DOCUMENTARY LINGUISTICS I

prof. Nicole Nau, UAM winter 2017/2018

Third lecture

17 October 2017


NALEŚNIKARNIA
GRAMOFON
NALEŚNIKARNIA
NIKI パンケーキ CREPES БЛИНЫ PFANNKUCHEN CRESPELLE 煎饼

TOPICS OF THE DAY

- Review of some essential requirements of modern language documentations
- Metadata
- Language archives

QUESTIONS YOU SHOULD BE ABLE TO ANSWER BY NOW

- What is (a) language documentation (LD)?
- What kind of records (data) does a LD contain?
- Why is it not possible to record **all** communicative events in a given speech community?
- For whom and for which purposes are languages documented?
- Why should native speakers take an active part in documenting a language?
- Why is it important to store primary data in open archives?

PROCESSES OF LANGUAGE DOCUMENTATION IDENTIFIED BY AUSTIN (2006)

1. *recording* – of media (audio, video, image) and text;
2. *capture* – moving analogue materials to the digital domain;
3. *analysis* – transcription, translation, annotation, and notation of metadata;
4. *archiving* – creating archival objects, and assigning access and usage rights;
5. *mobilization* – publication, and distribution of the materials in various forms.

SOME THOUGHTS ABOUT ADEQUACY: WHAT IS A «GOOD» LANGUAGE DOCUMENTATION? (BERGE 2010)

Adequacy in documentation must address the needs of non-linguists, particularly the needs of the users of the language being documented, as well as the needs of linguists.

One measure of adequacy in documentation and description might be how learnable the language is as a result, since acquiring fluency in a language requires enough data with enough descriptions to reproduce the language outside its normal context.

A GOOD LD IS…

(AUSTIN 2006, CITING WOODBURY 2003)

1. *diverse* – containing samples of language use across a range of genres and socio-cultural contexts, including elicited data;

2. *large* – given the storage and manipulation capabilities of modern information and communications technology (ICT), a digital corpus can be extensive and incorporate both media and text;

3. *ongoing*, *distributed*, and *opportunistic* – data can be added to the corpus from whatever sources that are available and be expanded when new materials become available;

4. *transparent* – the corpus should be structured in such as way as to be useable by people other than the researcher(s) who compiled it, including future researchers;

5. *preservable* and *portable* – prepared in a way that enables it to be archived for long-term preservation and not restricted to use in particular ICT environments;

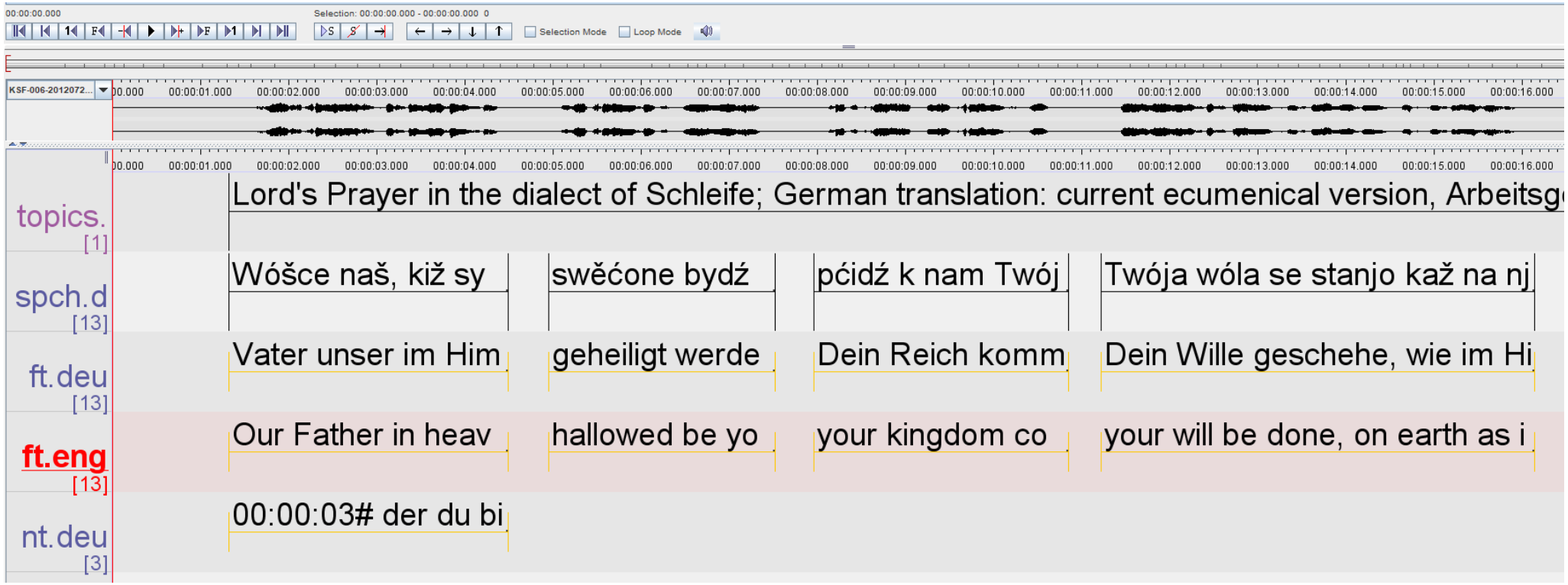
6. *ethical* – collected and analyzed with due attention to ethical principles (see Chapter 2) and recording all relevant protocols for access and use.

