
ENGHUM Conference 15 November 2017

Preserving and revitalising endangered languages and cultural heritage: the role of digital archives

Peter K. Austin

Department of Linguistics
SOAS, University of London

© Peter K. Austin 2017

Creative commons licence

Attribution-NonCommercial-NoDerivs

CC BY-NC-ND

Thanks and absolution to: Lise Dobrin, David Nathan,
Julia Sallabank, Nick Thieberger

Overview

- Some terminology and definitions
 - Who uses archives?
 - What do they use them for?
 - Do the data in and interfaces to digital archives support efforts to revitalise languages?
 - Changing models
 - Click bait?
 - Discussion
-

Terminology

- Language documentation
 - Revitalisation
 - Archiving
-

Language documentation

- “concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties” (Himmelfmann 1998)
 - Features:
 - *Focus on primary data*
 - *Accountability*
 - ***Long-term storage and preservation of primary data***
 - *Interdisciplinary teams*
 - *Cooperation with and direct involvement of the speech community*
 - *Narrow view*: outcome is **annotated and translated corpus** of archived representative materials on use of a language, cf. DoBeS/TLA, ELAR – separate from **description** (language as system)
 - *Broad view*: outcome is transparent records of a language with description and theorisation dependent (Woodbury)
-

Henke & Berez-Kroeker (2016:411)

“It is difficult to imagine a contemporary practice of language documentation that does not consider among its top priorities the digital preservation of endangered language materials. Nearly all handbooks on documentation contain chapters on it; conferences hold panels on it; funding agencies provide money for it; and even this special issue evinces the central role of archiving in endangered language work. In fact, archiving language data now stands as a regular and normal part of the field linguistics workflow (e.g., Thieberger & Berez 2011).”

Language revitalisation

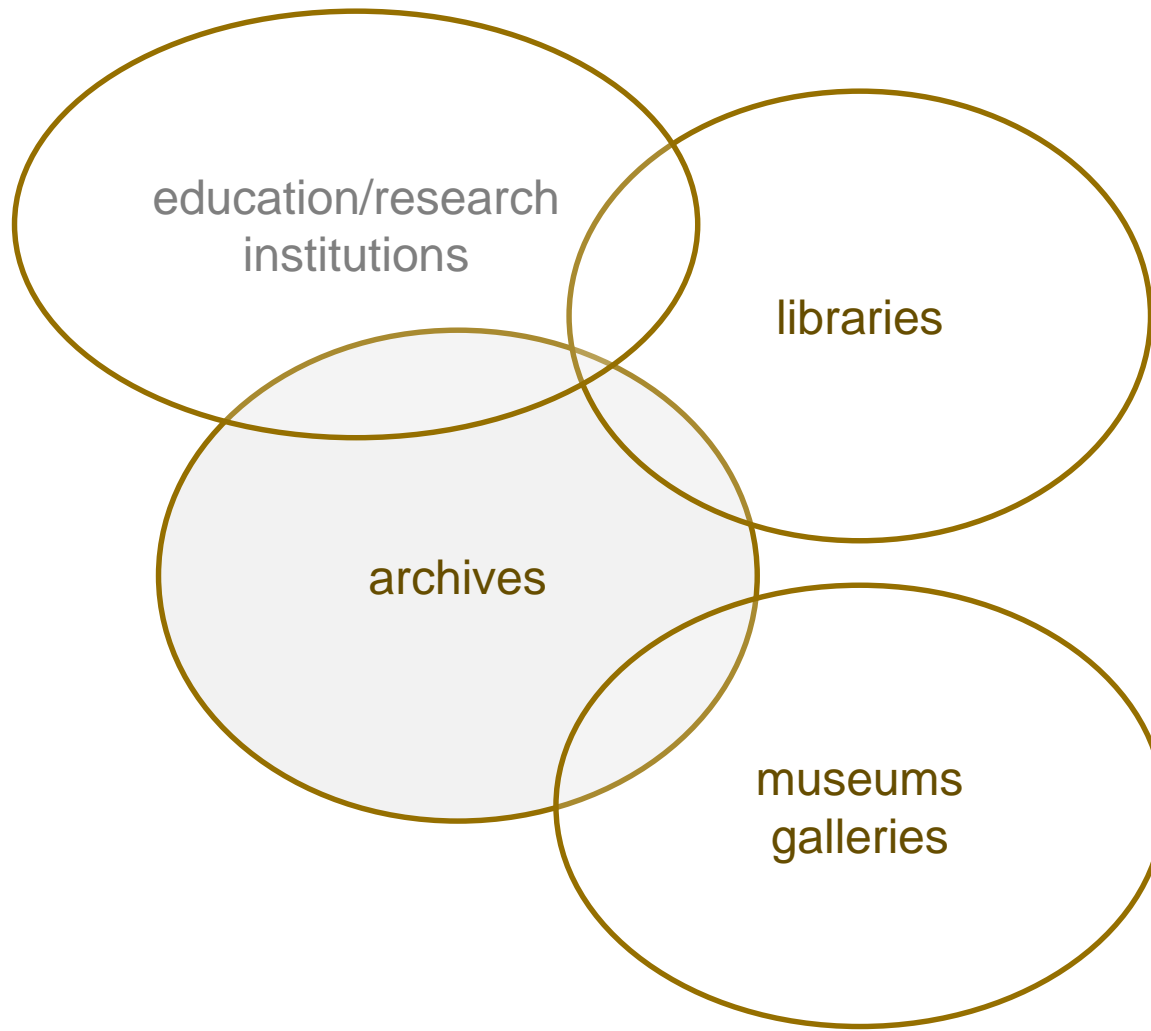
- efforts to increase **language vitality** by taking action to:
 - increase the domains of use of a language and/or
 - increase the number of speakers (often in the context of reversing language shift) both adults and children
 - older than language documentation (serious work began in 1970s and 1980s among Maori, Native American groups and others)
 - Speech/language community members are often more interested in revitalisation than documentation
 - Often assumed revitalisation = formal language learning (school lessons, immersion)
-

