DOCUMENTARY LINGUISTICS I

prof. Nicole Nau, UAM winter 2018/2019

Fourth lecture

23 October 2018

TOPICS OF THE DAY

From metadata to archives:

- What are metadata for?
- Metadata and catalogue
- Finding language documents on the Internet
- Challenges of legacy materials
- What are «undead voices» and how can they be revived?
- Language archives from different perspectives

METADATA ARE NEEDED ON VARIOUS LEVELS!

For each recording, metadata usually contain information on:

- participants (speakers and bystanders, their roles)
- time and location
- recorded by, recorded with
- …

Speaker metadata: age, sex, ..., dialect

Metadata for the whole documentation: information on the language

METADATA: MINIMUM ACCORDING TO JOHNSON (2004) => CONCERNING THE RECORD(ING)

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

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

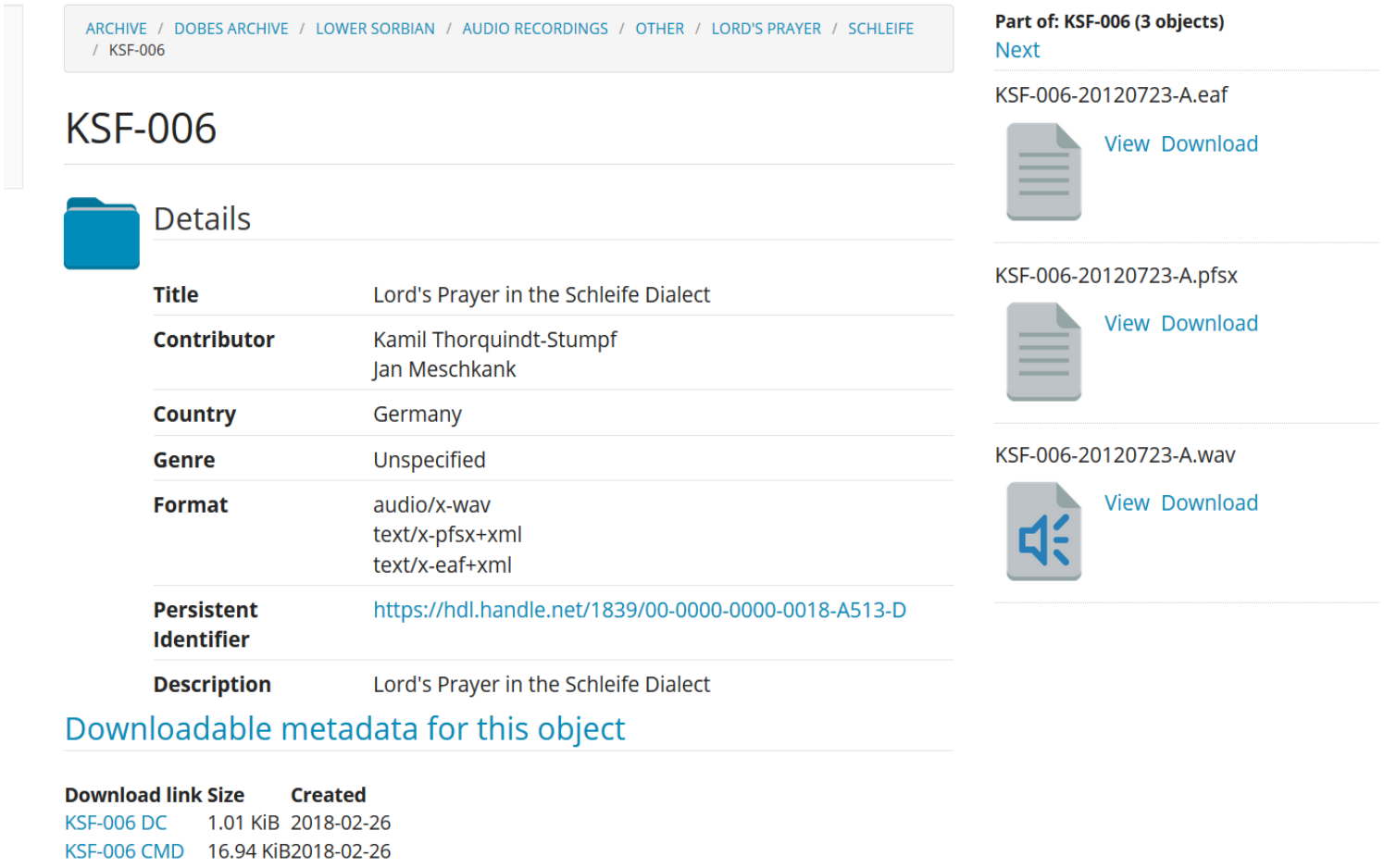
TYPES OF METADATA WITH EXAMPLE

(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

METADATA IN AN ARCHIVE: EXAMPLE



ARCHIVE / DOBES ARCHIVE / LOWER SORBIAN / AUDIO RECORDINGS / OTHER / LORD'S PRAYER / SCHLEIFE / KSF-006

KSF-006

Details

Title	Lord's Prayer in the Schleife Dialect
Contributor	Kamil Thorquindt-Stumpf Jan Meschkank
Country	Germany
Genre	Unspecified
Format	audio/x-wav text/x-pfsx+xml text/x-eaf+xml
Persistent Identifier	https://hdl.handle.net/1839/00-0000-0000-0018-A513-D
Description	Lord's Prayer in the Schleife Dialect