Primary data	Apparatus	
	Per session	For documentation as a whole
recordings/records of observable linguistic behavior and metalinguistic knowledge (possible basic formats: session and lexical database)	*Metadata* – time and location of recording – participants – recording team – recording equipment – content descriptors … *Annotations* – transcription – translation – further linguistic and ethnographic glossing and commentary	*Metadata* – location of documented community – project team(s) contributing to documentation – participants in documentation – acknowledgements … *General access resources* – introduction – orthographical conventions – ethnographic sketch – sketch grammar – glossing conventions – indices – links to other resources …

CONTENT OF A LANGUAGE DOCUMENTATION

(HIMMELMANN 2006)

THE LORD'S PRAYER IN A SORBIAN DIALECT, MODERN RECORDING WITH ANNOTATION (IN ELAN)



Where are the metadata? http://dobes.mpi.nl/

dobes.mpi.nl

FEATURED MEDIA

RESEARCH PORTAL | GENERAL INTEREST PORTAL | DEPOSIT YOUR DATA

WELCOME TO THE DOBES PORTAL

The DOBES Archive contains language documentation data from a great variety of languages from around the world that are in danger of becoming extinct. This portal gives access to the material in the archive and provides information about the DOBES endangered languages documentation programme.

North and Meso-America | South America | Eurasia | Africa | South East Asia and Oceania

This map displays all languages currently in the DOBES Archive. By clicking on a location you can go to the

Type your search term here

Search the DOBES archive

Show only results that are accessible to me

- Access and Registration
- Research Portal
- General Interest Portal
- Deposit your Data
- Documentation Projects
- Research Projects
- DOBES Programme
- Archive Information

BROWSE THE DOBES ARCHIVE

VolkswagenStiftung

http://dobes.mpi.nl/projects/

DOBES

DOCUMENTATION OF ENDANGERED LANGUAGES

LOWER SORBIAN

PROJECT STATISTICS

Sessions: 263
Audio recordings: 263
Video recordings: 1
Annotations: 263
Images: 0

