

Digital Archiving for Documentation of Endangered Languages

David Nathan

Endangered Languages Archive
SOAS University of London

3L Summer School, Lyon

July 9, 2011



Contents

- Archiving principles and concepts
- Data management
 - strategies
 - organising files
 - file naming
 - formats and encoding
 - metadata
- Archiving with ELAR
- Mobilisation of digital resources for language support



Archiving principles

- general archiving functions
 - acquire and preserve
 - add value
 - provide access
 - develop trust

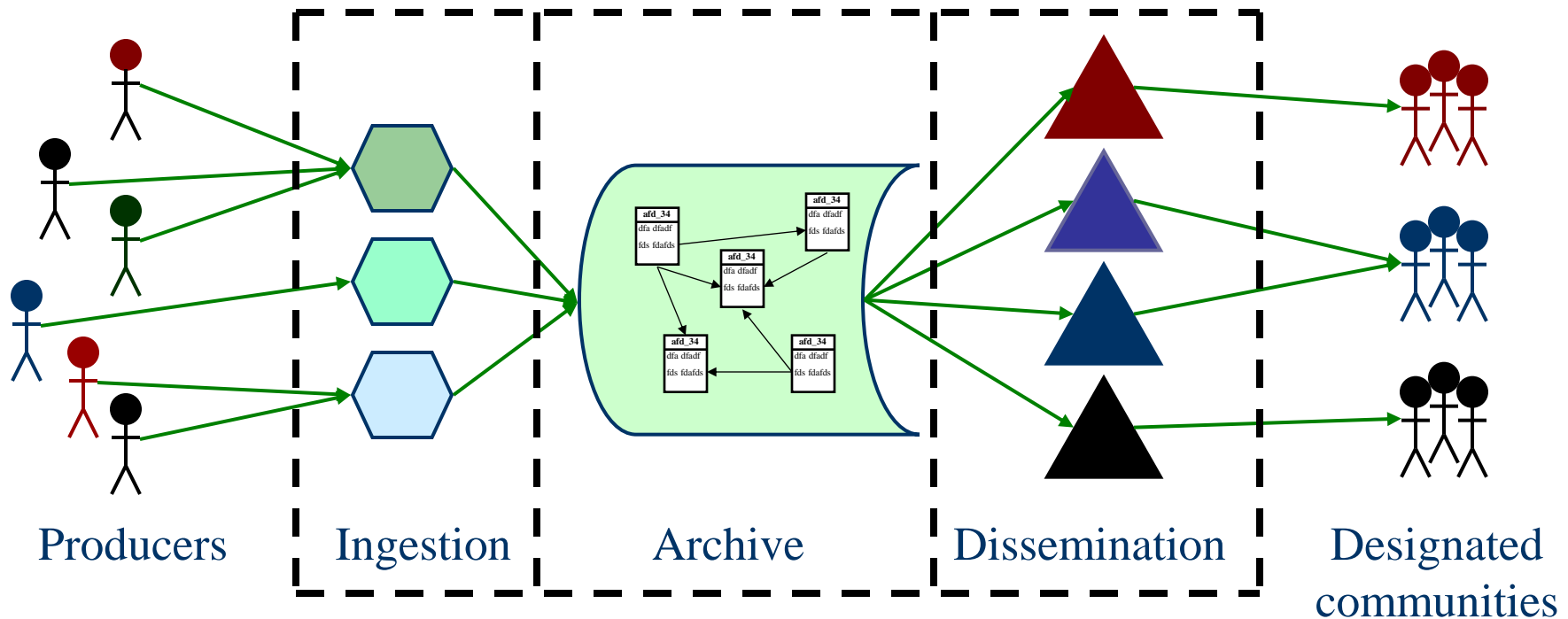


EL digital documentation archiving principles

- acquire and preserve
- support and curate
- develop trust
 - with depositors and users; via bodies and standards eg Data Seal of Approval, Ninch
- publish

