Metadata and Language-Resources

Jacquelijn Ringersma
Paul Trilsbeek

Documentation and Archival Training Workshop
Guwahati, Assam, India

Content

What is metadata?

Why should you create metadata?

Metadata for language resources

IMDI editor + exercise creating metadata + home work

Tomorrow:

Discuss home work

Arbil – Archive builder

Practical Arbil:

building a corpus and creating metadata with Arbil

Why archiving?

Misconceptions about archiving

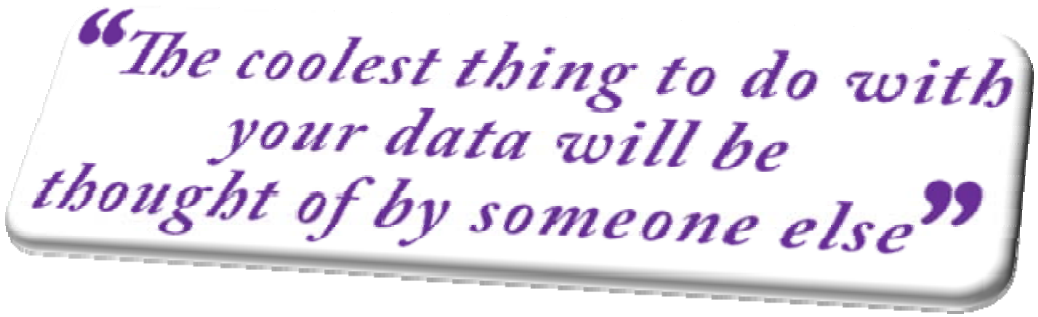
1. Your stuff is buried here and gone forever

Misconceptions about archiving

1. Your stuff is buried here and gone forever
2. Other linguists will take advantage of your hard work and take away your good ideas

But the actual truth is that:

1. The people who care about your work are the members of the speech communities –
   and they care about it in a different way than you do



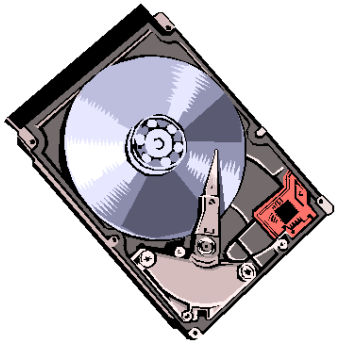
Is there a danger that we loose digital data?

YES,

UNESCO: 80% of our recordings is endangered

How much of your data and files on the notebook is organized, backed-up?

How long can media and formats be accessed?



Correct conceptions about archiving

1. It requires discipline
2. It creates a bit of techno noise

Show browser

The rewards:

Long term preservation and access

Different ways of presenting your data are possible

What is metadata?

Metadata in General

Metadata is “transcendental”

- Data about data
  (It is the ‘who, what, where and when’ of a document)
- Structured data about data
- *Internet*: machine readable data about data

Metadata is data describing a (set of) digital resource

Metadata in General

Structured

in a standardized fashion using a metadata model

Metadata model

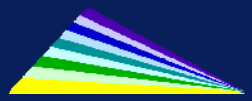
Dublin Core, OLAC, IMDI

- Metadata scheme (elements, structure)
- Metadata controlled vocabulary

Metadata in General

Dublin Core (DC) Metadata Set

Content	**Intellectual Property**	**Instance**
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Language	Rights	Identifier
Relation		
Coverage		
Source		



Metadata in General

DC example:

Content:	DC.Title = “The white tiger”
	DC.Language = “English”
IP:	DC.Creator = “Aravind Adiga”
Instance	DC.Format = “print”
	DC.Date = “2008-04-22”

Simple elements do not allow resource specific metadata, e.g. how to specify that the contributor role is ‘singer’?

When to use DC:

Interoperability:

You need to offer your data to other communities using commonly understood semantics

You only need a core description

… and find the DC vocabulary/names acceptable

You only have limited resources and can only manage to enter a few fields.

Why should you create metadata?

You do not want your data to float in the cyberspace and get lost!