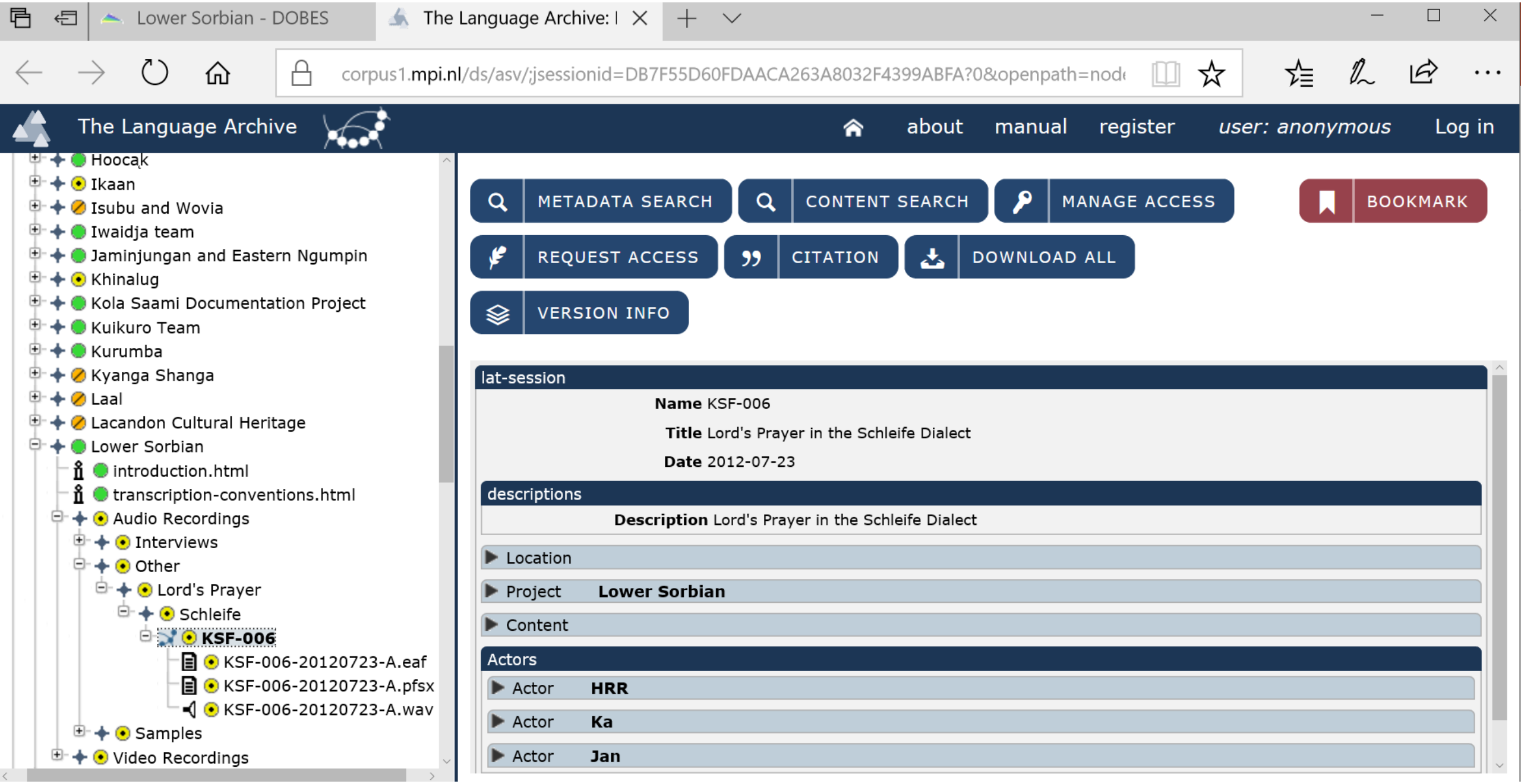
Last update: 2017-6-18

Welcome to the Lower Sorbian part of the DoBeS archive.

Lower Sorbian is a West Slavic language (see Language) spoken in Lower Lusatia, Germany (see People & Culture).

The presented project was carried out in 2010-2015 in the Sorbian Institute, Cottbus, Germany (see Team) with the aim to obtain and document speech performance of the native speakers of this endangered language.