Archiving

- Trusted repository with a collection policy and a commitment to:
 - ❑ appraise the value of certain materials
 - ❑ preserve selected items
 - ❑ make known their existence
 - ❑ enable access to them (or their 'content')
 - Johnson (2004), Conathan (2011)
-

Where does archiving fit in?

“traditionally”



libraries,
archives,
museums and
galleries are
“memory
institutions”

Archiving skill inputs (Nathan 2016)

Sources

speakers/performers

authors

historical and “legacy” providers

Recordists

audio and video experts

data collectors/annotators/analysts

THE ARCHIVE

Curators

content/area
specialists
cataloguers

Data managers

data scientists

Technical practitioners

IT, media &
communications

Co-ordinators

managers
governance

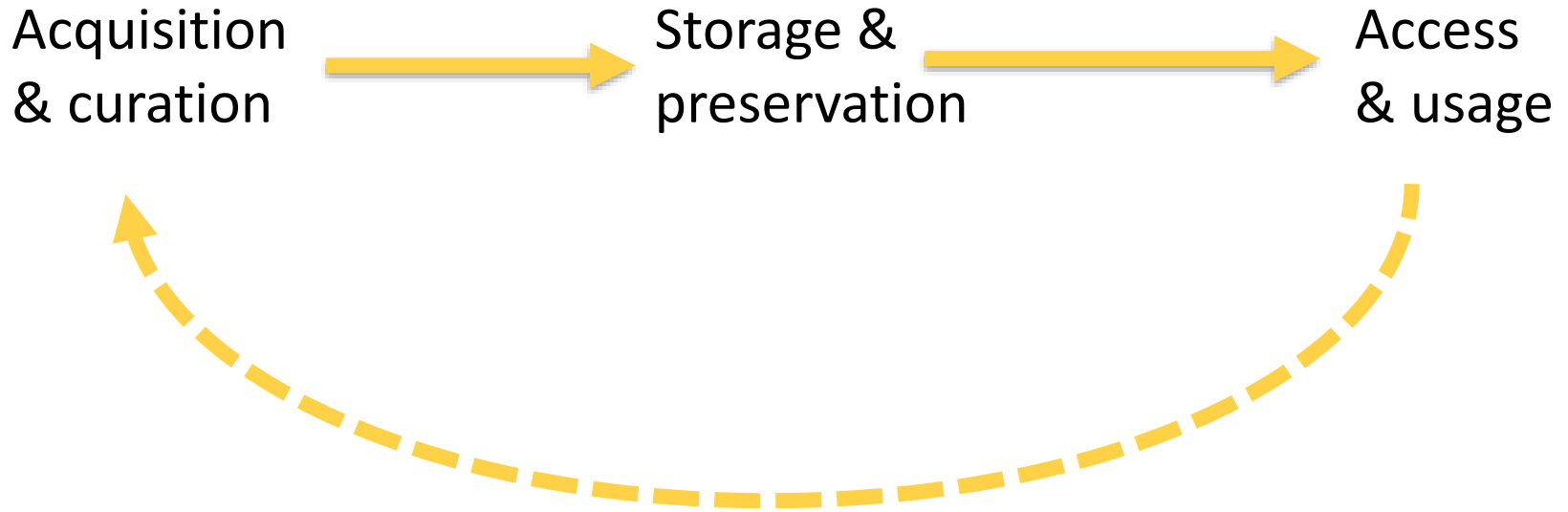
IT practitioners

programmers, installers

IT systems & software

cataloguing, storage,
preservation & access systems

Archiving and users (Nathan 2016)



The virtuous loop archives hope to achieve through serving their chosen community and through community participation

Table 2. Typology of language archives

Type	Type 1	Type 2	Type 3	Type 4
Descriptor	Large language archives with global collections	Large language archives with regional collections	Language archives embedded in larger digital repositories	Single language community archives
Contents	Global collection	Regional collection	Often regional but not always	Single language (or possibly 2 or 3 languages from a single community); may include cultural, historical, etc. materials
Primary Mission	Preservation/ documentation	Preservation/ documentation	Preservation/ documentation; may extend to language revitalization	Serving a language community; contributing to language revitalization
Source of Collections	Linguist depositors, typically tied to funded research projects	Linguist depositors	Linguist depositors	May come directly from linguist depositors; may be copies of collections in other archives
Intended Users	Broad base of users	Broad base of users	Broad base of users; parent repository may be required to serve certain constituents, e.g. university serves students, faculty	Language community; may also choose to be accessible to broad base of users
