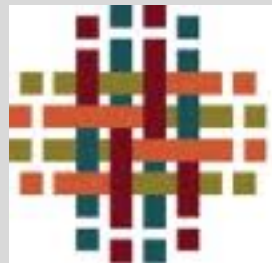


Language Documentation Methods

Nick Thieberger
Research Unit for Indigenous Language
School of Languages and Linguistics
The University of Melbourne

ARC Centre of Excellence for the Dynamics of Language



I acknowledge and pay my respect to the Woiwurrung original owners of the land we live and work on

Outline

9am - 10 am Linguistic data management, introducing
Lameta

Break

10.15 -11.15 Elan intro

Break

11.30-12.30 FLEEx intro

Break

1.00pm - open session for practical work and questions
(if needed)

Please sign up

- <https://go.coedl.net/LDocTraining2020>

Install:

Elan

<https://archive.mpi.nl/tla/elan/download>

Lameta

<https://github.com/onset/laMETA/releases>

FLEx. (Choose SE Minimal)

<https://software.sil.org/fieldworks/download/>

Audacity

<https://www.audacityteam.org>

Course overview

Why linguistic data management ?

Metadata

- File-naming conventions

- How much metadata is enough?

- Available linguistic metadata sets and how to select among them.

- Lameta – tool for creating metadata descriptions

The linguistic fieldwork workflow

- How each tool fits into the workflow (transcription, annotation, corpus-development, lexicon, interlinear glossed texts, *data conversion using regular expressions).

Processes and current tools

- Transcription with time-alignment

- Annotation

- Lexical database

- Interlinear text production

What is 'well-formed data'?

- Distinguishing the form and content of data to allow multiple outputs from the same underlying data

Principles

- Scale of data collection we are typically engaged in requires automating as much as possible of our workflow
- To automate our work we need to understand the nature of the data and the tools we can use to work with it

Principles

- Automating means getting the computer to do most of the work
 - Time-aligned transcripts
 - Annotation
 - Creating a corpus of texts
 - Creating a lexical database
 - Keeping track of all of that

Me, selwan ag kupi eñae, tiawi itraus
traus traus traus, natrauswen ga itaos
nlag. Itrausi pan kaipa. Me komam uta laap
kin uto mau, a? Malen umat, inom.

*But when you are far away the old man can
talk and talk and talk, his story is like
the wind. He tells it and it is gone. But
there aren't many of us left. When we die,
it will be finished.*



(NT1-98009-A, 1932.36 1942.16)

†Kalfañun Mailei, 1998, Erakor Village, Efate, Vanuatu

Methods

- The methods described here need not add too much to your work
- They will add a great deal to what you can do with your work
- You need to choose how much you can do, but at least be aware of what is offered by new and emerging methods

Methods

- Ask for advice – use networks like the RNLD/Living Languages mailing list
- If you know that some work that would take you two weeks can be done by someone with experience of the tools in a few minutes, then
 - learn those tools
 - or find that person!

Well-formed data

- Will endure into the future
- Will be reusable
- Characteristics
 - Text - Structured: Content and form separated
 - Media in standard and archival formats

Separating content and form

\lx faat
\ps n
\sn 1
\ge stone
\de stone
\ng ston
\xv Nskau nen iṗur ki faatfar.
\xe The reef has many calcium rocks.
\se faat ftak
\ps n
\de bottom stones of an oven
\se faat ni uum
\ps n
\de top stones of an oven
\xv Nafet faat ni uum rumiel malen ruftin.
\xe The top stones are red when they are heated.
\nt DTryonNumber:114
\nt DTryon:stone
\nt JCR:fat faar - calcium rock
\nt DTryon Form: fat
\so 001b