As a result, more than 100 hours of interviews and other recordings have been made publicly available through The Language Archive. All recordings are followed by an orthographic transcription and a German translation; selected samples are also transcribed phonetically and translated into English.


Lower Sorbian - DOBES
The Language Archive: |
corpus1.mpi.nl/ds/asv/;jsessionid=DB7F55D60FDAACA263A8032F4399ABFA?0&openpath=nod
The Language Archive
about
manual
register
user: anonymous
Log in
Hoocąk
Ikaan
Isubu and Wovia
Iwaidja team
Jaminjungan and Eastern Ngumpin
Khinalug
Kola Saami Documentation Project
Kuikuro Team
Kurumba
Kyanga Shanga
Laal
Lacandon Cultural Heritage
Lower Sorbian
introduction.html
transcription-conventions.html
Audio Recordings
Interviews
Other
Lord's Prayer
Schleife
KSF-006
KSF-006-20120723-A.eaf
KSF-006-20120723-A.pfsx
KSF-006-20120723-A.wav
Samples
Video Recordings
METADATA SEARCH
CONTENT SEARCH
MANAGE ACCESS
BOOKMARK
REQUEST ACCESS
CITATION
DOWNLOAD ALL
VERSION INFO
lat-session
Name KSF-006
Title Lord's Prayer in the Schleife Dialect
Date 2012-07-23
descriptions
Description Lord's Prayer in the Schleife Dialect
Location
Project Lower Sorbian
Content
Actors
Actor HRR
Actor Ka
Actor Jan

TYPES OF METADATA WITH EXAMPLE

(AUSTIN 2006)

Table 1. Different types of metadata associated with a computer file

Cataloguing	Title: Sasak.dic; Collector: Peter K Austin; Speakers: Yon Mahyuni, Lalu Hasbollah; Language code: SAS
Descriptive	Trilingual Sasak-Indonesian-English dictionary, linked to finderlists, morpheme forms link to Sasak text collection
Structural	Dictionary entries with headword, part of speech, gloss in Bahasa Indonesia and English, cross-references for semantic relations; SIL FOSF record format
Technical	Shoebox 5.0 ASCII text file
Administrative	Open access to all; Last edited version dated 2004-06-25; backup 2004-06-20 on DVD 012

WHAT ARE METADATA FOR?

- For users of the archive: finding and selecting records

«Metadata, or catalogue information, is what makes discovery possible.» (Johnson 2004)

- For later generations: have maximal information about the record

«Metadata catalogue information is especially vital for digital materials, because they are not amenable to direct inspection, as is a book or other printed matter.» (Johnson 2004)

- For documentators: keep your collection in order!
- For archivers: structure the archive in a logical way

MINIMUM ACCORDING TO JOHNSON (2004)

- creators’ full names
- name of the language
- date of creation
- place of creation
- access restrictions
- genre keyword

BUT…

«Evidently not all the information that can be specified with the proposed set of metadata elements is always available. This is specifically the case for legacy resources or very specialised resources. Therefore only those elements should be mandatory that are needed for the correct functioning of tools working with the metadata descriptions. For the session metadata only the session name is needed to distinguish between other sessions in the same corpus or sub-corpus.»

(Metadata elements for session description, ver. 3.0.4 (2003); https://tla.mpi.nl/?attachment_id=4532)

SOME PROBLEMS IN ORGANIZING METADATA: EXPERIENCE FROM A PROJECT

Example: recordings for a multilingual corpus of Baltic and Slavic dialects (http://www.trimco.uni-mainz.de/trimco-dialectal-corpus/)

- Language of metadata
- Filenames
- Completeness, problem of adding data later
- Metadata relating to recordings, to speakers, to places …

METADATA: WHERE TO FIND ADVICE

«A gentle introduction to metadata» by Jeff Good (2002):

http://linguistics.berkeley.edu/~jcgood/bifocal/GentleMetadata.html

IMDI (ISLE Metadata Standard):

https://tla.mpi.nl/imdi-metadata/

OLAC Metadata Standard – explanations:

http://www.language-archives.org/NOTE/usage.html

LDC Filename conventions and metadata:

https://www.ldc.upenn.edu/data-management/providing/filenames-metadata

WHAT IS OLAC?

