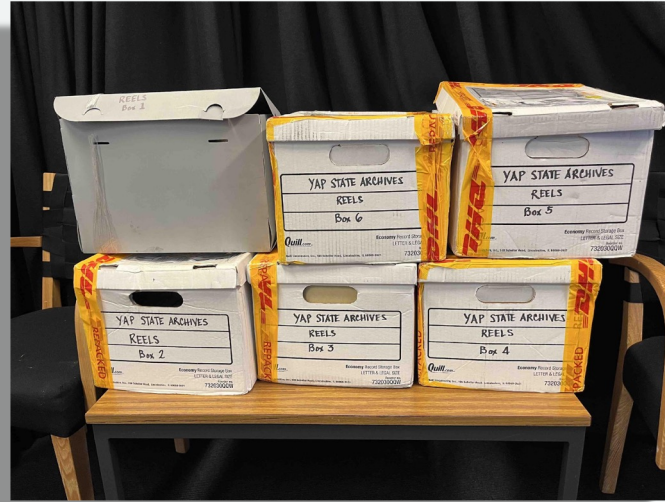
Countries represented in PARADISEC

Afghanistan (2), American Samoa (5), [Australia \(9814\)](#), Austria (2), Azerbaijan (3), Bangladesh (32), Bhutan (66), Bolivia (7), Botswana (2), Brunei (65), Bulgaria (3), Cambodia (4), Cameroon (15), Chad (1), Chile (25), [China \(1178\)](#), Cook Islands (201), Czech Republic (2), Denmark (2), East Timor (118), Egypt (1), Estonia (2), Ethiopia (143), Fiji (762), Finland (2), France (19), French Polynesia (104), Georgia (3), Germany (4), Ghana (34), Greece (3), Greenland (1), Guam (2), Hong Kong (2), Hungary (3), Iceland (3), India (741), [Indonesia \(4028\)](#), Iran (1), Italy (380), Japan (190), Kiribati (30), Korea, North (2), Korea, South (3), Laos (346), Latvia (4), Lebanon (2), Libya (1), Lithuania (2), Macedonia (2), Madagascar (18), Malaysia (307), Maldives (5), Marshall Islands (9), Mauritius (35), Mexico (6), Micronesia (158), Monaco (14), Myanmar (3427), Nauru (8), Nepal (577), Netherlands (2), New Caledonia (116), New Zealand (100), Nigeria (132), Niue (3), Northern Mariana Islands (31), Norway (3), Pakistan (27), Palau (5), Palestinian West Bank and Gaza (1), [Papua New Guinea \(9952\)](#), Philippines (342), Poland (2), Portugal (1), Romania (2), Russian Federation (167), Réunion (1), Samoa (166), Singapore (33), Slovakia (2), [Solomon Islands \(2105\)](#), Somalia (1), South Africa (8), Spain (2), Sri Lanka (19), Sudan (204), Sweden (2), Switzerland (3), Taiwan (36), Thailand (113), Tonga (100), Tuvalu (1), Uganda (2), United Kingdom (7), United States (205), [Vanuatu \(4477\)](#), Viet Nam (7), Wallis and Futuna (32)

Working with cultural agencies in the region

- to return materials recorded there in the past
 - Typically by Australian researchers (linguists, musicologists, anthropologists)
 - to digitise their analog tapes
 - No local playback machines, little capacity to do this work in situ
- Vanuatu Kaljoral Senta – Digitising 320 tapes
 - University of New Caledonia – Digitisation of mouldy field recordings in Drehu
 - 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
 - Divine Word University – Madang, PNG – digitising open reels
 - Solomon Islands National Museum – digitising 400 tapes
 - University of French Polynesia – training courses, digitising tapes, helping set up a new archive there