\sd natural_features
\sn 2
\ps n
\ge vatu
\de vatu, money
\ng mane
\so 086:013 NT1-98010-B.wav 512.634 520.218
\xv Ale ipan sor tete sernale hotel, nen kesol faat sees.
\xe Then he goes and sells something at the hotel, to get a little money.
\sd human_artefact
\dt 20/Apr/2011

faat *n.* 1) stone. **Nskau nen iṗur ki faatfar.** The reef has many calcium rocks.
faat ftak *n.* bottom stones of an oven.
faat ni uum *n.* top stones of an oven. **Nafet faat ni uum rumiel malen ruftin.** The top stones are red when they are heated. 2)
— *n.* vatu, money. **Ale ipan sor tete sernale hotel, nen kesol faat sees.** Then he goes and sells something at the hotel, to get a little money.

We want

- To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)
- Excellent data will endure, can be accessed and can be made into various forms for delivery to various users

Central tenet of language documentation that we create
data for posterity
Formats need to have longevity

- To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)
- Excellent data will **endure**, can be accessed and can be made into various forms for delivery to various users

It has to be described using standard terms in a catalog
that can be searched

It has to have persistent identification

Permission to use it

- To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)
- Excellent data will endure, **can be accessed** and can be made into various forms for delivery to various users

Automatically (not laboriously by hand)

For our own research into the future

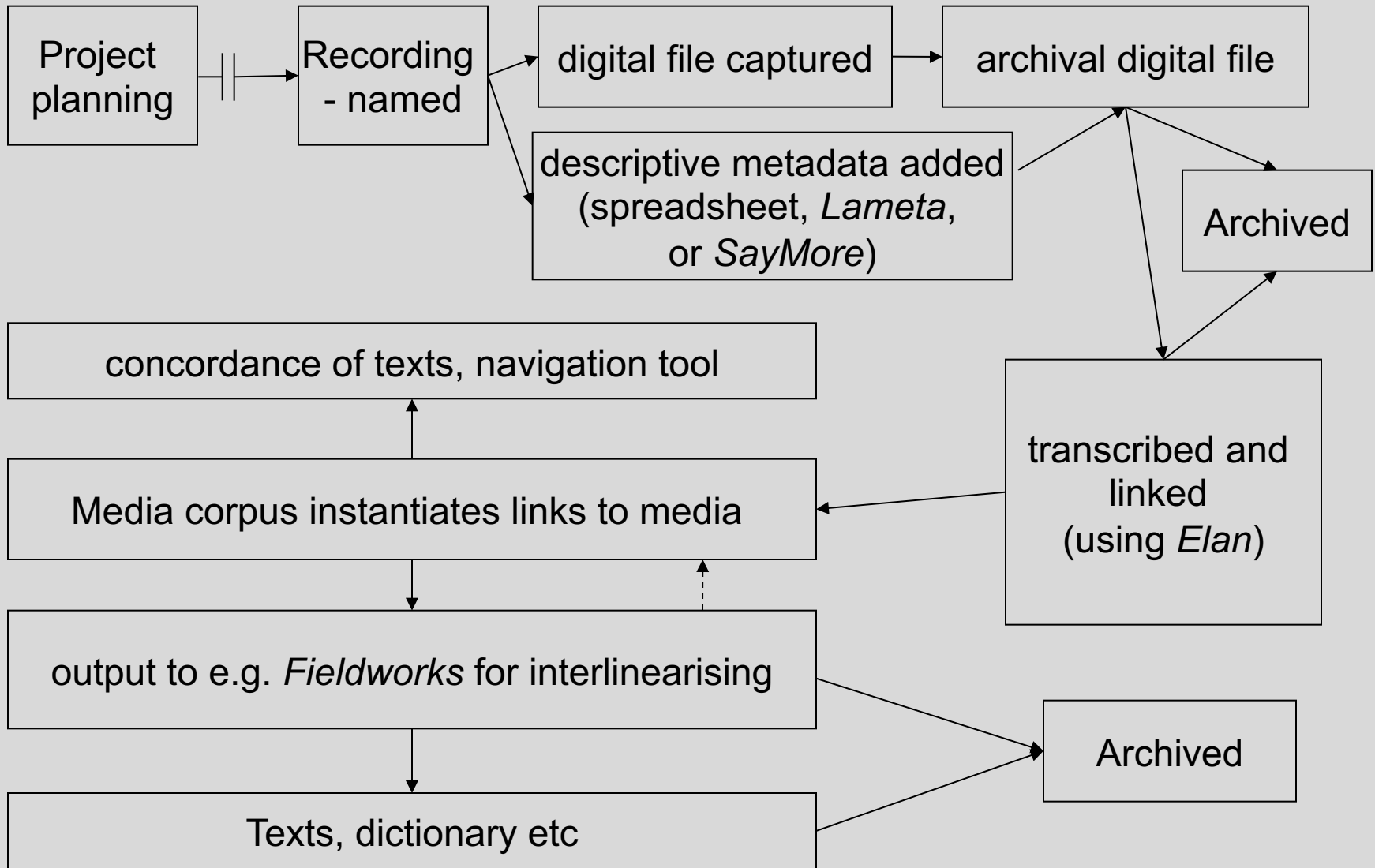
So that others can produce material based on our data