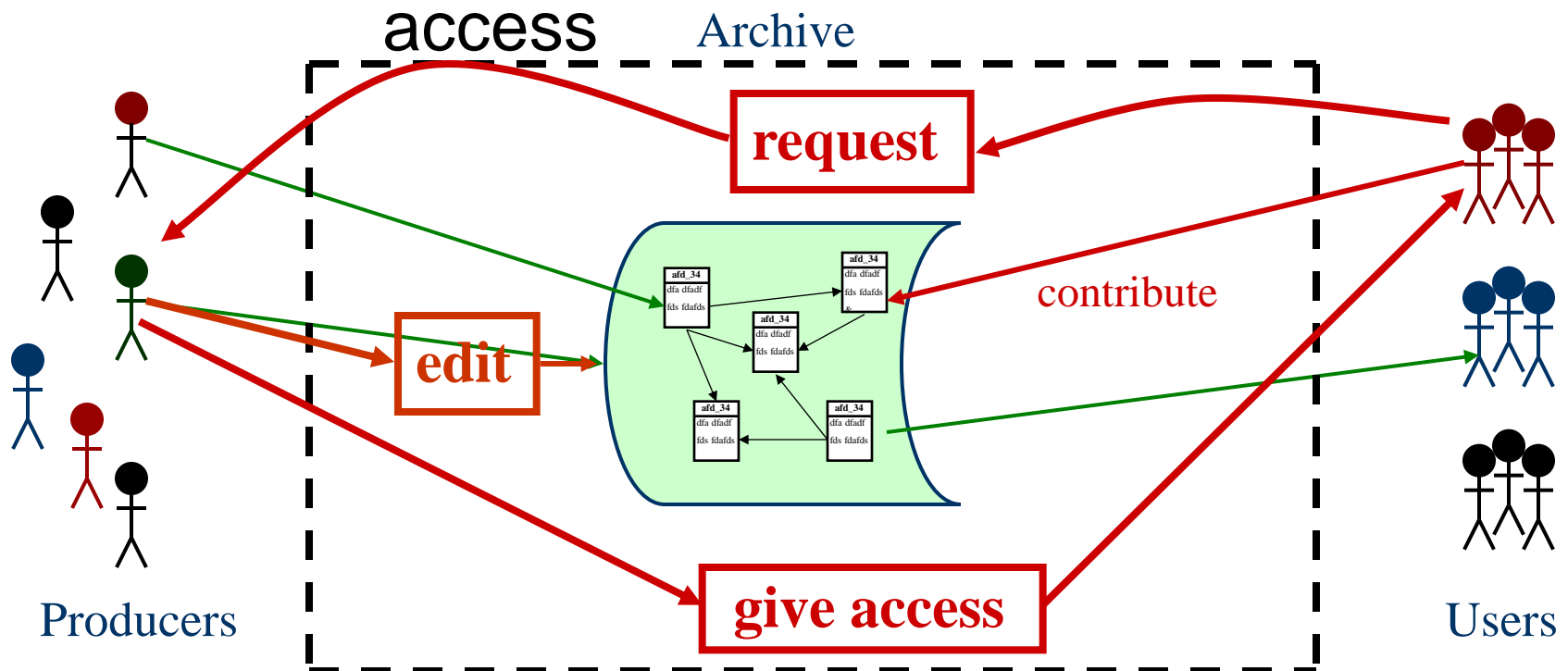
OAIS model

- OAIS archives define three types of 'packages'
ingestion, archive, dissemination:



ELAR - architecture

- Boundary between depositors, users and archive:
 - users add, update content; negotiate





Redefining the digital EL archive

- a platform for developing and conducting *relationships* between knowledge *producers* and knowledge *users* – a social networking archive
- level the playing field between researchers and community members/other stakeholders
- encourage, recognise and cater for *diversity*



On diversity, skills and practices

- on diversity, tools, standards, archivism, and the boundaries between archiving, documentation and researcher skills



Data management and archiving

- the following slides about data management are NOT strictly speaking part of archiving
 - documenters should use good data management practices whether or not they plan to archive their materials
 - good data management practices will make a future archiving process easier and better



Data management: 3 most valuable strategies

- work out your fundamental units of documentation and the relationships between them
 - design and use a filename system
 - choose “tools” to suit your purpose, desired outcomes, skills, working styles , existing materials, context
- if you get these right, they will do the “heavy lifting” of your data management strategy



Data management: 3 more strategies

- use appropriate and conventional data encoding methods (e.g. Unicode)
- be explicit and consistent
- document steps, decisions, conventions, structures
- plan for flow of data, working with others, across different systems
 - see Bird and Simons, 'Seven Dimensions of Portability'



Data management: 2 important (non) distinctions

- a spectrum: *data* and *metadata* are intertwined, points in a spectrum rather than different things
- a distinction: distinguish between *machine readable* information and other information



Data management: 3 important qualities

- machine readability

computer programs can act on your data in terms of its proper structures and categories

processes will be scalable (will work equally well on large quantities)

example example example example example example

- consistency
- documentation of conventions, structures, methods



Tell-tale signs of likely problems

- information carried by
 - colours
 - typography (italic, font, size etc)
 - MS Word document
- conflation of
 - information
 - information structure
 - presentation



Simple strategies and checks

- if you want to, you can use MS Word to prepare documents, but:
 - use ¶ to show all hidden formatting etc
 - transfer regularly (and finally) to plain text
 - use defined styles properly
- view materials in plain text and/or in a browser
- if using spreadsheet, sort columns and check for consistency of values



Managing data in your computer

- design a well-organised system of folders and files so that you (and others) can always find your stuff according to *what it is*, **not**:
 - where the software decided to put it
 - what the software decided to call it
 - when/where you last used it
 - what someone else called it
- design so that you will *always* be able to find things



Organising files

- design folder structure as a logical hierarchy that suits your goals, content and work style
- recommendations:
 - have materials gathered within one overall folder (e.g. for backup)
 - make folders for *relevant categories*, e.g. sessions, media types, participants, dates, or ...
- you may need to *restructure* at different points in your project, e.g. move from



On identifiers

- real world objects are inherently identified because they are physically unique - an unlabelled cassette is only *poorly* identified
- digital objects have no such physical independence - they depend on the *identifiers* that we give them
- three types of identifiers:
 - *semantic*
 - *keys*
 - *relative*



On identifiers

- *semantic*, e.g.
 - Nelson Mandela
 - The Sound of Music
 - SA_JA_Bongo_Palace_Land Dispute Trial_015_29-04-2010.wav *

* SA_JA_Bongo_Palace_Land Dispute Trial_015_29-04-2010.wav



On identifiers

- *keys* (disambiguators), e.g.
 - 1137204 (a student number)
 - 0803 211 6148 (a telephone number),
p12893fh23.pdf (some system's
reference number)



On identifiers

- *relative*, e.g.
 - 67 High Street
 - the secretary
 - index.html
 - metadata.xls