Funders	Private grants	Government or private grants or university support	Parent repository, most often a state institution	Community, maybe grants
Examples	ELAR DoBeS/Language Archive Cologne	PARADISEC Calif Lang Archive AILLA Sam Noble ANLA	U Oregon Libraries U Hawai'i Kaipuleohone	Dinji Zhuh K'ya Myaamia Center FirstVoices is conglomeration of Type 4s

Wasson et al.
(2016: 655)

Who uses digital archives, and for what?

Austin (2011a): “regionally-oriented archives like those in Alaska and California [Type 3] are essentially used by speaker communities or their descendants to access materials for cultural, historical or language-learning purposes. [For Type 1] The DOBES archive is primarily used by researchers, particularly its depositors. The ELAR archive has only been operating for a relatively short time, but most users seem to be depositors or other researchers.”

And today?

Paradisec

Booker (2017)

1st September 2017: 31TB of archived material, in 1,116 languages

“In the last 4 years PARADISEC has had 16,375 downloads, with 1058 registered users. Our catalog has had 11,000 sessions over the past 12 months. Significantly, these include 95 from PNG, 89 from Vanuatu, 33 from Fiji, 23 from the Solomon Islands, 20 from French Polynesia and 18 from New Caledonia.”

ELAR usage 2017-2017

Public statistics

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' at the centre.

Find a collection

[Browse](#)
[List](#)
[Map](#)

Public statistics

[Statistics](#)

See also

[ELAR at SOAS Library](#)
[Using ELAR](#)
[Depositing with ELAR](#)
[ELAR on Facebook](#)
[ELAR on Twitter](#)

Map



Public statistics

From:

31/10/2016

To:

31/10/2017

Filter

Data from 2016 October 31 to 2017 October 31

Page hits	1519948
User logins	2332
Clicks per visit	92
Average time per visit	0:19
Downloads of O resources	22384
Downloads of U resources	8714
Downloads of S resources	3208

Number of resources uploaded

O resources	5479
U resources	22195
S resources	6569

Who are the ELAR users?