«OLAC, the **Open Language Archives Community**, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by:

(i) developing **consensus on best current practice** for the digital archiving of language resources, and

(ii) developing a **network of interoperating repositories** and services for housing and accessing such resources.»

http://www.language-archives.org/index.html

HOW TO USE THE OLAC CATALOGUE

The OLAC catalogue is a «catalogue of catalogues».

Information about archives containing records of or information about individual languages.

Rich metadata about the record.

Link to the archive were the record can be found.

Google search: OLAC language name, for example:

OLAC Guwamu

OLAC resources in and about the Guwamu language

ISO 639-3: gwu

The combined catalog of all OLAC participants contains the following resources that are relevant to this language:

- Primary texts
- Lexical resources
- Language descriptions
- Other resources about the language

Use faceted search to explore resources for Guwamu language.

Primary texts

1. ONLINE *Guwamu vocabulary and notes.* Gavan Breen (compiler). 1970. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). oai:paradisec.org.au:GB07-001
2. ONLINE *Guwamu vocabulary.* Gavan Breen (compiler). 1967. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). oai:paradisec.org.au:GB07-002
3. ONLINE *Guwamu vocabulary and elicitation sentences.* Gavan Breen (compiler). 1967. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). oai:paradisec.org.au:GB07-003
4. ONLINE *Guwamu, Goodooga, 1955.* Stephen (S.A.) Wurm (compiler); Stephen (S.A.) Wurm (recorder). 1955. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). oai:paradisec.org.au:SAW2-128

Lexical resources

1. ONLINE *Guwamu Swadesh List.* n.a. n.d. The Rosetta Project: A Long Now Foundation Library of Human Language. oai:rosettaproject.org:rosettaproject_gwu_swadesh-1

Language descriptions

1. ONLINE *Glottolog 3.0 Resources for Guwamu.* n.a. 2017. Max Planck Institute for the Science of Human History. oai:glottolog.org:guwa1243

Other resources about the language

1. ONLINE *Guwamu, a language of Australia.* n.a. 2017. SIL International. oai:ethnologue.com:gwu

OLAC Record

oai:paradisec.org.au:SAW2-128

Metadata

Title:	Guwamu, Goodooga, 1955
Access Rights:	Open (subject to agreeing to PDSC access conditions)
Bibliographic Citation:	Stephen (S.A.) Wurm (collector), Stephen (S.A.) Wurm (recorder), 1955; Guwamu, Goodooga, 1955, MPEG/X-WAV, 2017-09-29. I
Contributor (compiler):	Stephen (S.A.) Wurm
Contributor (recorder):	Stephen (S.A.) Wurm
Coverage (Box):	northlimit=-6.158; southlimit=-28.077; westlimit=145.894; eastlimit=149.083
Coverage (ISO3166):	AU PG
Date (W3CDTF):	1955-01-01
Date Created (W3CDTF):	1955-01-01
Description:	Tape on loan from AIATSIS: WURM_S03-002895A -- Guwamu, Goodooga, 1955 (Track B features some non-Australian material, Language as given: Guwamu, Oiana
Format:	Digitised: yes Media: Small Kodak R to R Tape Audio Notes: A). 3 3/4ips. Mould removed. Levels vary. Some accidental pausing/s variation. Some analogue clipping. B). 3 3/4ips. Mould removed. Levels vary. Some accidental pausing/stopping by recordist. Some clipping.
Identifier:	SAW2-128
Identifier (URI):	http://catalog.paradisec.org.au/repository/SAW2/128
Language:	Gadsup