On identifiers

- your collection will have a mix of these but it is important to be aware of the differences and limitations, for example:
 - semantic identifiers: invite name clashes
 - keys: a program or process might depend on the identifier to work properly
 - relative identifiers: if you move them you typically change or destroy their meaning



Digital objects and identities

- a file's identity = path + filename
- the path is a representation of the volume and the directory (folder) hierarchy
- if the full identity is *unambiguous* then everything can be fine, compare:
 - c:\\dogs\\spaniels\\rover.jpg
 - c:\\cars\\british\\rover.jpg
 - or
 - lectures\\syntax\\20091103\\lecture.doc



Objects and identities

- semantic identifiers are potentially ambiguous, because just adding more chunks to disambiguate them will not work:
 - `my\rover.jpg`
 - `my\white_rovers.jpg`
- so objects that are not semantically unique need identifiers which are either keys, or relative



Segue to file names

- (having said all that)
- filenames are filenames, and do not necessarily identify other entities
- common mistaken assumptions:
 - a filename “dp_verbs_39.wav” means there is an entity “dp_verbs_39”
 - files are linked by sharing some part of their filenames
 - eg PL_conv_02.wav, PL_conv_02.txt



File naming

- we tend to be unsystematic in naming files. This *might* be OK, if you have a large amount of files and a method that already does everything you need to do (and will need to do in the future)
- but filenames that are unsystematic or are non-standard will cause problems, eventually



Filename “good practice”* rules

- all filenames should have correct extensions
- each filename should have only one ".", before the extension
- do not use characters other than letters, numbers, hyphen - and underscore _
- avoid non-ASCII characters
- keep filenames short, just long enough to contain the necessary identifier - don't fill them up with lots of information about the content (that is metadata!)



How about these file names?

1. ready.audio.wav
2. ReALLyOdDtOReAd.txt
3. éclair.jpg
4. éclair_fr.jpg
5. e'clair.jpg
6. french-cake.jpeg
7. french-cake.jaypeg
8. lexicon-master
9.   I  .eaf
10. ice cream.doc
11. OBAMA.TXT
12. Obama.txt



Make filenames sortable

- make filenames usefully sortable:
 - 20100119lecture.doc
 - 20100203lecture.doc

gr_transcription_1.txt	gr_transcription_001.txt
gr_transcription_2.txt	gr_transcription_002.txt
gr_transcription_9.txt	gr_transcription_009.txt
gr_transcription_53.txt	gr_transcription_053.txt



Associating files

- you can make resources sortable together by giving them the same filename root (the part before the extension), or part of the root

- *document* if you mean to establish

gr_reefs.wav

gr_reefs.eaf

gr_reefs.txt

paaka_photo001.jpg

paaka_photo002.jpg

paaka_txt_conv203.wav

paaka_txt_conv203.eaf

paaka_txt_lex.doc



Avoid metadata in filenames

- avoid stuffing *metadata* into filenames. A filename is an identifier, not a data container
- better to use a simple (semantic) filename or a key (i.e. meaningless) filename, and then create a metadata table to contain all the relevant information
- a table can properly express all the information, contain links etc, and is extensible for further metadata



Avoid metadata in filenames

- e.g. Paaka_Reefs_Dan_BH_3Oct97.wav
- better:
 - paaka_063.wav
plus
 - paaka_063.txt

filename: paaka_063.txt

language	topic	speaker	location	date
Paakantyi	Reefs at Mutawintyi	Dan Herbert	Broken Hill	1997-10-03



A file naming *system*

- carefully design a filename **system** for your data and *document the system so that somebody else can understand it*
- one documenter's new system:

aaa_bb_cc_yyyy-mm-dd_nnn.wav



A filenaming *system*

- `aaa_bb_cc_yyyy-mm-dd_nnn.wav`
 - aaa = village **code**
 - bb = (main) speaker **code**
 - cc = genre/event **code**
 - yyyy-mm-dd = date (why this order?)
 - nnn = optional number (e.g. 001)
 - .wav = correct extension for file content type



Documenting the filename system

- describe the system
 - how would you describe it?
 - where would you put the description?
- document the **codes** – this is probably part of your metadata



On changing file names

- decide if it's possible, benefits and side effects (e.g. loss of links in ELAN files)
- design a system first
- don't change names *in situ* – copy data set and gradually migrate it to your new system
- document file name changes



Tools for listing and changing filenames

- if possible, automate or copy/paste filenames
- if possible, use machine processes, e.g. filename listings, XLS formulas, filenaming utilities
 - pFrank
 - Karen's Directory Printer
 - DOS *cmd*
 - *Run* (Windows + R)
 - type *cmd* to open "DOS box"



STOP! did you first model your data?