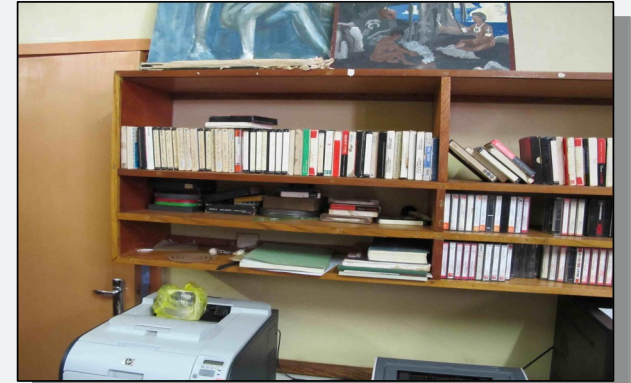
120 YAP STATE ARCHIVES TAPES DIGITISED IN 2023



UNESCO

Memory of the World Committee for the Asia-Pacific, Asia Culture Center
Grant of \$USD 5,000

Solomon Islands National Museum



> 600 more tapes need digitising, most include oral tradition in local languages
2014 – Endangered Archives Programme funding to digitise 200 tapes

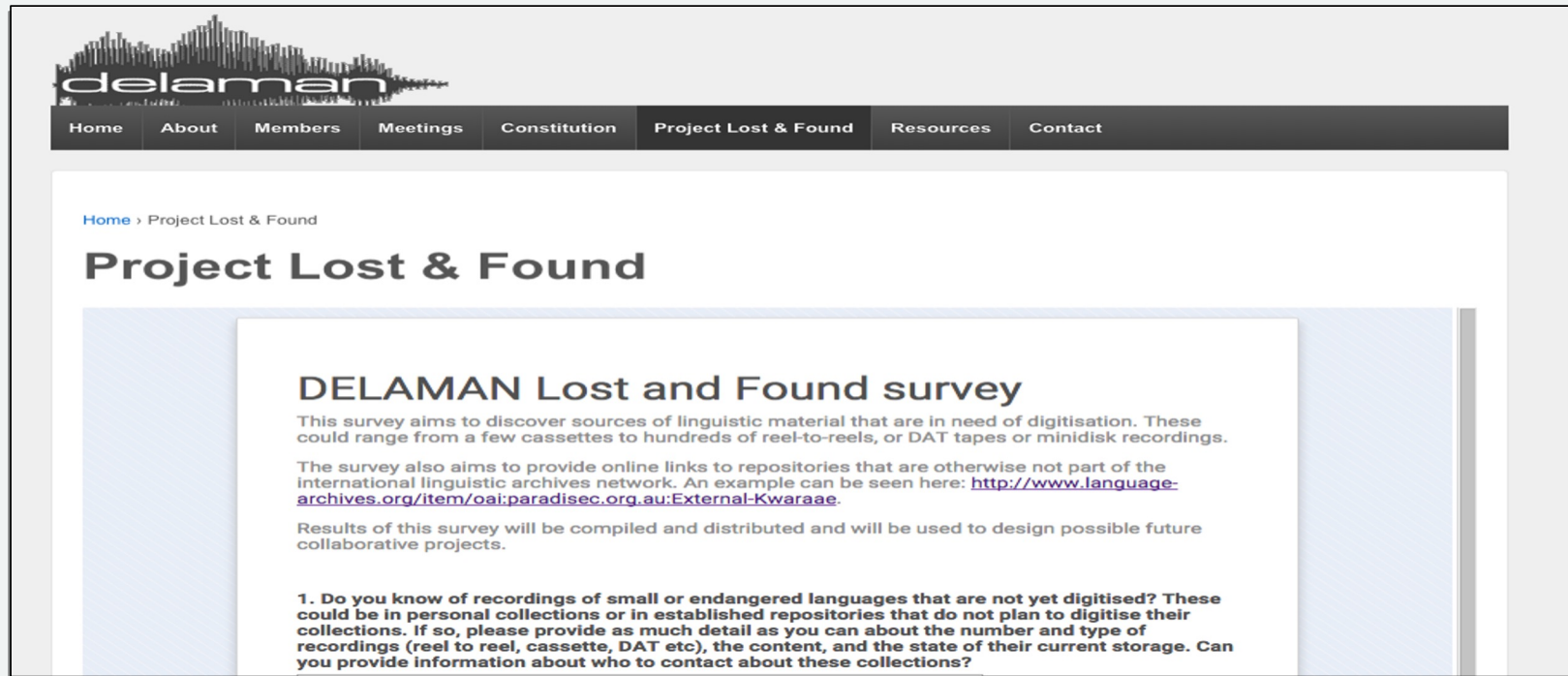
Creating archive-ready material in the field