PARADISEC Catalog

Item details

Item ID	SAW2-128 (Collection Details)
Title	Guwamu, Goodooga, 1955
Description	Tape on loan from AIATSIS: WURM_S03-002895A -- Guwamu, Goodooga, 1955 (Track B features some non-Australian material, Oiana recorded at Kainantu, PNG)
Origination date	1955-01-01
Origination date free form	1955, 1958
Archive link	http://catalog.paradisec.org.au/repository/SAW2/128
URL	
Collector	Stephen (S.A.) Wurm Find similar
Countries	Australia - AU Papua New Guinea - PG *To view related information on a country, click its name*
Language as given	Guwamu, Oiana

Content Files (4)

View file contents

Filename ▲▼	Type ▲▼	File size ▲▼	Duration ▲▼	File access
SAW2-128-A.mp3	audio/mpeg	10.7 MB	00:11:38.431	
SAW2-128-A.wav	audio/x-wav	385 MB	00:11:39.967	
SAW2-128-B.mp3	audio/mpeg	10.8 MB	00:11:46.405	
SAW2-128-B.wav	audio/x-wav	389 MB	00:11:47.982	
4 files	--	**795 MB**	--	--

Show 10 | Show 50 | Show all 4

Collection Information

Collection ID	SAW2
Collection title	Recordings from Solomon Islands, French Polynesia, Papua New Guinea and Australia
Description	A large collection of audio recordings made by Stephen Wurm in a number of countries on a variety of languages. Solomon Island data comprises the majority of the collection. Languages recorded include

INFORMATION ABOUT THE EXAMPLE (GUWAMU RECORDING BY STEPHAN WURM)

Austin, Peter K. 2013 Language documentation and meta-documentation. In Sarah Ogilvie and Mari Jones (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*. Cambridge: Cambridge University Press. [a preprint version can be found online]

Is this record a «Zombie voice»?

ZOMBIE VOICES

«As the last speaker utters her/his last words, the ‘expert’ is there to record this important moment and preserve it for all time. Among the benefits from such preservation efforts is the ability to play back the recordings at any time in any place. In popular media this process is described as ‘saving the language’ through recording and documentation. Unfortunately, these recordings are not living voices. Rather, they are zombie voices—undead voices that are disembodied and techno-mechanized. They are cursed with being neither dead nor alive.» (Perley 2012)

WHAT CAN YOU DO TO REVIVE «UNDEAD» VOICES?

Program «**Breath of Life**» of the University of Berkeley, California

- short video: https://www.youtube.com/watch?v=xquUir5mn28
- a bit longer (and broader topic): RECOMMENDED WATCHING

Archiving for Speakers and Linguists (Nijmegen, The Netherlands, and Berkeley, California) - third video on this site:

http://www.pbs.org/thelinguists/For-Educators/Video-Extras.html#Archiving

- talk by Leanne Hinton (about 45 minutes plus discussion):

https://www.youtube.com/watch?v=DDn2VhHjoNY

ARCHIVES BETWEEN DEPOSITORS AND USERS

TRILSBEEK & WITTENBURG (2006)