- to model = to explore and be explicit about ontology
- even a cursory attempt will benefit your project
- Lenore Grenoble's example:
 - Greenlandic names
 - Latin (scientific or binomial) name



And then

- ideally, following modeling, you work out how to represent and manage the information in terms of your model, using some of:
 - file organisation and names
 - tables with rows and columns (relational)
 - tagged data
 - multipurpose software such as spreadsheets, databases, XML authoring
 - (if appropriate) specialist software

filename: sessions.xls

ID	audio	transcription
1	TRS00065.wav	bjt_02.txt
2	TRS00066.wav	krs_43.txt

relational

filename: [sessions.xml](#)

```
<sessions>
  <session id="1">
    <audio>TRS00065.wav </audio>
    <transcription>bjt_02.txt</transcription>
  </session>
  <session id="2">
    <audio>TRS00066.wav</audio>
    <transcription>krs_43.txt</transcription>
  </session>
</sessions>
```

tagged



Formats/encoding

- format choices at these levels:
 - representation of information
 - representation of characters
 - how characters are assembled into files (file formats)



Characters

- use UTF-8 (aka Unicode ISO 10646)
- be aware of using characters outside ASCII (common US keyboard characters) – these can break if UTF-8 is not used
- distinguish character encoding and fonts (a font is simply a set of images for a “character set”)
 - something may be coded perfectly in UTF-8 but there is no suitable font applied
 - some fonts may display special



Useful tools for character encoding

- **Notepad++ (download via SourceForge)**
`http://notepad-plus-plus.org`
- **Fileformat website**
`http://www.fileformat.info`
- **SIL View-Glyph**
- **web browsers (they are UTF-friendly)**



File formats

- audio
 - WAV
 - (what if original is not WAV??)
 - resolution: 16 bit, 44.1KHz, stereo or better
- video
 - changing frequently
 - MPEG2 or MTS/H264/AVCH
 - resolution: depends on ...
 - get advice and check with your archive!



File formats

- images
 - TIFF ****OR**** original from device
 - resolution: archive quality is 300dpi or better



File formats

- text
 - best is plain text
 - PDF/A often acceptable, may pose problem
 - if MS-Word or ODF, check with archive
- structured data (spreadsheets, databases)
 - original format should be supplied
 - provide a preservable derivative as well (eg csv, PDF)
- common linguistic software (ELAN, Transcriber, Toolbox, Praat etc)
 - their file formats are generally



Standards

- we have already mentioned some standards – UTF-8, WAV etc
- there are other relevant standards, eg
 - ISO 639-3
 - metadata systems
- you can also establish project-local standards, eg
 - to handle special characters (eg \e = schwa)
 - data field names
 - document them! – for your usage and for correspondence to wider standards



Express yourself - Metadata

- metadata is *data about data*
 - for *identification, management, retrieval* of data
 - provides the *context* and *understanding* of that data
- carries those understandings into the *future*, and to *others*



Express yourself - Metadata

- metadata reflects the *knowledge* and *practices* of data providers
- ... and therefore *defines* and *constrains* audiences and usages for the data
- *all* value-adding to recordings of events (annotations transcriptions, translations, glosses, comments, interpretations, part of speech tagging etc) are actually metadata



Express yourself - Metadata

- you need to choose
 - a set of metadata categories applying across whole collection
 - additional metadata where possible
 - ways of expressing and encoding all that metadata



Common metadata standards

- OLAC: Open Language Archives

Title	Date
Community:	Description
Identifier	Format
Creator	Type
Contributor	Rights
Language	Coverage
Subject.language	Relation

- IMDI: ISLE Metadata Initiative (IMDI)
more categories, software specific

- ELAR: for endangered language
documentation, metadata framework is to



Types of metadata

- *people* metadata – creator's / delegate's details
- *descriptive* metadata – content of data
- *administrative* metadata – eg. date of last edit, relation to other data
- *preservation* metadata – character encoding, file format
- *access and usage* protocols



Examples

- [example](#) - XLS
- [example](#) - XML
- [example](#) – key
- [example](#) – key XML
- [example](#) – summary and requests
- [example](#) - notes



Meta-documentation

- Nathan (2010): “think of metadata as meta-documentation, the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced. Such meta-documentation should be as rich and appropriate as the documentary materials themselves.”



Meta-documentation

- identity of *stakeholders* involved, and their roles
- *attitudes* of language consultants, towards their languages and towards the documenter and documentation project
- *relationships* with consultants and community (Good 2010 mentions what he called ‘the 4 Cs’: ‘contact, consent, compensation, culture’);
- *goals and methodology* of researcher, including research methods and tools, some of the orientations (Woolley 2011)



Meta-documentation

- *project and researcher biography*: knowledge and experience of the researcher and consultants (eg. researcher's knowledge at beginning of project, what training researcher and consultants received)
- for funded projects: grant application, reports, email communications
- *agreements* entered into – formal or informal (eg. Memorandum of Understanding, compensation arrangements) and *promises* made to



Archiving with ELAR

The Endangered Languages Archive

Preserving and publishing documentation of endangered languages

[browse collections](#)

[search collections](#)

[get an account](#)

[advice for depositors](#)

[technical resources](#)

The Endangered Languages Archive (ELAR) is a digital repository for documentation of endangered languages. We aim to:

- provide a safe long-term repository for language documentation collections
- enable access to documentation collections according to the wishes of documenters and originating speakers and communities
- encourage co-operation between documenters and users of their collections
- provide advice and collaboration



A programme of the Hans Rausing Endangered Languages Project



School of Oriental and African Studies, University of London

ARCADIA

Funded by Arcadia

[ELAR access protocol](#)
[About ELAR accounts](#)

[Hans Rausing Endangered Languages Project \(HRELP\)](#)
[Online Resources for Endangered Languages \(OREL\)](#)

[Forum](#)
[ELAR staff](#)
[Contact us](#)



The Endangered Languages Archive at SOAS, London

Browsable collections

Africa
Asia
Australia
Europe
Middle East
North America
Pacific
South America

Click a region above to view browsable collections. To see all collections, click ['Search'](#) below.