Hits October 2016-2017

Russia 149,000

Germany 198,610

UK 381,179

USA 678,520

ELAR has no way to identify types of users but most likely the bulk are their own depositors and other researchers (especially Europe and Russia) doing theoretical or typological linguistics

Who are the users?

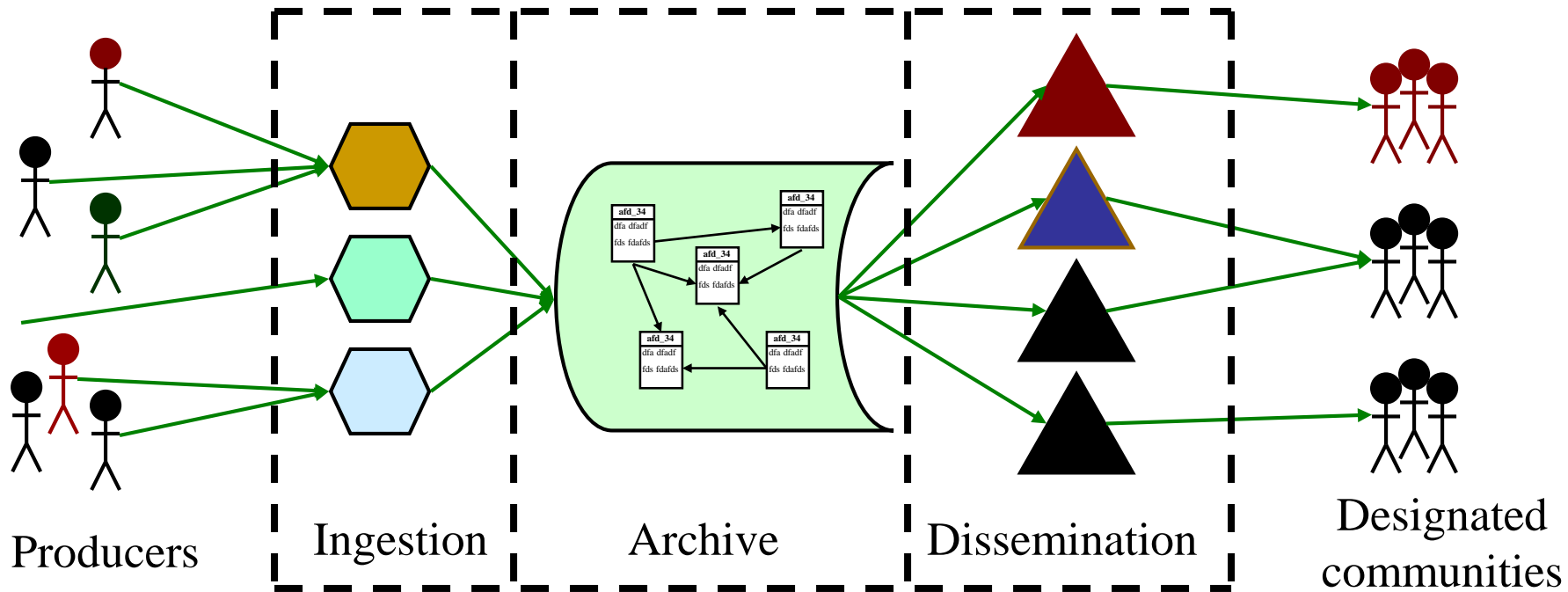
Wasson et al. (2016: 667) quoting Susan Kung, AILLA:

“As someone who runs an archive, the biggest issue I have, since it’s a digital archive, is knowing who’s using the archive. When somebody contacts me by email or phone...then I know who they are and what they’re trying to do, but otherwise I can track the downloads, I know people are logging in every day and downloading materials, but **I have no idea who they are or what they’re using these materials for**, what their agendas are, what they’re researching. So I feel like I’m just totally disconnected from most of my users, I have no insight into their needs or their wants.”