Devise naming conventions for files

Use tools to help create structured metadata in the field



Looking for analog tape collections 2013 -



The screenshot shows the website for DELAMAN's Project Lost & Found survey. At the top is the DELAMAN logo, which features a stylized waveform above the word "delaman". Below the logo is a dark navigation bar with white text for the following menu items: Home, About, Members, Meetings, Constitution, Project Lost & Found, Resources, and Contact. The main content area has a breadcrumb trail "Home > Project Lost & Found" and a large heading "Project Lost & Found". Below this is a section titled "DELAMAN Lost and Found survey". The text in this section describes the survey's purpose: to discover linguistic material in need of digitisation, such as cassettes, reel-to-reels, DAT tapes, or minidisk recordings. It also mentions that the survey aims to provide online links to repositories not part of the international linguistic archives network, with an example link provided: <http://www.language-archives.org/item/oai:paradisec.org.au:External-Kwaraae>. The text states that the results will be compiled and distributed to design future collaborative projects. A numbered list item follows: "1. Do you know of recordings of small or endangered languages that are not yet digitised? These could be in personal collections or in established repositories that do not plan to digitise their collections. If so, please provide as much detail as you can about the number and type of recordings (reel to reel, cassette, DAT etc), the content, and the state of their current storage. Can you provide information about who to contact about these collections?"

Digital Languages and Musics Archives Network
<http://www.delaman.org/project-lost-found/>

Online catalogue: <https://catalog.paradisec.org.au>

The screenshot displays the PARADISEC Catalog interface. At the top left is the PARADISEC logo. The main header reads "PARADISEC Catalog" and includes a user profile "Nick Thieberger | Sign out". A navigation bar contains links for Home, Dashboard, Collections, Items, and Contact. Below the navigation bar are buttons for "Previous item", "Edit item", "Duplicate Item", and "Next item".

The "Item details" section on the left shows the following information:

- Item ID:** NT1-20002 (Collection Details)
- Title:** Audio recordings in Erakor village
- Description:** a: Toukolau Takau; John Maklen and William Wayne (rain noise); b Dick Lauto (see video version at NTS-DickLauto). (JPG files are images of the handwritten transcript of this audio file by Manuel Wayne).
- Origination date:** 2000-03-31
- Origination date free form:** 2000-04-04
- Archive link:** <http://catalog.paradisec.org.au/repository/NT1/20002>
- URL:**
- Collector:** Nick Thieberger (Find similar)
- Countries:** Vanuatu - VU (To view related information on a country, click its name)

The "Content Files (64)" section on the right features a "View file contents" link and a table of files:

Filename	Type	File size	Duration	File access
NT1-20002-001.jpg	image/jpeg	1.75 MB		View
NT1-20002-002.jpg	image/jpeg	1.82 MB		View
NT1-20002-003.jpg	image/jpeg	1.81 MB		View
NT1-20002-004.jpg	image/jpeg	1.74 MB		View
NT1-20002-005.jpg	image/jpeg	1.85 MB		View
NT1-20002-006.jpg	image/jpeg	1.84 MB		View
NT1-20002-007.jpg	image/jpeg	1.83 MB		View
NT1-20002-008.jpg	image/jpeg	1.83 MB		View
NT1-20002-009.jpg	image/jpeg	1.81 MB		View
NT1-20002-010.jpg	image/jpeg	1.82 MB		View
NT1-20002-011.jpg	image/jpeg	1.83 MB		View
NT1-20002-012.jpg	image/jpeg	1.84 MB		View
NT1-20002-013.jpg	image/jpeg	1.78 MB		View
NT1-20002-014.jpg	image/jpeg	1.84 MB		View
NT1-20002-015.jpg	image/jpeg	1.83 MB		View
NT1-20002-016.jpg	image/jpeg	1.84 MB		View