Find a collection

Search
Map
List

Support

Help
Forums

Information at the [HRELP website](#):

ELAR @ HRELP
The Raising Room
Equipment info & reviews
Info for Depositors
Deposit forms

Popular collections

Choguita Rarámuri description and documentation

Gabriela Caballero

... audio and video recordings with transcriptions and photos of speakers from Chihuahua, Mexico. Includes myth and historical narratives, oratory, interviews, conversations, activities, ritual song and prayer, language teaching.

U R C S



Pite Saami: documenting the language and culture

Joshua Karl Wilbur

... audio and video recordings of Pite Saami speakers, northern Sweden, with transcriptions and translations in Swedish and English. Topics include reindeer roundup and traditional handicrafts.

U R C S



Ayutla Mixe documentation data

Rodrigo Romero Méndez

... around 150 audio and video recordings of speakers from Oaxaca, Mexico, with transcriptions. Includes texts of legends/folk tales and local histories, folk definitions and elicitation.

U R C S





Featured collection

We recently released *Vanishing Voices of the Great Andamanese*, a collection of audio and video recordings, texts, books, articles and images. As well as linguistic description and analysis, there are narratives, folktales, conversations, discussions, songs, ancestral knowledge, language learning resources, photos, and paintings by children. Together, these provide a linguistic and social ethnography of the Great Andamanese people.

We have provided an accessible navigational aid to the collection, signaling our new directions for collection presentation.



Access protocol

Throughout the site you see ELAR's access protocol graphics, like this: . A green box like this  means that you can access data.

Only registered users can access data; conditions also apply to each collection or files within it ... [read more](#)



User login

Username or e-mail:

Enter your username or email address.

Password:

Enter your ELAR password. ([Forgot it?](#))

[Create new account](#)

[Request new password](#)

About ELAR accounts

Why apply for an account?

If you don't have an account, you can read about the deposits here but you will not be able to access their resources. By registering, you will be able to access resources, apply for special access to restricted items, and create your own bookmarked collection of deposits.


To apply for an account, you must provide your real name. We do not allow anonymous accounts or fictitious identities. Accounts for ordinary User access are automatically granted on application.

For further information, see [Getting an account with ELAR](#) and for details of our account types and access system, see [ELAR's access protocol](#).



ELAR's users

- currently about 700 registered users
- users include anthropologists, archivists, artists, ethnographers, ethnomusicologists, filmmakers, folklorists, historians, journalists, language activists, language community members, language speakers, language teachers, librarians, linguists, poets, students, and “generally interested”
- ... from over 60 countries
- registrations from endangered language-speaking community members running at about 15%



ELAR's holdings

- currently online
 - 100 collections
 - 32,000 'bundles'
 - 60,000 files
 - about 55% are 'open'
 - 4 TB



ELAR holdings

- data types:
 - media files (sound, video) 19,050
 - graphics files (images, scans) 1,857
 - text files (fieldnotes, grammars, description, analysis) 3,407
 - structured data files (aligned and annotated transcriptions, databases, lexica) 1,893
 - metadata (structured, standardised contextual information about the materials)



Search

Found 17670 resources in the archive (page 1 of 2209)

1 2 3 4 5 6 7 8 9 ... [next >](#) [last >>](#)**Anvita Abbi***Jawaharlal Nehru University*

Associated deposits:

Great Andamanese Dictionary
Vanishing Voices of the Great Andamanese**Great Andamanese Dictionary***Anvita Abbi*

... video and audio recordings of songs, narrations, and sentences; interlinearized data; transcription and translation in English and Hindi; pictures with captions.

**Vanishing Voices of the Great Andamanese***Anvita Abbi*

... Great Andamanese is a highly endangered language of the Andaman Islands, south-east of India in the Bay of Bengal.



URCS

Primary data of Daohua*Yeshe Vodgsal Acuo*

... audio and video recordings, photos, lexical entries, texts, software dictionary, and research report for Daohua which has about 2,600 speakers in eastern Tibet.

**Primary data of Wutunhua***Yeshe Vodgsal Acuo***How to use search**

You can search in two ways:

- enter text in the search box and press 'Search'. Search is not case sensitive, and variations of words are found, e.g. 'Village' finds 'villages' and 'Indian' finds 'India'; or
- click a keyword in the left panel to find a set of resources. Click another keyword to refine the results (a black keyword) or to find a new set (a brown keyword)

To refine your search:

- enter two or more words for results containing all those words; e.g. entering 'nigeria' and 'audio' finds the deposit *Damakawa wordlist* which includes recordings made in northern Nigeria.
- use the keywords in the left panel to browse and select further categories; e.g. if you search for 'nigeria' and 'audio', a list (under 'Tags') includes place and language names: Akoko, Ikaann, Damakawa and Sakaba. Click one to find a resource pertaining to that name

To reset search and display all keywords, press 'Reset keywords'.

Colour coding of results

Search results can include deposits, bundles (file groups within deposits) and people. These are colour coded:

- A deposit: click on its title to view the deposit
- ▶ A bundle (group of files within a deposit): click on the icon or title to see the files and more metadata
- A person (usually, a depositor): click on the name to see more details, or click on one of the associated deposits.