Downloadable metadata for this object

Download link	Size	Created
KSF-006 DC	1.01 KiB	2018-02-26
KSF-006 CMD	16.94 KiB	2018-02-26

Part of: KSF-006 (3 objects)
Next

KSF-006-20120723-A.eaf
View Download

KSF-006-20120723-A.pfsx
View Download

KSF-006-20120723-A.wav
View Download

https://archive.mpi.nl/islandora/object/lat%3A1839_00_0000_0000_0018_A513_D

METADATA: WHERE TO FIND ADVICE (ON TECHNICAL QUESITONS)

«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 where 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)

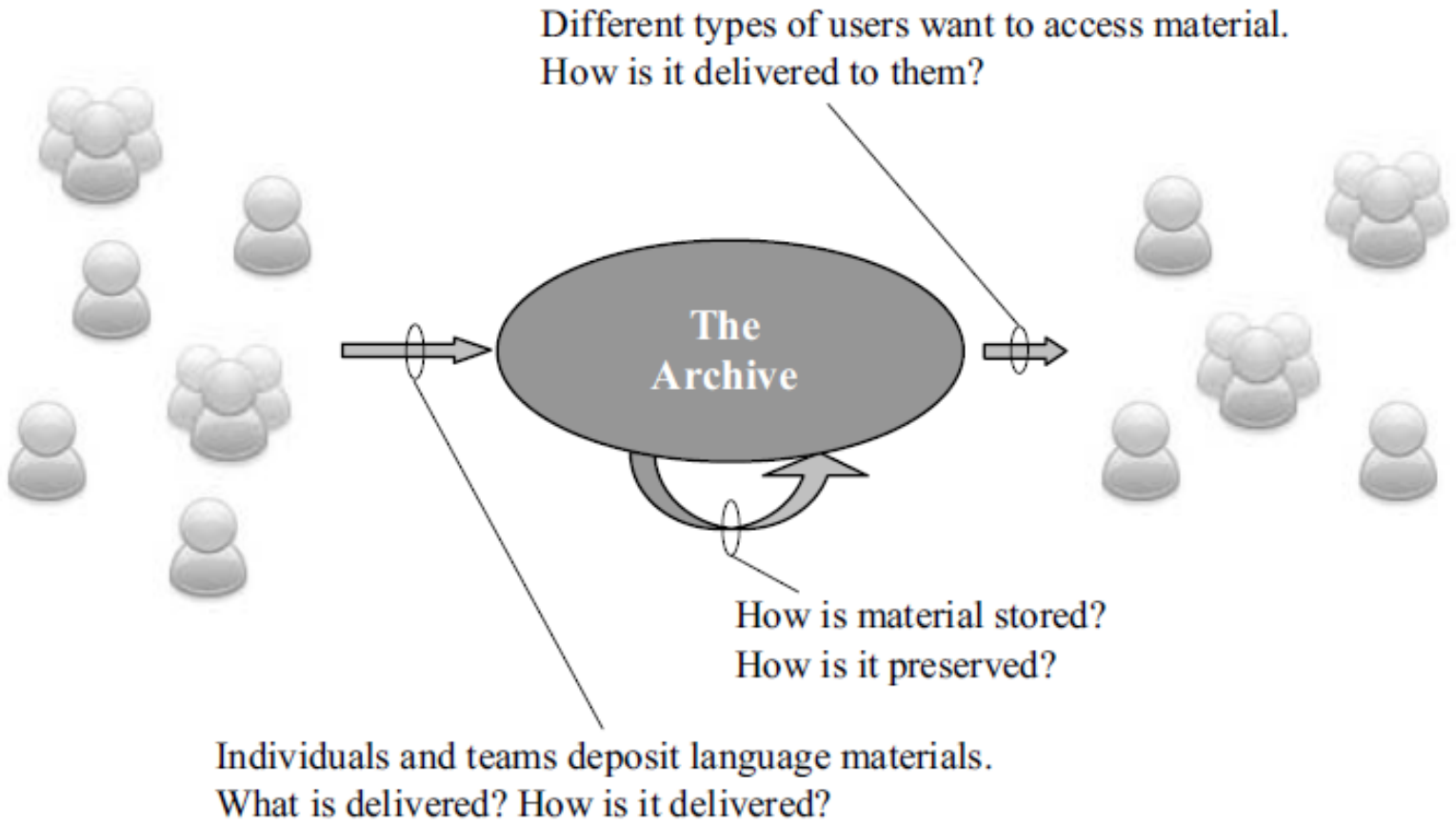


Figure 1. Different kinds of interactions with the archive

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

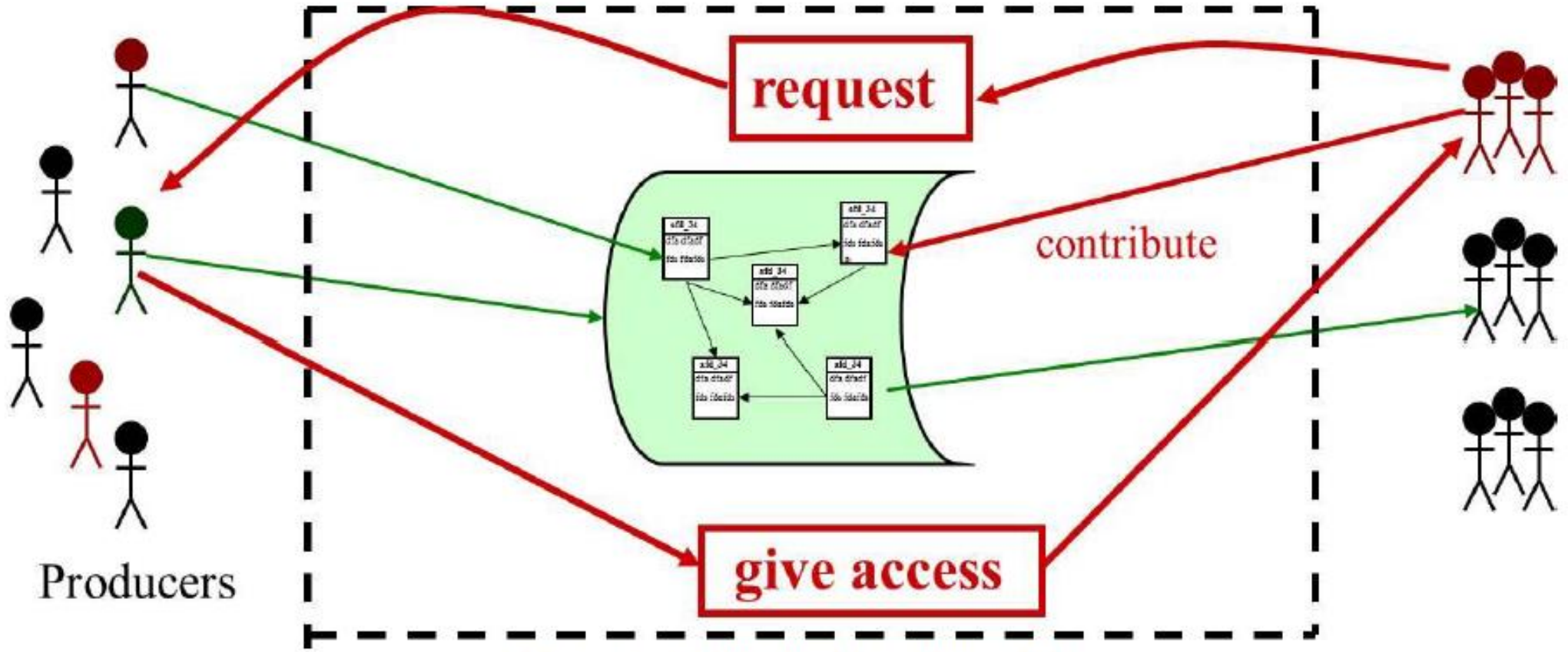


Figure 3: ELAR Archive 2.0 model

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."

JOHNSON (2004): WHAT KINDS OF THINGS ARE GOOD CANDIDATES FOR ARCHIVAL PRESERVATION?

- public events: ceremonies, oratory, dances, chants;
- narratives: historical, traditional, myths, personal, children's stories;
- instructions: how to build a house, how to weave a mat, how to catch a fish;
- literature: oral or written, poetry, any creative work that people may offer;
- conversations: anything that's not gossip or too personal, e.g. conversations about a recent school event or holiday;
- (continued next slide)

CONTINUED

- transcriptions, translations, and annotations of recordings, in which anonymity is preserved if necessary;
- field notes, elicitation lists, orthographies - anything other people might find useful;
- datasets, databases, spreadsheets and other secondary (unpublishable) materials;
- sketches of all kinds: grammar, ethnography;
- photographs of speakers and public events.

A) YOUR FIRST TASK FOR GRADING => ELECTRONIC HANDOUT

B) Reading for the next two weeks (23. + 30.10.)

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. Language Documentation and Description, ed. Peter K. Austin, vol. 2, 140-153. London: SOAS. Available at: http://www.elpublishing.org/PID/026.

Mosel, Ulrike. 2006. Fieldwork and community language work (in *Essentials of Language Documentation*)

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