API feeds picked up by various services to increase reach of catalog, Digital Pasifik, NLA Trove, Open Language Archives Community (OLAC), Research Data Australia, google, etc

API from our catalog

PARADISEC API Documentation

The PARADISEC catalog exposes a number of APIs for harvesting the public data of the site. PARADISEC makes use of [OAI-PMH](#) for its harvesting APIs. The public data is also available through a [GraphQL API](#).

Harvesting Collections

RIF-CS
A RIF-CS feed is available at <http://catalog.paradisec.org.au/oai/collection>
For example, you can get a feed of all publicly available collections of PARADISEC:
<http://catalog.paradisec.org.au/oai/collection?verb=ListRecords&metadataPrefix=rif>
About RIF-CS: <http://ands.org.au/guides/cpguide/cpgrifcs.html>

Harvesting Items

OLAC
A OLAC feed is available at <http://catalog.paradisec.org.au/oai/item>
For example, you can get a feed of all publicly available items of PARADISEC:
<http://catalog.paradisec.org.au/oai/item?verb=ListRecords&metadataPrefix=olac>
Or you can get all the details of a single item of PARADISEC like this:
<http://catalog.paradisec.org.au/oai/item?verb=GetRecord&identifier=oai:paradisec.org.au:AA1-002&metadataPrefix=olac>
Just replace the item identifier with the identifier that you are after to get its metadata.
About OLAC: <http://www.language-archives.org/documents.html#Standards>

GraphQL API

[GraphQL](#)

<https://catalog.paradisec.org.au/apidoc>

Our catalog via
the National
Library of
Australia

The screenshot displays the Trove website interface. At the top, the Trove logo is visible, along with navigation links for 'Explore', 'Categories', 'Community', 'Research', and 'First Australians'. A search bar contains the text 'efate kalsarap'. Below the search bar, the results for 'NT1-20001 - Recordings in South Efate' are shown. The page includes a 'Data set - 2000' label, a large green placeholder image, and buttons for 'Get' and 'Cite this'. A 'Report culturally sensitive content' link is also present. The summary text at the bottom provides details about the recordings, including names like Toukoulau/Harris and Metu Josef, and mentions of time-aligned transcripts and handwritten transcripts by Manuel Wayane.

<https://trove.nla.gov.au/work/243668228>

Our catalog via Research Data Australia

The screenshot displays the Research Data Australia website interface. At the top, the logo for Research Data Australia is visible, along with navigation links for 'EXPLORE', 'ABOUT', and 'MYRDA'. A search bar is present with a dropdown menu set to 'All Fields' and a search button. Below the search bar, there is a checkbox for 'Publicly accessible online' and links for 'Advanced Search' and 'Map Search'. The main content area features the PARADISEC logo and the title 'NT1-20001 - Recordings in South Efate'. It indicates the dataset is funded by the Australian Research Council. A 'Dataset' tag and social media icons are also present. A large blue button labeled 'Access the data' is prominent. Below it are 'Cite' and 'Save to MyRDA' buttons. The 'Licence & Rights' section shows 'Open' access. The 'Full description' provides details about the recordings, including participants and content. A sidebar on the right lists 'Similar datasets you may be interested in' with several entries.