Search ELAR Search[Reset keywords](#)**Access protocol**

URCS (546)
URCS (6281)
URCS (9)
URCS (212)
URCS (426)
URCS (149)
URCS (9733)
URCS (127)

Country [more](#)

Argentina (2)
Australia (21)
Bolivia (3)
Brazil (2)
Brunei (1)
[more ...](#)

Language [more](#)

Adelaide dialect (175)
Aiton (2)
Arapaho (128)
Assamese (2)
Auga (5)
[more ...](#)

Type [more](#)

Audio (13927)
Deposit (136)
Document (1721)
ELAN (1087)
Image (804)
[more ...](#)

Tags [more](#)

Show deposits: all full partial received proposed

Find a deposit:
[List](#)
[Map](#)

[Help](#)
[Forums](#)
[Home](#)



POWERED BY Google

Map data ©2012 MapLink, Tele Atlas - [Terms of Use](#)

Singpho Language of North East India (including Turung)

Home Resources Comments

Search this deposit

[Reset keywords](#)

Access protocol

- URCS (1206)
- URCS (2)
- URCS (21)

Language [more](#)

- Aiton (2)
- Assamese (2)
- English (50)
- Singpho (Diyun Hkawng) (17)
- Singpho (Diyun) (8)
- [more...](#)

Type

- Audio (1207)
- Document (175)

Genre

- Failed Recording (7)
- Lexicon (5)
- Primary Data (34)
- Primary Text (1158)
- Song (19)

Participants [more](#)

- A Seng (7)
- Aboni Kanta Shyam (8)
- Aboni (10)
- Adhon (6)
- Ai Mya Seng (27)
- [more...](#)

Singpho Language of North East India (including Turung)

Language: Singpho, Turung [sgp, try]

Depositor: Stephen Morey

Location: India



Summary of deposit

This deposit consists of audio recordings of speakers of Singpho, spoken in Arunachal Pradesh and Assam, India, resulting from fieldwork conducted between 2001 and 2006. The recordings include Turung, a variety of Singpho spoken in the Golaghat, Karbi Anglong and Jorhat Districts of Assam.

Deposit contents

The deposit comprises over 1200 audio files, as well as transcriptions and English translations of some of the recordings.

The recordings consists of traditional stories, some of stories of the Buddha and his previous lives; traditional songs, such as rice pounding songs and sagas; procedural texts, such as how to make certain types of rice cake or traditional customs relating to birth, marriage and death; cultural information, such as the preparation and usage of opium in former times; and a number of grammatical texts, elicitations and discussions about grammaticality.

Many of these texts are accompanied by Word documents and hmtl web page files that contain transcriptions.

The transcribed Singpho and Turung texts can be searched at the Tai and Tibeto-Burman languages of Assam website, maintained by the Centre for Research in Computational Linguistics, <http://sealang.net/assam>.

Acknowledgement

This recording was made by Dr. Stephen Morey in Assam. It is part of the heritage of the Singpho or

Depositor

Stephen Morey



Nationality: Australian
Affiliation: Research Centre for Linguistic Typology, La Trobe University

Access

Default access protocol: [URCS](#)
Your access roles: [URCS](#)

Deposit

Group represented: Singpho, Turung
Location: Karbi Anglong, Golaghat and Jorhat Districts in Assam (Turung) and Tinsukia District in Assam and Changlang and Lohit Districts of Arunachal Pradesh (Singpho - Numhpuk Hkawng, Tieng Hkwang and Diyun Hkawng)



- Topic**
- Talk (62)
- Chatting (9)
- Sickness (9)
- Bird Story (7)
- Where Wild Things Are (5)
- Coconut Oil (3)
- Cardinal (2)
- Devilish Pig (2)
- Flying Fox and Parrot (2)
- Laplap (2)
- Linguo-labials (2)
- Prawn (2)
- Swadesh (2)
- Turtle and Shark (2)
- Ais Island (1)
- Akalao Bird and Daughter (1)
- Akalao and Mother (1)
- Aore Island (1)
- Before Going to War (1)
- Circumcision (1)
- Conch and Sea Snail (1)
- Directions (1)
- Dying (1)
- Engagement (1)
- First Coconut (1)
- Five Fingers (1)
- Numbers (1)
- Pig Attack (1)
- Pig-killing Ceremony (1)
- Piria (1)
- Pledge (1)
- Plover and Red-head Bird (1)
- Rat, Short-leg and Octopus (1)
- Six Sisters (1)
- Surae (1)
- Troll (1)
- Turtle and Old Man (1)
- Tutuba Wild Man (1)
- Two Wild Men (1)
- Wedding (1)
- White Heron (1)
- Wild Apple (1)

007mavea

Not logged in. [Login](#) | [New user](#) | [Feedback](#) | [ELAR catalogue](#)

Documentation of Mavea

[Home](#) | [Metadata](#) | [Discussion](#) | [Resources](#)

Contributor

Valerie Guerin (162)
Gabriel Torno (2)

Participants

Sera Lima Lowet (88)
Allan Natu Lowet (18)
Elsie Fopua Kaman (15)
Mosela Vomei Kaman (10)
Valerie Guerin (5)
Fred Kaman (4)
Jo Tavon Livo (4)
PFL (4)
Paul Sope Livo (4)
James Sesei Livo (3)
Lowet Daldal Morris (3)
Pupu Moldovo Morris (3)
Rogen Molavea Lowet (3)
Gabriel Torno (2)
John Molsi Livo (2)
Judy Vokarae Livo (2)
Morris Tov'aoi Kaman (2)
Peter Vuropaitia Lowet (2)
Alfred Moltas Kaman (1)
Johnatan Nono Livo (1)
Jona Parparu Morris (1)
Lina Vatari Simptia (1)
Rolin Vofti (1)

Depositor

Valérie Guérin


Nationality: French
Affiliation: University of Hawai'i
Manoa

Your access

Default access protocol: **U R C S**
Your access roles: **U R C S**

Deposit

Group represented: Mavea speaking community, located on Mavea Island, and Deproma, Espiritu Santo Island, Vanuatu.
Location: Recordings were all created in Vanuatu. Locations include: Vunopuma, Saoroi, and Vunopua (on Mavea Island), Deproma (Espiritu Santo Island), Port Vila (Efate Island), Aore and Tanna Islands.