- To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)
- Excellent data will endure, can be accessed and **can be made into various forms** for delivery to various users

What tools and processes do we need to do this?

- Tools that do not lock our data into proprietary formats
- Processes that allow our data to flow from recording through transcription to analysis and further reuse

Typical workflow resulting in well-formed data



Working form

The form in which information is stored as it is created and edited

Archival form

The form in which information is stored for access long into the future

Presentation form

The form in which information is presented to the public

Working form

The form in which information is stored as it is created and edited

Can include notes, not all of which may be useful later

E.g., files being processed (in Elan, Toolbox or Flex) and ancillary files (.typ, .prj, pfsx, etc) that are only necessary while working to create the annotated data

Archival form

The form in which information is stored for access long into the future

Highest resolution form of the data

Presentation form

The form in which information is presented to the public

Derived from working or archival form.

May be compressed and arranged to make it easier to deliver and interpret

Issues with workflows

- Data is primary and needs to be able to flow through the various tools
- Need to adapt as new tools are developed
- Understand how each tool deals with the data

Issues with workflows

- Lossless conversion of primary data is crucial

BUT derived views of the primary data *can* be compressed and reduced for delivery, e.g.,

- Subtitled video clips may not have interlinear glosses
- Dictionaries in various forms derived from lexical databases
- Streaming media is in lower resolution formats

Filenaming

- What needs to be named? What constitutes an *object* for our purposes?
 - Primary data
 - Tapes
 - Files - audio, video
 - Derived items
 - Transcripts
 - Texts

Filenaming

- Select a convention that works for you
- Find out what the archive uses
- Be parsimonious (don't get carried away!)
- Use simple characters
 - Various computer systems will have to be able to read your filenames

Filenaming

Simplicity is best:

200601A.wav

NT1-200601-A.wav

Not so good:

ERK	200512	ERAKORTOKELAUF
-----	--------	----------------

.wav

Language code

Date

Place

Person

Gender

Filenaming

Every citation of a piece of data should be able to resolve to that data

Filenames must be consistent over time:
“Persistent Identification”

References:

PARADISEC filenaming:

<http://www.paradisec.org.au/deposit/file-naming/>

File naming

- Unique names: digital recorder names STE-001, and your camera will produce images with names like IMG_0086
- Take note of what the name is in your fieldnotes and then be sure to assign the correct metadata to them when you can
- Be aware that hyphens and underscores (and other non-alphanumeric characters) may be 'reserved' characters in some archives, so they will need to be converted later
- To ensure the greatest legibility and persistence of your filenames it is still best to use ASCII characters
- Be consistent in using upper and lower case—for some computer systems upper and lower case characters are treated equally, but in others they are not
 - ideally all extensions should be lowercase (e.g., .wav not .WAV)

Principles

- Scale of data collection we are typically engaged in requires automating as much as possible of our workflow.
- To automate our work we need to understand the nature of the data and the tools we can use to work with it

Metadata

Use terms that allow you to find your records:

- Date created
- Place (lat / long; placename)
- People involved and their roles
 - speaker, singer, recorder ...
- Kind of record (text, sound, moving image ..)