<https://researchdata.edu.au/nt1-20001-recordings-south-efate/1576056>

Our catalog via the Virtual Language Observatory

The screenshot displays the VLO interface for a specific record. The header includes the VLO logo, navigation links (Search, Contributors, Help), and the CLARIN logo. The breadcrumb trail shows 'VLO / Faceted search / Record: Recordings in South Efate'. The main title is 'Recordings in South Efate'. Below the title are tabs for 'Record details', 'Links (1)', 'Availability', 'All metadata', and 'Technical Details'. The 'Record details' tab is active, showing a table of metadata:

Name	Recordings in South Efate
Description	a: Toukolau/Harris (rowat) (vid); Metu Josef; b Kalsarur (also on video at NT5-KalsarurNawen); Kalsarap. Endis intonation. There are time-aligned transcripts of this item. NT1-20001. Story #107, Litrapog. Story #108, The spring at Epakor (video at NT5-Kalsarap). Story #117, Nkapmat go Nkapfag. Stories can be seen at NT8-TEXT. Handwritten transcripts by Manuel Wayane.. Language as given: Nafsan
Collection	Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) <input type="text"/>
Language	Bislama <input type="text"/> South Efate <input type="text"/>
Country	Vanuatu <input type="text"/>
Subject	language_documentation <input type="text"/> text_and_corpus_linguistics <input type="text"/>
Resource type	audio <input type="text"/>
Data provider	Other <input type="text"/>

To the right of the metadata table is a thumbnail image of a document icon with the number '20001' below it.

https://vlo.clarin.eu/record/oai_58_paradisec.org.au_58_NT1-20001

Automated processes mean we can run with limited staff

New items

Depositor fills out deposit form – access conditions

Depositor fills out metadata spreadsheet or uses the tool *Lameta*
(<https://lameta.org>)

Access determined by the depositor

Items and collections can be open or closed

All metadata can be 'private' so it is not visible to the public, but can be made visible to selected users

Some users can be given rights to view, edit, and download items that are otherwise closed

Automated processes mean we can run with limited staff

Files are uploaded into our system, automatically checked and go through a series of automated steps to get put into the collection

Transcode to archival format and delivery format (e.g., tif and jpg, wav and mp3, mov and mp4)

Deposited material can be online the same day or shortly after

With this automated system we are able to keep operating with minimal funding

Linked data standards and Research Object Crates

What is RO-Crate?

- lightweight approach to packaging research data with their metadata
- schema.org annotations in JSON-LD
- collection 'at rest' is completely self-describing

```
{
  "@context": "https://w3id.org/ro/crate/1.1/context",
  "@graph": [
    {
      "@type": "CreativeWork",
      "@id": "ro-crate-metadata.json",
      "conformsTo": { "@id": "https://w3id.org/ro/crate/1.1" },
      "about": { "@id": "./" }
    },
    {
      "@id": "./",
      "@type": ["Dataset", "Person"],
      "repositoryIdentifier": "http://catalog.cherokee.org/person/Marco_La_Rosa",
      "name": "Marco La Rosa",
      "identifier": "https://orcid.org/0000-0001-5383-6993",
      "description": { "@id": "http://catalog.cherokee.org/item/MyItem" }
    }
  ]
}
```

<https://www.researchobject.org/ro-crate/> - RO Crate spec

Home Advanced Search Transcription Search About

Show OCFL inventory file Show RO-Crate Show Data files

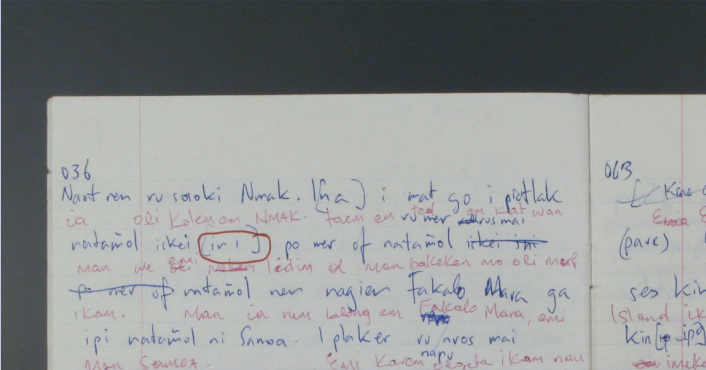
Versions: v1

Metadata Content

You have agreed to the conditions of access for viewing the content of

Images Audio XML Files

NT1-98007-001.jpg



Handwritten text on a piece of paper, likely a transcription or notes. The text is written in a cursive script. A red circle highlights a word in the middle of the page. The text is arranged in several lines, with some words written in red ink.