Wasson et al. (2016: 668) ‘[Felix] Rau noted that at the Language Archive Cologne, the majority of users are, in fact, depositors: “it’s bordering to the ridiculous sometimes how **the whole thing is focused on the producer side.**”

Traditional OAIS model

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



Changing models

Since its inception in 2005, ELAR aimed at a new Web 2.0 social model of archiving – Nathan (2010, 2011; see also Haske & Berez-Kroeker 2016)

“the archive is reconceived as a platform for conducting relationships between information providers (depositors) and information users” (Nathan 2010: 111).

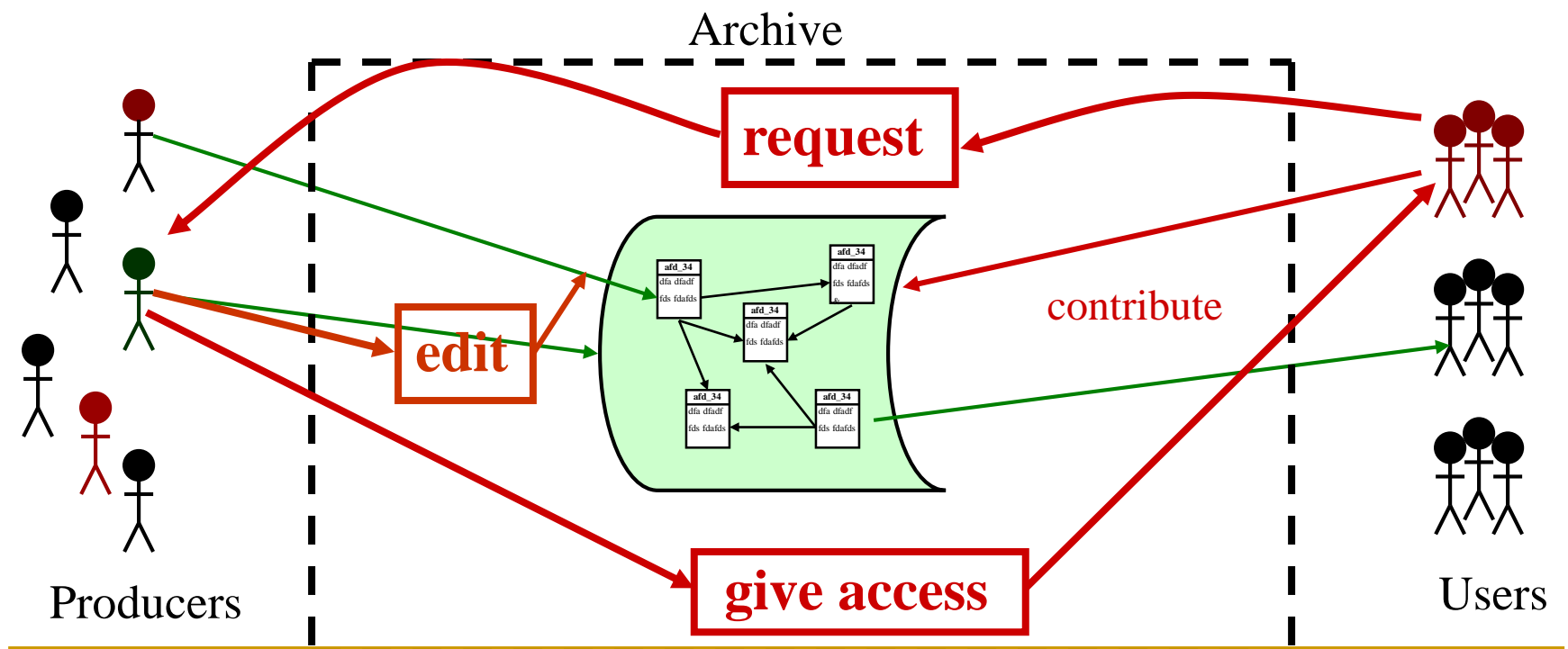
Introduction of managed access

U R C S

- U – resource available to *all* registered users
 - R – resource available to users registered as *researchers*
 - C – resource available to users endorsed as members of relevant *language community*
 - S – resource available to users who have been given *individual* access rights for that resource
-