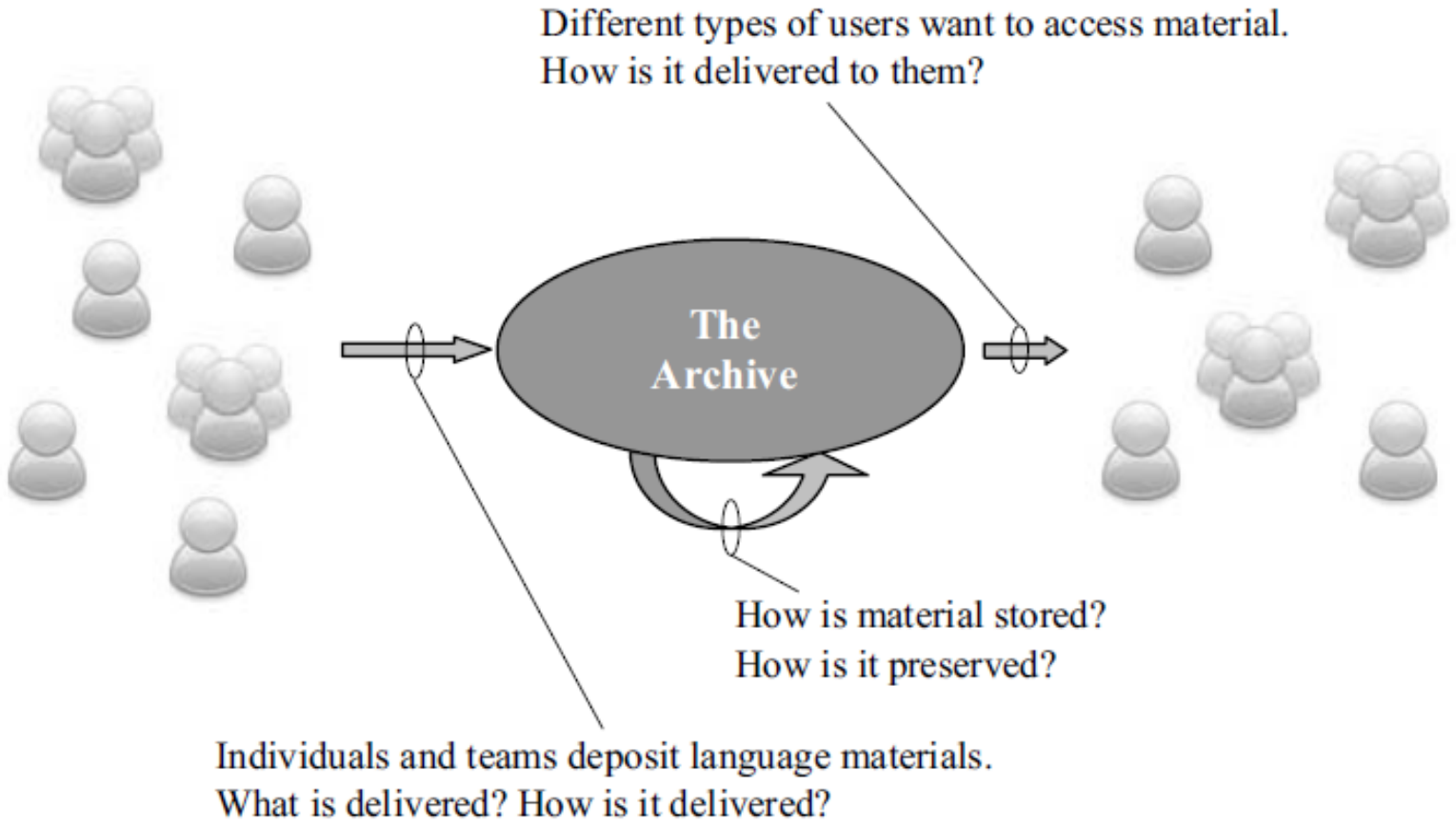


Figure 1. Different kinds of interactions with the archive

ANOTHER MODEL: NATHAN (2010), AFTER AUSTIN (2014)