Home Advanced Search Transcription Search About

Show OCFL inventory file Show RO-Crate Show Data files

Versions: v1

Metadata Content

You have agreed to the conditions of access for viewing the content of this item. To review the conditions [click here](#).

Images Audio XML Files

NT1-98007-98007A

0:22 / 45:29

PLAY (00:20)
{SPEAKER=John Maklen} Tiawi nen ruto Enieltafra tetwei.

PLAY (00:23)
Ipiatlak natanio! rulap wes.

PLAY (00:29)
Ore tenen ruto go ruto nrus ruto nrus frafer rumai pak Erfat,

PLAY (00:33)
Kin nlaen maarik naot ni Enieltafra ipreglu namfer nen kin ruto preg nawesien sa

https://mod.paradisec.org.au/view/NT1/98007?ocfl_version=v1&type=audio#NT1-98007-98007A

Home Advanced Search Transcription Search About

Must match

Erakor

Title

Manuel Wayane

Description

+

Should match

Select a field to add to the query

- Title
- Description
- Contributor Name
- Date Created
- Date Modified
- University
- License

Must not match

Search

Total 3 < 1 >

1. Item
Audio recordings in Erakor village

a: Tou
(rain r
DickL
transc
/view/I

2. Item

Erako
Court
this it
Waya
single
/view/I

3. Item

Imag
Villag
Image
works
Endis

nasap


keyword search keyword search phrase search

Total 6 < 1 >

- nasap?

</view/NT5/200801?transcription=NT5-200801-2.eaf&begin=595.767&end=598.564#NT5-200801-2>
- Ko namër ni etog, nasap? nen rumai, ka fo tli tefla, nasap?, rupi nasap?

</view/NT1/20003?transcription=NT1-20003-20003A.eaf&begin=76.808&end=79.845#NT1-20003-20003A>
- Ko namër ni etog, nasap? nen rumai, ka fo tli tefla, nasap?, rupi nasap?



</view/NT5/DickLauto?transcription=NT5-DickLauto-Vid2.eaf&begin=76.808&end=79.845#NT5-DickLauto-Vid2>
- nana nafet nasap?, ruk taulu namër ni ena

</view/NT1/98014?transcription=NT1-98014-98014A.eaf&begin=706.586&end=712.345#NT1-98014-98014A>
- Go tete natamöl nen na kutae na, kulek-, pälek naklun mäs me kuipe tae na natamöl nen tu ipi natamöl nasap? nen tu.

</view/NT1/20003?transcription=NT1-20003-20003A.eaf&begin=876.385&end=883.12#NT1-20003-20003A>

<https://mod.paradisec.org.au/>

Returning files



JL5-002_4_2_SA2.wav
JL5-002-Ma_2_1a_SAR1.wav
JL5-002-Ma_2_1b_SAR1.wav
JL5-002-Ma_2_2_AS10.wav
JL5-002-Ma_3_1_AMAS1.wav
JL5-002-Ma_3_2_IN2.wav
JL5-002-Ma_4_1_AA12.wav
JL5-002-Ma_4_2_NA2.wav
JL5-002-SH_1_1_MAD1.wav
JL5-002-SH_1_2_MH4.wav
JL5-002-SH_2_1_SMH1.wav
JL5-002-SH_2_2_FAD1.wav
JL5-002-SH_3_1_FM2.wav
JL5-002-SH_3_2_SH4.wav
JL5-002-SH_4_1_AZ3.wav
JL5-002-SH_5_1_AM6.wav
JL5-002-SH_5_2_MN7.wav
JL5-004-SH_4_0.wav

PARADISEC collection

42,266 items

440,332 files

Subcollections

Copying the files leaves them orphaned, with no catalog description.