...ne 2005 and December

...riel Torno, entitled 'The

...at. The deposit


...ing this story: to listen,

...ch come with time-

...ories.

...resented in the

[Map](#) | [Terrain](#)





Access Protocol - URCS

- "full-size" URCS in black and white which show the status of either a deposit OR a user

U R C S

- all Users can access.

U R C S

- Researchers and Community members are allowed access

U R C S

- only Community members are allowed access (normally requires application to Depositor)

U R C S

- only Subscribers are allowed access (requires application to Depositor)

U R C S

- only the Depositor and delegate can access



URCS enhanced

- coloured URCS which display an overlay of deposit PLUS current user status, for example: "mini-urcs" URCS

URCS no account or not logged in

URCS User account, resource available to Users

URCS User account, resource available to Subscribers

URCS Community member account (for this deposit), resource available to Researchers and Community members

URCS Depositor (for this deposit), resource available to Subscribers (but Depositor can access all of their own deposit)

Singpho Language of North East India (including Turung)

Home Resources Comments

Search this deposit

 Search

Reset keywords

Access protocol

URCS x
 URCS (1)
 URCS

Language more

- Turung (Tibeto-Burman) (1)
- Turung (1)
- Aiton
- Assamese
- English
- [more...](#)

Type

- Audio (2)
- Document (1)

Genre

- Primary Text (2)
- Failed Recording
- Lexicon
- Primary Data
- Song

Participants more

- Ai Mya Seng (1)
- Doga (1)
- Kon Kham (1)
- Munindra (1)
- Palash Nath (1)
- [more...](#)

Singpho Language of North East India (including Turung)

Language: Singpho, Turung [sgp, try]

Depositor: Stephen Morey

Location: India



Summary of deposit

This deposit consists of audio recordings of speakers of Singpho, spoken in Arunachal Pradesh and Assam, India, resulting from fieldwork conducted between 2001 and 2006. The recordings include Turung, a variety of Singpho spoken in the Golaghat, Karbi Anglong and Jorhat Districts of Assam.

Deposit contents

The deposit comprises over 1200 audio files, as well as transcriptions and English translations of some of the recordings.

The recordings consists of traditional stories, some of stories of the Buddha and his previous lives; traditional songs, such as rice pounding songs and sagas; procedural texts, such as how to make certain types of rice cake or traditional customs relating to birth, marriage and death; cultural information, such as the preparation and usage of opium in former times; and a number of grammatical texts, elicitations and discussions about grammaticality.

Many of these texts are accompanied by Word documents and hmtl web page files that contain transcriptions.

The transcribed Singpho and Turung texts can be searched at the Tai and Tibeto-Burman languages of Assam website, maintained by the Centre for Research in Computational Linguistics, <http://sealang.net/assam>.

Acknowledgement

This recording was made by Dr. Stephen Morey in Assam. It is part of the heritage of the Singpho or

Depositor

Stephen Morey



Nationality: Australian
Affiliation: Research Centre for Linguistic Typology, La Trobe University

Access

Default access protocol: URCS
Your access roles: URCS

Deposit

Group represented: Singpho, Turung
Location: Karbi Anglong, Golaghat and Jorhat Districts in Assam (Turung) and Tinsukia District in Assam and Changlang and Lohit Districts of Arunachal Pradesh (Singpho - Numhpuk Hkawng, Tieng Hkwang and Diyun Hkawng)



ELAR Singpho Language of North East India ELAR

elar.soas.ac.uk/deposit/morey2007turungsingpho#q=%3Ffilters%3Dtype%253Abundle%2520im_og_gid%253A14%2520sm_eas%253A3

Singpho Language of North East India (including Turung)

Home Resources Comments

Found 2 bundles in this deposit with keyword URCS x (page 1 of 1)

Search this deposit

[Reset keywords](#)

Access protocol

URCS x

URCS (1)

URCS

Language *more*

Turung (Tibeto-Burman) (1)

Turung (1)

Aiton

Assamese

English

more...

Type

Audio (2)

Document (1)

Genre

Primary Text (2)

Failed Recording

Lexicon

Primary Data

Song

Participants *more*

Ai Mya Seng (1)

Doga (1)

Kon Kham (1)

Munindra (1)

Palash Nath (1)

more...