Look at what existing archives use

Metadata

Be consistent!

Use the same format all the way through your metadata

Dates: 20190215, 2019-02-15

Names, first last ; last, first

Places – always spelled the same

A database system can help enforce consistency

Principles

Automating means getting the computer to do most of the work

- Time-aligned transcripts
- Annotation
- Creating a corpus of texts
- Creating a lexical database
- Keeping track of all of that

Principles

- Reuse
 - Of primary data
 - Of derived analysis
- By us
- By others, including the speakers/performers

Timeliness

- Leaving some tasks until later places an intolerable burden on the linguist who then does not manage to catch up with the backlog
 - types of tasks include: filenaming; entering metadata into a standard format (preferably a database management system)

Project management

- Planning fieldwork and follow-up
- What kind of information will be recorded?
 - notes
 - audio/video
 - narratives; multi-participant conversations; songs
 - still images
 - locations – GPS
 - genealogies
 - Other?

Project management

- What kind of equipment will be required?
 - e.g., music may require a better mic than spoken word
 - hard disks
 - backup equipment in case of failure?

Research equipment, get advice

Planning

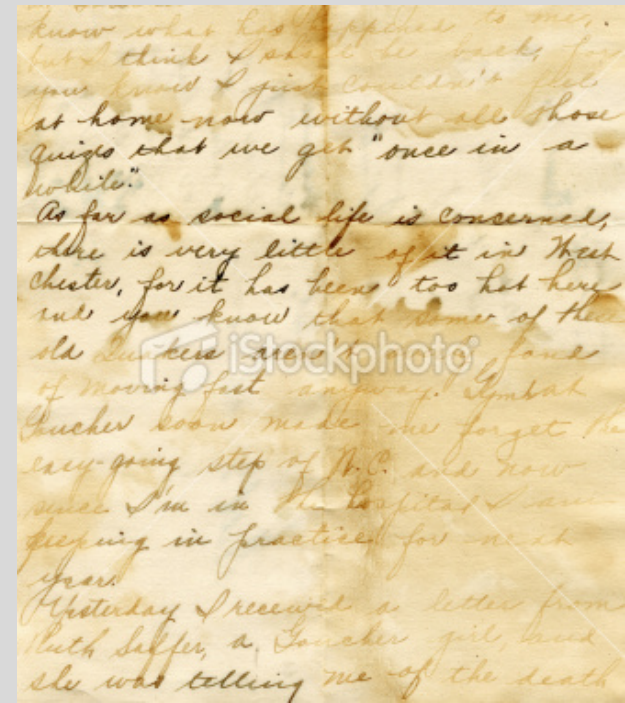
- Consent forms for speakers to sign
 - Make it clear what can be done with your recordings and other outcomes of your fieldtrip

Planning

- Recording methods
 - learn to use your equipment, what it can and can't do
 - in different conditions – background noise, wind etc

Planning

- Recording methods
 - Write notes, using a waterproof pen! Carry notebooks:
 - identify each recording in your notes
 - say 'who, when' intro to recordings
 - keyboard and scan written notes
 - use a laptop / portable device?