Why metadata

Creation of data bases according to metadata structure

(Re) Finding resources

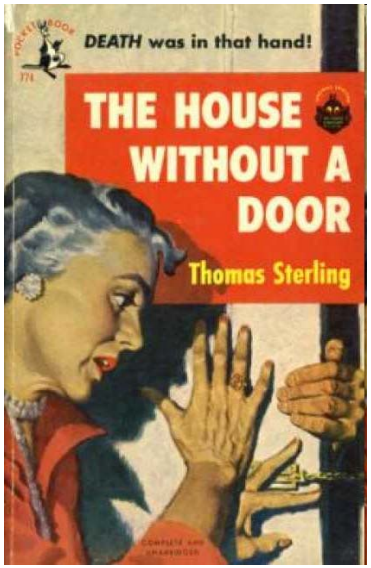
Using free text “key words”

Search engines that work in a structured manner

Those who don’t take care over their metadata are doorless.

Content is expensive to create

Just what good is your content if the people who need to read it can’t find it?



Metadata for language resources

Language resources that make up corpora:

- (Digital) video or audio recordings, photographs
- Digitisations of images used as stimuli
- Transcription files
- One or more analysis files
- Field notes and experiment descriptions

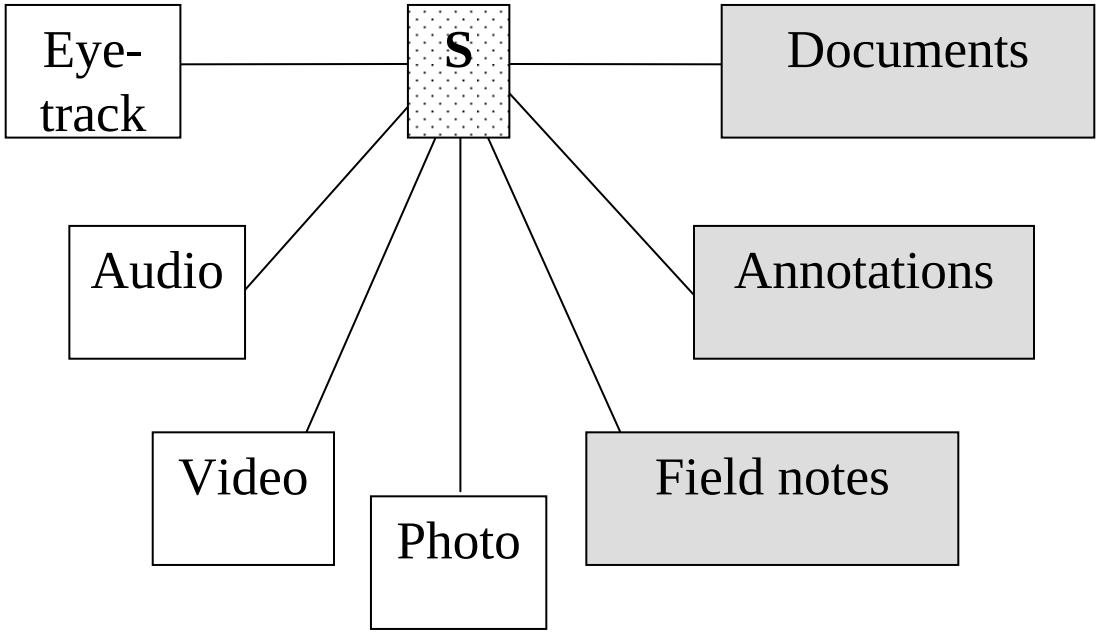
Special to Language Resources

- In the linguistic domain often *clustered* resources
- Clustered because they refer to or result from the *same linguistic event/performance*.
- In our archive terminology: **session** or **resource bundle**

Metadata for Language resources

Session or ‘Resource bundle’ concept:

Bundle of tightly related resources
Basic unit of linguistic analysis
Described with the same set of metadata (S)



IMDI metadata set (1)

Aim: describe a session or resource bundle

- with a sufficiently rich metadata set
- using domain specific names

IMDI = ISLE Metadata Initiative
ISLE = International Standards of Language Engineering

IMDI metadata set (2)

Categories

Administrative – (date, tool, version …)
General – (project, location …)
Content – (language, genre, modality …)
Actors/Participants – (biographic/contact information)
Resources – (URL, type, format, accessibility …)

Context