Research Object Crates enable the creation of subcollections. For example, all items related to a particular village and putting them on a hard disk, with a catalog of just that set of items.



Raspberry Pi is a low powered webserver
wifi transmitter

The USB drive stores the data

Delivery of archival files avoiding bandwidth cost

We wrote a service (dataloader) to convert sets of
items from PARADISEC into a new sub-collection,
with a catalog, loaded to hard disk or to a Raspberry Pi

<https://www.paradisec.org.au/blog/2024/02/using-raspberry-pi-in-ranongga/>

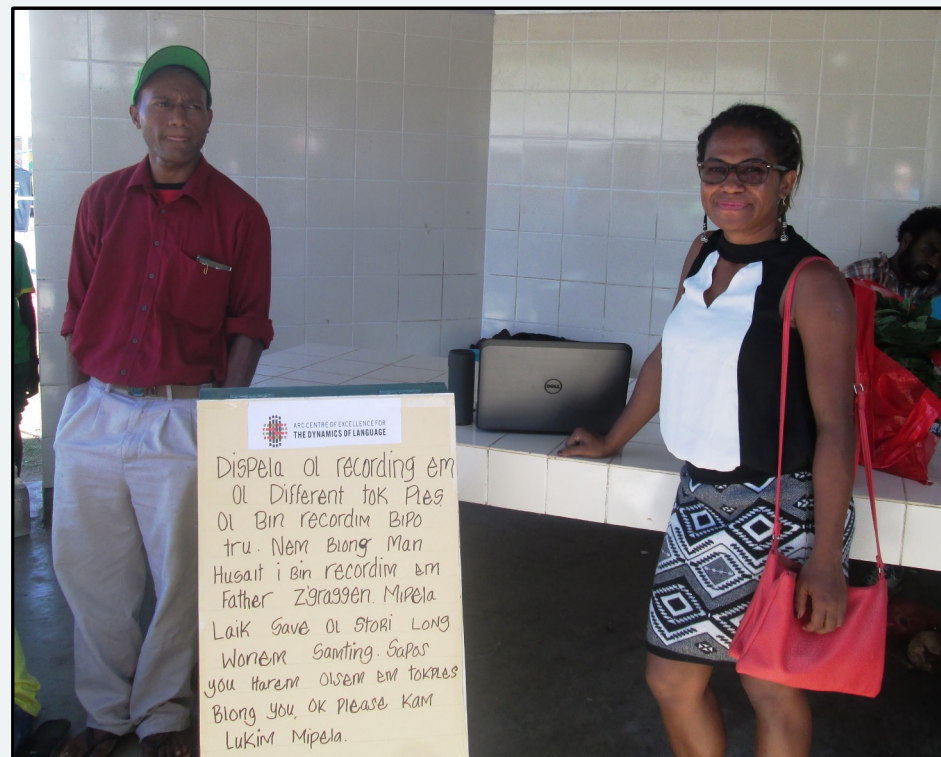


Elisa Alphonse in Erakor village



Kalonuk Albert & Gino Joseph in Erakor village

Enriching metadata



VIRTUOUS CIRCLE

- Research recordings are made available for the source communities and re-used, perhaps to learn forgotten traditions, or just to hear ancestors talking
- Research is properly grounded in the recordings, that can be cited and re-used
- New research can be based on recordings that have licences for re-use
- Collaboration with museums and archives in the Pacific that need help preserving analog tape.

Thanks to:



Australian Research Council DPo450342, DPo984419,
FT140100214, CE140100041, LE220100010



Australian Research Data Commons



Language Data Commons of Australia