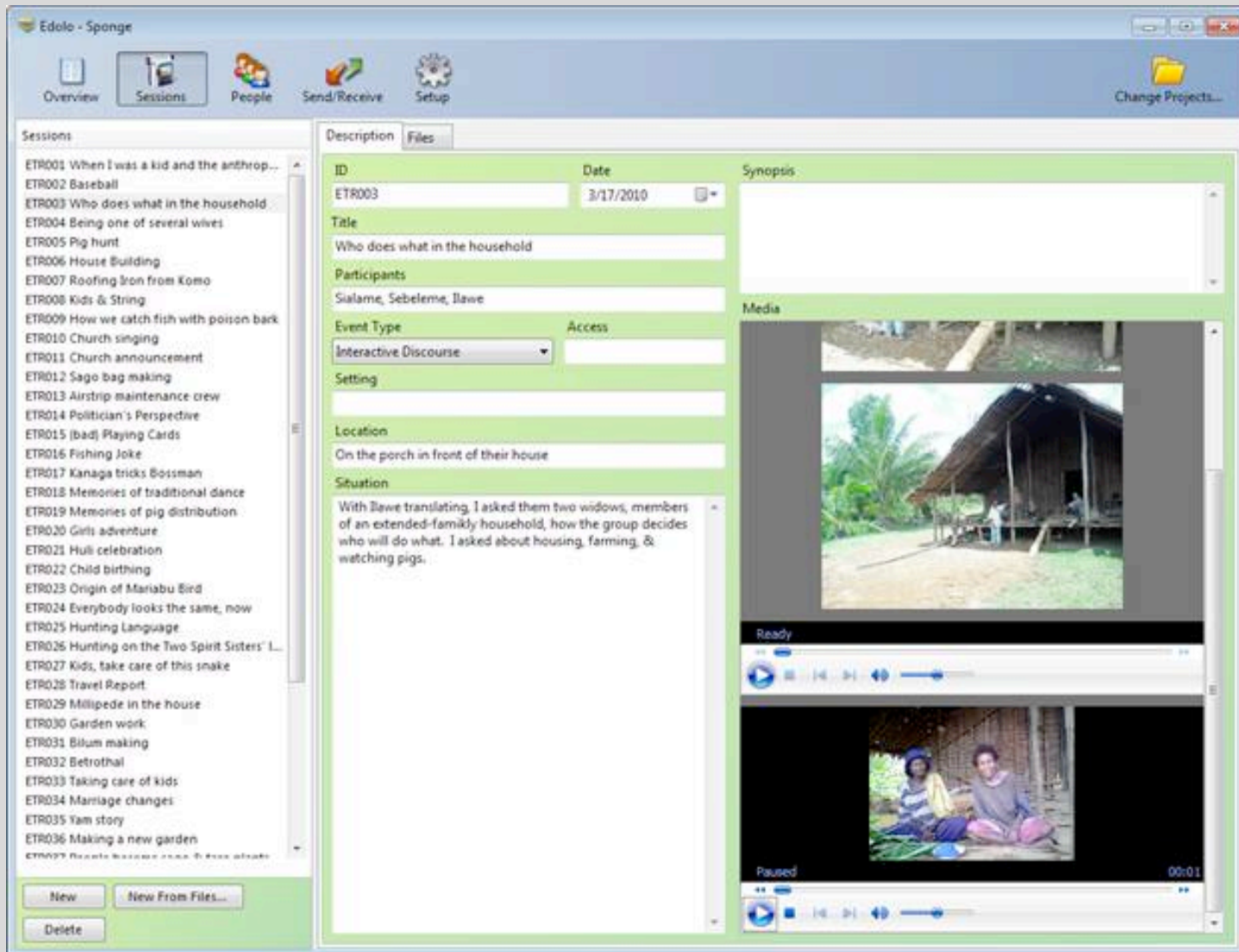


SayMore and Lameta

- Metadata entry, file management

Saymore

<https://software.sil.org/saymore/>



Saymore

<https://software.sil.org/saymore/>

C:\Users\John\Documents\SayMore\EdoloSample\EdoloSample.sprj - SayMore

File

Sessions People Progress

People

Name	Type	Date Modified	Size	Duration
Awile Heole.person	Person	6/24/2010 3:35:5...	782 B	
Awile Heole_Consent.JPG	Image	6/24/2010 3:13:2...	207 KB	
Awile Heole_Photo.JPG	Image	3/20/2010 5:41:1...	1.05 MB	

Add...