ELAR - architecture

- reduced boundaries between depositors, users and archive:
 - users add, update content; negotiate access



Changing models

Haske & Berez-Kroeker (2016: 425): ‘This integration changes the nature of both access and distribution by allowing parties to negotiate directly with each other—rather than always going through an archivist/archive—which helps address problems such as accessing sensitive materials as well as managing the complexities of growing collections stewarded by small numbers of dedicated staff (Nathan 2010, 2011). This model, of course, shatters traditional boundaries of archiving: The digital archive is not just a place for preserving data; it has been reconceptualized as “a forum for conducting relationships between information providers (usually the depositors) and information users (language speakers, linguists and others)” (Nathan 2011:271)

Changing archive models

All this changed in 2015 when management of ELAR changed and the software system was jettisoned and replaced by the TLA (Nijmegen) system – no longer based on social networking but on file management

At the same time, a push for “open access” by the funder meant changes to the usage protocol system:

1. Removal of need to register as a user
 2. Introduction of “O” (open) category available to any user, registered or unregistered
 3. Removal of “C” (available to community members only) means elimination of one signal to identification of community member users
-

Archive models

Then, beginning in 2016 in an attempt to reduce curation (and staff) all depositors were required to use particular software tools for metadata management and to upload their collections themselves

In 2017 this became part of the funding requirements for the grants distributed by ELDP, whose outcomes had to be archived in the new ELAR

New archiving skill inputs

Sources

speakers/performers

authors

historical and “legacy” providers

Recordists

audio and video experts

data collectors/annotators/analysts

THE ARCHIVE

Curators

content/area
specialists
cataloguers

Data managers

data scientists

Technical practitioners

IT, media &
communications

Co-ordinators

managers
governance

Software Tools

IT practitioners

programmers, installers

IT systems & software

cataloguing, storage,
preservation & access systems

The baddies – recently in the news



Signals

- These are all essentially social media companies whose main business and products are **the users** – their identities (demographics, networks, follow(er)s), their actions ('like', vote, share, post, click, buy, view, download), their histories ('people who bought this also bought that', 'your friends also like')
 - User data and actions are collected and algorithmically processed to generate **signals**, that can then be used in search engines and advertising
 - This produces a rich picture of perceived needs, goals, choices, and uses of data and products
-

Archives and signals

- Most archives, including the Type 1 and Type 2 relatively well-funded ones like AILLA, TLA and Paradisec **lack** the infrastructure to collect signals
 - ELAR had the capability, but recent changes mean the **loss of signals** to identify users and user actions and interactions
 - But signals, exactly like those of Facebook, Google, Amazon and Apple could be of great assistance for making archives **more accessible and responsive** to users wanting to use them for revitalisation purposes – this could be a great space for **participant action research** and engaged humanities!
-

Archive content and interfaces

Wasson et al (2016: 669): 'In their presentations, the archivists provided a rich list of problems that might be encountered by users of language archives. The most frequently mentioned items were:

- A lack of contextual information at the deposit level, or metadata
- Incomplete materials—missing annotation, missing translations
- Inadequate search/browse functions
- Problems with the interface/information display
- Users may be frustrated when they don't have access to data; it may be hard for the archivist to get hold of a collection owner to request access for a user
- Technology issues—outdated, broken scripts, Flash/Java problems, etc.
- Interface language(s) may not [be] ... spoken by would-be users'

For language revitalisation purposes

I looked at the collections in ELAR, TLA, AILLA and Paradisec and identified the following additional issues for users interested in revitalisation:

1. Materials are often in a transcription that does not match community orthographies, even when such versions could easily be produced in many cases
 2. Special software is needed to view/play materials
 3. Materials are not tagged for use in education or for level (cf. Nathan & Fang 2009)
 4. Search interfaces on the archives do not return useful results – see Austin 2011b on searching for “Educational materials”
-

For language revitalisation purposes

- 5. Content of materials is often inappropriate for teaching purposes, because of genre, taboo lexicon, etc. but not labelled as such in the metadata (Austin & Sallabank 2018)
 - 5. Materials are culturally and/or pragmatically inappropriate, e.g. recordings of 'Frog Stories' while traditional stories are missing!
-

ELAR search for “Frog Story”



Search

Found 196 resources in the archive (page 1 of 25)

1 2 3 4 5 6 7 8 9 ... [next >](#) [last >](#)

Documentation and Analysis of Kabardian as Spoken in Turkey

Ayla Applebaum Bozkurt

... documentation of Kabardian, a typologically rare and threatened Northwest Caucasian language as spoken by the Turkish Kabardian community.