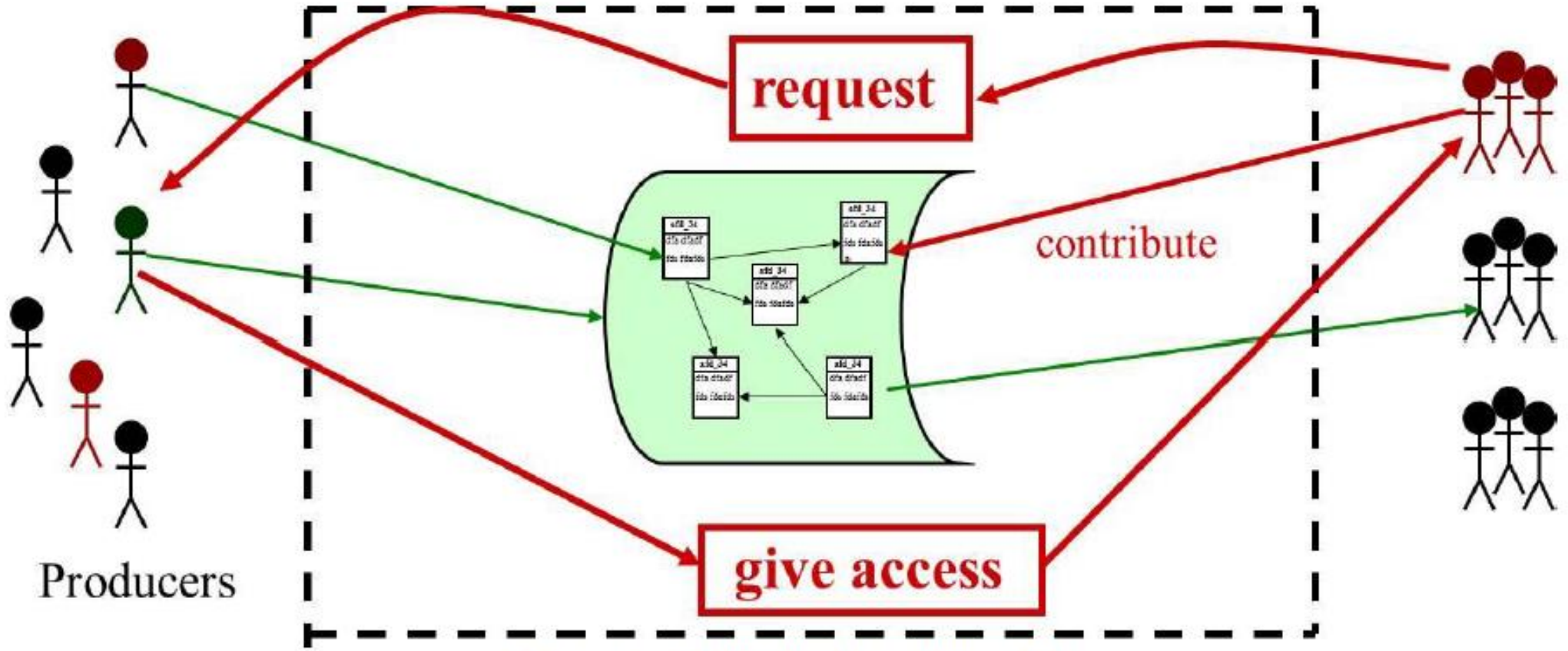


Figure 3: ELAR Archive 2.0 model

LANGUAGE ARCHIVES

http://dobes.mpi.nl/ (DOBES = **Do**kumentation **be**drohter **S**prachen)

https://elar.soas.ac.uk/ (ELAR = **E**ndangered **L**anguages **Ar**chive)

https://www.ailla.utexas.org/ (**AILLA** is a digital archive of recordings and texts in and about the indigenous languages of Latin America)

http://catalog.paradisec.org.au/ (**PARADISEC** = Pacific And Regional Archive for Digital Sources in Endangered Cultures)

http://lacito.vjf.cnrs.fr/pangloss/index_en.htm (**PANGLOSS** collection)

http://siberian-lang.srcc.msu.ru/ Siberian Lang (МАЛЫЕ ЯЗЫКИ СИБИРИ: НАШЕ КУЛЬТУРНОЕ НАСЛЕДИЕ**)**

http://inne-jezyki.amu.edu.pl/Frontend/ Poland’s Linguistic Heritage

OTHER INITIATIVES (SITES OFTEN CONTAIN FURTHER USEFUL LINKS!)

CLARIN ERIC

CLARIN = **C**ommon **La**nguage **R**esources and Technology **In**frastructure

ERIC = **E**uropean **R**esearch **I**nfrastructure for Language Resources and Technology

https://www.clarin.eu/content/services

DELAMAN = **D**igital **E**ndangered **La**nguages and **M**usics **A**rchives **N**etwork

http://www.delaman.org/

Linguistic **D**ata **C**onsortium

https://www.ldc.upenn.edu/