Person Notes

Full Name: Awile Heole

Birth Year: 36/4

Primary Language: Edolo

Gender: Male

Learned In: Huya


Other Languages: Tok Pisin, Huli

Education: Grade 2

Primary Occupation: Subsistence Farmer

How to Contact:

New Delete



Saymore

<https://software.sil.org/saymore/>

C:\Users\John\Documents\SayMore\EdoloSample\EdoloSample.sprj - SayMore

File

Sessions People Progress

Sessions

Name	Type	Date Modified	Size	Duration
ETR009.session	Session	6/24/2010 3:28:5...	733 B	
ETR009_Careful.wav	Audio	3/27/2010 10:52:...	352.79 MB	00:21:24
ETR009_Careful_Part2.wav	Audio	3/27/2010 11:05:...	159.09 MB	00:09:39
ETR009_MonoExtract.wav	Audio	3/27/2010 9:57:0...	89.53 MB	00:10:51
ETR009_Original.MOV	Video	3/20/2010 5:25:4...	505.76 MB	00:10:51
ETR009_Original.wav	Audio	4/2/2010 7:50:15 ...	59.68 MB	00:10:51
SceneAroundCamera.JPG	Image	3/20/2010 5:45:3...	1.06 MB	
SceneHouse.JPG	Image	3/17/2010 11:45:...	1.05 MB	

Add...

Session Notes

Id
ETR009

Date
6/24/2010

Title
The story behind how we catch fish with poison bark

Setting
Sitting on their fron porch

Participants
Awi Heole, Ilawi Amosa

Location
Huya

Event Type
Narrative

Access
Open

Situation
Walking with Ilawi, we passed the plant they use to make poison for fishing, and I asked him if he'd tell me about how to do it. We went to his house, where Awi joined us, and the two of them took turns telling parts of the "whole story".

Synopsis

Custom Fields

Field	Value
*	

New

New From Files...

Delete

Lameta (2020)

<http://go.coedl.net/lameta>

Nafsan/ Nafsan - Digame 0.8.3 Beta

Project Sessions People

ID	Title
20200212Im=ages	Pics from Erakor

Name	Type	Modified	Size
20200212Im=ages.session	Session	2020-02-22T06:43:39.6...	19 B
IMG_3997.JPG	Image	2020-02-22T06:43:46.6...	4 MB
IMG_3998.JPG	Image	2020-02-22T06:43:46.6...	4 MB
IMG_3999.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4000.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4001.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4002.JPG	Image	2020-02-22T06:43:46.6...	3 MB

Session Contributors Status Notes

ID20200212ImDate2020-02-12YYYY-MM-DD

TitlePics from Erakor

GenreSubgenre

AccessAccess Explanation

Content LanguagesSouth Efate

Working Languages

Custom Fields

People

Description

TopicKeyword

Neighborhood/Town/Village

More Fields

Field	Value
Researcher Inv...	
Region	
Country	

New Session


Nafsan/ Nafsan - Digame 0.8.3 Beta

Project Sessions People

ID	Title
20200212Im=ages	Pics from Erakor

Name	Type	Modified	Size
20200212Im=ages.session	Session	2020-02-22T06:43:39.6...	19 B
IMG_3997.JPG	Image	2020-02-22T06:43:46.6...	4 MB
IMG_3998.JPG	Image	2020-02-22T06:43:46.6...	4 MB
IMG_3999.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4000.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4001.JPG	Image	2020-02-22T06:43:46.6...	2 MB
IMG_4002.JPG	Image	2020-02-22T06:43:46.6...	3 MB

Image Properties Contributors Notes



Lameta

- Export of metadata to csv, xml formats relevant to each archive
- Free, still in development