Documentation of Rongga

I Wayan Arka

... Rongga has 4,000 speakers in the villages Tanarata, Bamo, Watunggene and Waelengga, on Flores Island, Indonesia. Data collected includes audio and visual recordings of interviews and observations and linguistic descriptions ...

The painter's eye, the painter's voice: language, art and landscape in the Gija world

Frances Kofod

The painter's eye, the painter's voice: language, art and landscape in the Gija world

Search ELAR

[Reset keywords](#)

Language

!Xo (1)
Adelaide dialect (10)
Avatime (Dominant) (9)
Avatime (1)
Brisbane dialect (10)
Cashibo-Cacataibo (1)
Chinese (1)
Choguita Rarámuri (4)
Ecuadorian Siona (1)
English (Dominant) (1)
English (9)
French (5)
Gija (6)
Guari-Guari (8)
Gurindji Kriol (10)
Hupa (1)
Ju|'hoan (1)
Kabardian (2)
Kibena (Kisovi) (2)
Kibena (Nga?veta) (1)
Kibena (8)
Koyi rai (1)
Kubokota (4)
Melbourne dialect (10)
Nalu (6)
Northern dialect (20)

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:

Corpus accessibility – I found it, what now?

Cicipu documentation


Home Resources

Found 60 bundles in this deposit with keyword **ELAN x** (page 1 of 8)
1 2 3 4 5 6 7 8 next > last >>

Discussion of chieftancy

svgd001.eaf Access protocol: **URICIS**
[Download](#)

svgd001.001.mpg Access protocol: **URICIS**



00:05 00:17

[Download](#)

Search this deposit

[Reset keywords](#)

Access protocol

URICIS (60)

Language

[more](#)

- Cicipu (58)
- Tidipo (5)
- Tikula (3)
- Tirisino (6)
- Damakawa

[more...](#)

Type

ELAN x

- Audio (60)
- Image (8)
- Transcriber (3)
- Video (10)
- Document
- Text
- XML
- Zipped collection

Tags

- Kezzeme (2)
- Photo
- Photos

Genre


[more](#)

Deposit status

✓ **Curated:**
Resources online and curated

Depositor

Stuart McGill




Nationality: UK
Affiliation: School of Oriental and African Studies

Your access

Your roles: **URICIS**

Tools

[Download metadata](#)
[Add to My Bookmarks](#)



Corpus accessibility – I can't even find it

The screenshot displays the interface of The Language Archive. On the left is a hierarchical tree view of corpora, including 'DoBeS archive' and 'Bainouk'. The main panel on the right shows the details for a session named 'DJI10312CDD'. The session title is 'Tree list' and the date is '2012-03-01'. The description is 'Verification of the pronunciation and agreement patterns of all tree names'. The content details include: Genre: Elicitation; SubGenre: lexical elicitation; Task: Tree list; Modalities: speech; Subject: Tree list; Interactivity: interactive; PlanningType: planned; Involvement: individual; SocialContext: individual; EventStructure: individual; Channel: individual. The languages are listed as 'Bainouk Gubeeher (c)' and 'French (c)'. The actors are 'Alexander Cobbinah' and 'Jean Marie Sagna'. The media file is an audio file in 'audio/x-wav' format, 460 MB in size, with unspecified quality. The recording conditions and time position are also listed as unspecified.

The Language Archive

about manual register user: anonymous Log in

METADATA SEARCH CONTENT SEARCH MANAGE ACCESS REQUEST ACCESS

CITATION DOWNLOAD ALL VERSION INFO

BOOKMARK

Session

Name DJI10312CDD
Title Tree list
Date 2012-03-01

Description

Verification of the pronunciation and agreement patterns of all tree names.