▼ SDM07-2006-122

SDM07-2006-122.wav Access protocol: URCS

This file is not available to all users. In order to be considered for access, you must apply to the depositor for individual access rights. You will receive an email with the outcome of your request. Note that even if your request is approved, you still may not be allowed to view this particular file. [Apply for access](#)

ID:	SDM07-2006-122
Title:	SDM07-2006-122
Date created:	27/1/06
Location:	Rengmai
Devices:	M-Audio Microtrack 24/96 - Sony ECM ZS90
Description:	Differences between the language in different villages

Keywords: Turung (Tibeto-Burman) - Primary Text - Kon Kham - Munindra - Palash Nath

► SDM07-20050703-00Z

Primary Text


Recorded on 2/2/05

Basapathar history

Keywords: Turung - Primary Text - Ai Mya Seng - Doga - Sukhen

Depositor

Stephen Morey



Nationality: Australian
Affiliation: Research Centre for Linguistic Typology, La Trobe University

Access


Default access protocol: URCS

Your access roles: URCS

Deposit

Group represented: Singpho, Turung

Location: Karbi Anglong, Golaghat and Jorhat Districts in Assam (Turung) and Tinsukia District in Assam and Changlang and Lohit Districts of Arunachal Pradesh (Singpho - Numhpuk Hkawng, Tieng Hkwang and Diyun Hkawng)



Singpho Language of North East India (including Turung)

Home Resources Comments

Search this deposit

 Search

Reset keywords

Access protocol

- URCS x
- URCS (1)
- URCS

Language

- Turung (Tibeto-Burman) (1)
- Turung (1)
- Aiton
- Assamese
- English
- more...

Type

- Audio (2)
- Document (1)

Genre

- Primary Text (2)
- Failed Recording
- Lexicon
- Primary Data
- Song

Participants

- Ai Mya Seng (1)
- Doga (1)
- Kon Kham (1)
- Munindra (1)
- Palash Nath (1)
- more...

Found 2 bundles in this deposit with keyword **URCS** (page 1 of 1)

SDM07-2006-122

SDM07-2006-122.wav

Access protocol: **URCS**

This file is not available to all users. In order to be considered for access, you must apply to the depositor for individual access rights. You will receive an email with the outcome of your request. Note that even if your request is approved, you still may not be allowed to view this particular file.

Depositor

Stephen Morey



Nationality: Australian
Affiliation: Research Centre for Linguistic Typology, La Trobe University

Apply for access

Concerning Edward Garrett's application

Tell the depositor who you are and why you would like access to restricted material in the deposit.

I met Ai Mya Seng when I travelled with my family to India in 1987. I would like to hear his voice again and perhaps eventually to contact him and his community.

Send request

access protocol: **URCS**
access roles: **URCS**


represented: Singpho,
on: Karbi Anglong, Golaghat
rhat Districts in Assam
g) and Tinsukia District in
and Changlang and Lohit
s of Arunachal Pradesh
no - Numhpuk Hkawng,
tkwang and Diyun Hkawng)





Archiving process

- look at ELAR
- contact ELAR
- send samples, summary
- send resources in suitable form
 - preservable
 - negotiate problems for best outcomes



Archiving process – what to provide

- deposit form (online)
- at least some description or annotation for all media
- inventory/catalogue/metadata covering ALL files
- metadata should cover at least these minimal categories:

<i>Category</i>	<i>Definition</i>	<i>Example</i>
Filename	The name of the file with its extension.	ejosm001.wav
Path	The path to the file in the folder structure of your deposit.	c:\recordings\ejosm001\ejosm001.wav
Identifier	The name of the file without its extension or filetype number.	[2] ejosm001 [#]
Title	A descriptive title for the session.	The Old Man and the Sea [#]
Topic(s)	The topic/subject matter of the session.	old man sea fish ^{#†}
Genre(s)	The genre of the session.	narrative retelling ^{#†}
Participant(s)	People involved in the session (may include the speaker and/or the person who made the recording).	John Smith Jane Saunders ^{#†}
Language(s)	The language(s) used in the session.	English Spanish ^{#†}
Date	The date when the session happened.	2012-03-06 [#]
Location	The location where the session happened.	Euston Tap, Euston Road [#]
Description	A description of the content of the session.	John retells the story of <i>The Old Man and the Sea</i> by Ernest Hemingway. [#]
Access Rights	An indication of who can access the deposit: (U, R, C or S – see below)	U



Current mode

- “progressive deposits”, to deal with backlog of deposits; appear sooner and incrementally curated


deposited data



resources available online



published collection



Archiving process – working with ELAR

- answer questions and help modify if necessary
- provide information (text, images) for general introduction
- if access restricted, respond to requests
- manage protocol over time
- send updated and additional materials
- give us feedback, report problems



End of archiving session

- end of archiving slides !!



Mobilisation

- documentation should be useful for a variety of purposes, including language teaching/learning
- may involve recording, collecting, managing materials differently, different metadata etc
- involves multiple skills and is best done by a team
- exploit 80/20 rule
 - only 20% of the user's perception of value comes from 80% of the work
 - 80% of the user's perception of value



Karaim – from CD to YouTube

- Spoken Karaim [link](#)
- annual summer schools
 - games [link](#) – crossword
 - games [link](#) – memory
 - resources [link](#) – texts
- drama work [link](#) – performances
- subsequently – the kids have posted their own videos on YouTube



Other examples

- Gayarragi, winangali - adding value to linguistic materials [link](#)
- created in training contexts [link](#) [link](#)
- Wunderkammer mobile phone dictionaries [link](#)
- speech bubble player [link](#)
 - conversing in Pite Saami are Henning Rankvist (left) and Elsy Rankvist (right). Video and texts from an ELAR collection deposited by Joshua Wilbur. Speech bubble player created by Edward Garrett.



End

