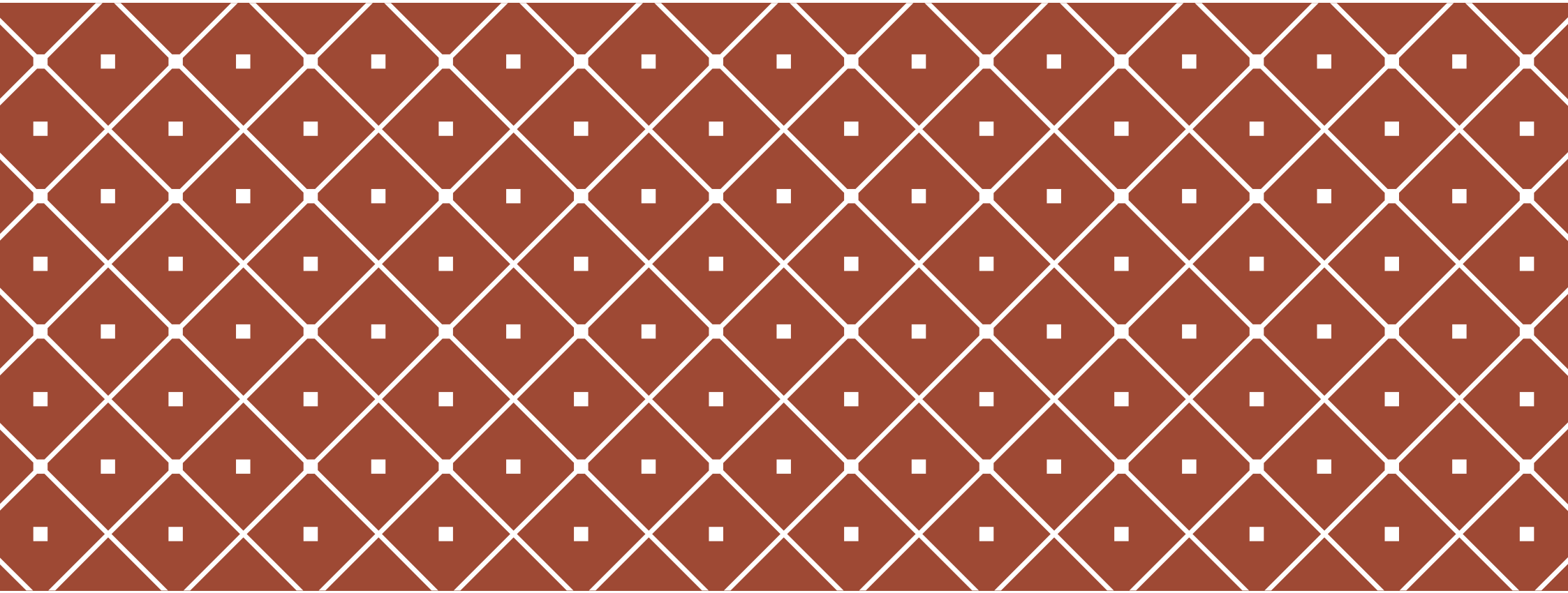

DOCUMENTARY LINGUISTICS I

prof. Nicole Nau, UAM winter 2016/2017

Third lecture

24 October 2016

TOPICS OF THE DAY

More on

- metadata
- archives
- standards

QUESTIONS FOR WARMING UP (3 MINUTES FOR REFLECTION / CONSULTATION)

1) Why is it important to store primary data in open archives?

2) Why is metalinguistic knowledge a necessary part of primary data, and how can we access a speaker's metalinguistic knowledge?

3) What are functions of metadata? (short answer)

EXAMPLES OF LANGUAGE DOCUMENTATION

1) Non-professional documentation (small samples of languages)

http://languagelandscape.org/

What is recorded? Why?

Which metadata are given?
Do you think this is enough? What is missing?

TYPES OF METADATA (AUSTIN 2006)

There are several types of metadata:

1. *Cataloguing* – information useful to identify and locate data, e.g. language code, file ID number, recorder, speaker, place of recording, date of recording, etc.
2. *Descriptive* – information about the kind of data found in a file, e.g. an abstract or summary of file contents, information about the knowledge domain represented.
3. *Structural* – for files that are organized in a particular way, a specification of the file structure, e.g. that a certain text file is a bilingual dictionary.
4. *Technical* – information about the kind of software needed to view a document, details of file format, and preservation data.
5. *Administrative* – background information such as a work log (indicating when the files were last saved or backed up), records of intellectual property rights, moral rights, and any access and distribution restrictions imposed by researcher and/or community.

AUSTIN'S EXAMPLE

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: WHERE TO FIND ADVICE

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

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

OLAC Metadata Standard – explanations:

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

(Standard, more formal:
http://www.language-archives.org/OLAC/metadata.html)

LDC Filename conventions and metadata:

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

ON THE BENEFIT OF STANDARDIZING METADATA AND ADDITIONAL INFORMATION

You want to find records of the language Aleut.
Compare the following web-pages:

Non-standardized ressource list:
http://www.native-languages.org/aleut.htm

OLAC ressource catalogue:
http://www.language-archives.org/language/ale

OLAC ressources for Basque:

http://www.language-archives.org/language/eus

BACK TO METADATA: SOME PROBLEMS IN ORGANIZING METADATA

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 …

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

OLAC STATISTICS

(MAYBE SIZE IS A PROBLEM AFTER ALL?)

Name	Value
Number of Archives	60
Archives with Fresh Catalogs	30
Archives with Five-star Metadata	21
Number of Resources	143666
Number of Resources Online	97862
Distinct Languages	8056
Distinct Linguistic Subfields	28
Distinct Linguistic Types	3
Distinct DCMI Types	13
Average Elements Per Record	47.4
Average Encoding Schemes Per Record	9.8
Average Metadata Quality Score	8.3
Record views per month	10731
Click-throughs per month	2392
Last Updated	2017-07-14
Known Integrity Problems	6149

Note: Record views and click-throughs are for the month of 2016-09.

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/ see also: https://vlo.clarin.eu/?4

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/

BEST PRACTICES – SEVEN DIMENSIONS IDENTIFIED BY BIRD & SIMON (2003)

- Content (of the documentation)
- Format (of files, characters, structures)
- Discovery (make the resource visible)
- Access (regulate who can use the resource, and how)
- Citation (how to cite the resource or parts of it)
- Preservation (ensure the resource will be accessible in the future)
- Rights (who owns the resource and who is allowed to use it)

example for regulation of access and rights: http://www.ailla.utexas.org/site/use_conditions.html
citation guidelines: http://ailla.utexas.org/site/citation.html

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.

LANGUAGE ARCHIVES: HOW USERFRIENDLY ARE THEY?

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

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

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

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

HOMEWORK

Do the **first task** (to be submitted 7 November, or before; see handout of the second lecture).

Background reading: in the section «Guide to language archives, examples of language documentation» at the bibliography page

http://elldo.amu.edu.pl/?page_id=229

and maybe have a look at Bird & Simon (2003) for ideas about best practices