Location

Project DoBeS 3P

Content

Genre Elicitation
SubGenre lexical elicitation
Task Tree list
Modalities speech
Subject Tree list
Interactivity interactive
PlanningType planned
Involvement individual
SocialContext individual
EventStructure individual
Channel individual

Languages

Language Bainouk Gubeeher (c)
Language French (c)

Actors

Actor Alexander Cobbinah
Actor Jean Marie Sagna

MediaFile

Type audio
Format audio/x-wav
Size 460 MB
Quality Unspecified

RecordingConditions

TimePosition

Start Unspecified
End Unspecified

Conclusions – some lessons for ENGHUM?

1. There is very **little role**, if any, for Type 1 & 2 digital archives in their current form in supporting revitalisation of endangered languages and cultural heritage – possibly just for preservation
 2. Such archives **lack** the skills, expertise and interest in revitalisation or identifying and supporting such users
 3. Recent changes in ELAR in particular have **reversed** the relationship-based model (at exactly the time this is the core of successful big business models) which might have been useful for signals collection and interpretation, and selected and directed communication between users and depositors
 4. Elimination of archival staff and skills and reliance on software solutions puts pressure on depositors, reducing the already limited time they have to support revitalisation
-

Conclusions

5. Local archives, and local museums, are likely to be the **best candidates** for a role in revitalisation, but they are often hamstrung by lack of resources and skills (Wilbur 2014) and would need to partner with a larger organisation for preservation and infrastructure
 6. There is **great potential** for future research involving ideas around signals collection and processing (derived from social media companies) via participant action research investigating the potential roles of archives in language revitalisation and the support they could provide for minority and endangered languages
-

References

- Austin, Peter K. 2011a. Who uses digital language archives?
<http://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/>.
- Austin, Peter K. 2011b. Searching in Endangered Languages Archives.
<http://www.paradisec.org.au/blog/2011/05/searching-in-endangered-languages-archives/>
- Austin, Peter K. & Julia Sallabank. 2018. Language documentation and language revitalisation: some methodological considerations. In Leanne Hinton, Leena Huss & Gerard Roche (eds.) *Handbook of Language Revitalisation*, 207-215. London: Routledge.
- Conathan, Lisa. 2011. Archiving and language documentation. In Peter K. Austin & Julia Sallabank (eds.) *Cambridge Handbook of Endangered Languages*, 235-254. Cambridge: Cambridge University Press.
- Himmelman, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–95.
-

References

Himmelman, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelman & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 1–30. Berlin: Mouton de Gruyter.

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelman, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.) *Language Documentation & Conservation Special Publication No. 3, Potentials of Language Documentation: Methods, Analyses, and Utilization*, 111–117. Honolulu: University of Hawai'i Press.

<https://scholarspace.manoa.hawaii.edu/handle/10125/4523>.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language Documentation and Description Volume 2*, 140–153. London: SOAS.

References

- Nathan, David. 2010. Archives 2.0 for Endangered Languages: from Disk Space to MySpace. *International Journal of Humanities and Arts Computing* 4.1–2, 111–124.
- Nathan, David. 2011. Digital archiving. In Peter K. Austin & Julia Sallabank (eds.) *The Cambridge handbook of endangered languages*, 255–273. Cambridge: Cambridge University Press.
- Nathan, David. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. In David Nathan & Peter K. Austin (eds.) *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 187–208. London: SOAS.
- Nathan, David. 2016. Archiving. Lecture slides, DocLing Tokyo.
- Nathan, David & Meili Fang. 2009. Language documentation and pedagogy for endangered languages: a mutual revitalisation. In Peter K. Austin (ed.) *Language Documentation and Description*, vol 6, 132–160. London: SOAS.
- Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.
-

References

Wasson, Christina, Gary Holton & Heather S. Ross. 2016. Bringing User-Centered Design to the Field of Language Archives. *Language Documentation and Conservation* 10, 641-681.

Wilbur, Joshua. 2014. Archiving for the community: Engaging local archives in language documentation projects. In David Nathan & Peter K. Austin (eds.) *Language Documentation and Description*, vol 12, 85-102. London: SOAS.
[http://www.elpublishing.org/docs/1/12/lidd12_06.pdf]

Thank you!
