

MAX-PLANCK-GESELLSCHAFT

Language Archiving Technology
Archiving challenges and metadata

Alexander König
The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Content





1. What is Metadata?
2. Why Metadata
3. Metadata for Language Resources
4. Organizing metadata for Language Resources: IMDI
5. Other metadata model: CMDI
6. Arbil metadata editor

1) What is **METADATA**?



What is metadata?

- Metadata is “transcendental”
  - Data about data: the ‘who, what, where and when’ of a document
  - Structured data about data
  - *In the computer age*: machine readable data about data
- Metadata is data describing a (set of) digital resource(s)





2) **Why**

Metadata?



Why metadata?



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

Why metadata?

What good is your content if the people who need to read it can't find it?

(Re) Finding Resources:

MAX-PLANCK-GESELLSCHAFT

Ways of (Re) Finding resources:

1. Using free text “key words” (i.e., Google)
2. Special search engines that exploit the structure of metadata
   - IMDI Browser
   - VLO





3) Metadata for Language Resources

Language Resources

Your corpus might consist of different

Language Resources:

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

Language Resources

Language Resources are often ***clustered***, i.e. they might refer to or result from the *same linguistic event/performance.*





4) How to organize metadata for language resources?

Metadata Models

In general, Metadata is organized according to a ***Metadata Model***

Some Metadata models

- Dublin Core (DC)
- OLAC
- IMDI
- CMDI

Metadata Model for Language Resources

MAX-PLANCK-GESELLSCHAFT

IMDI:

- The ISLE Meta Data Initiative
- A metadata model to describe multi-media and multi-modal language resources.
- It provides interoperability for browsable and searchable corpus structures and resource descriptions with help of specific tools.
- Website: http://tla.mpi.nl/imdi-metadata





Why IMDI is so suitable for Language Resources?

Session

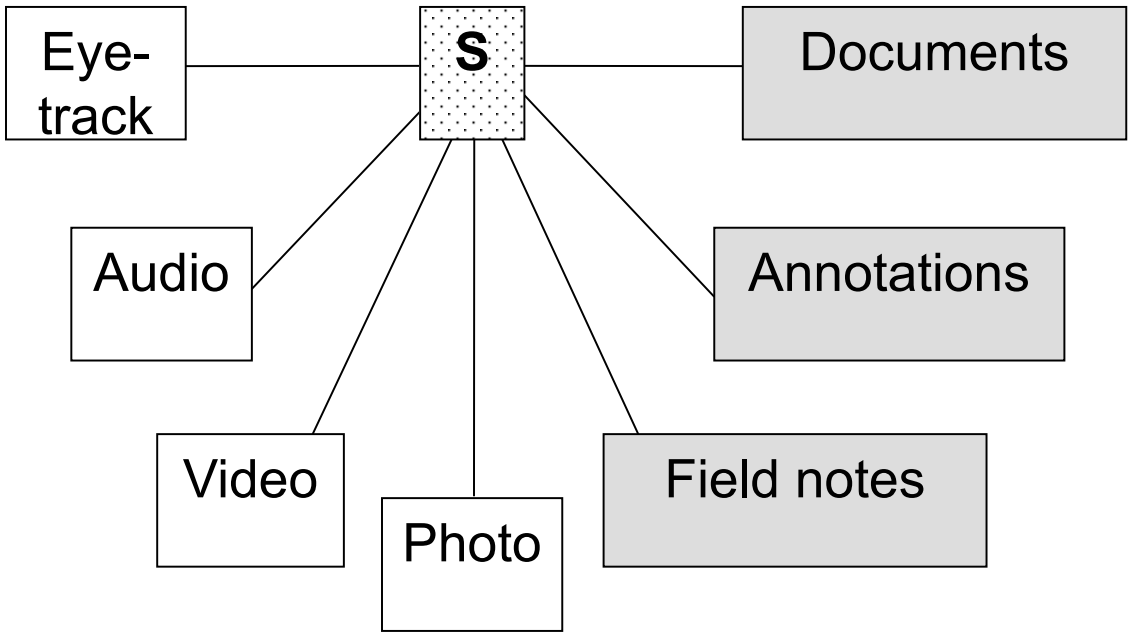
Session (or ‘Resource bundle’):

- Collection of tightly related resources
- It contains all the resources that pertain to the same linguistic event (i.e. *clustering* of linguistic resources)

Structure of a Session (S)

Two Types of Resources:

1. Media Resource: Eye-track, Audio, Video, Photo
2. Written Resource: Documents, Annotations, Field Notes



Viewing IMDI Metadata

MAX-PLANCK-GESELLSCHAFT

The ***IMDI BROWSER***

→ (http://corpus1.mpi.nl/ds/imdi_browser/)





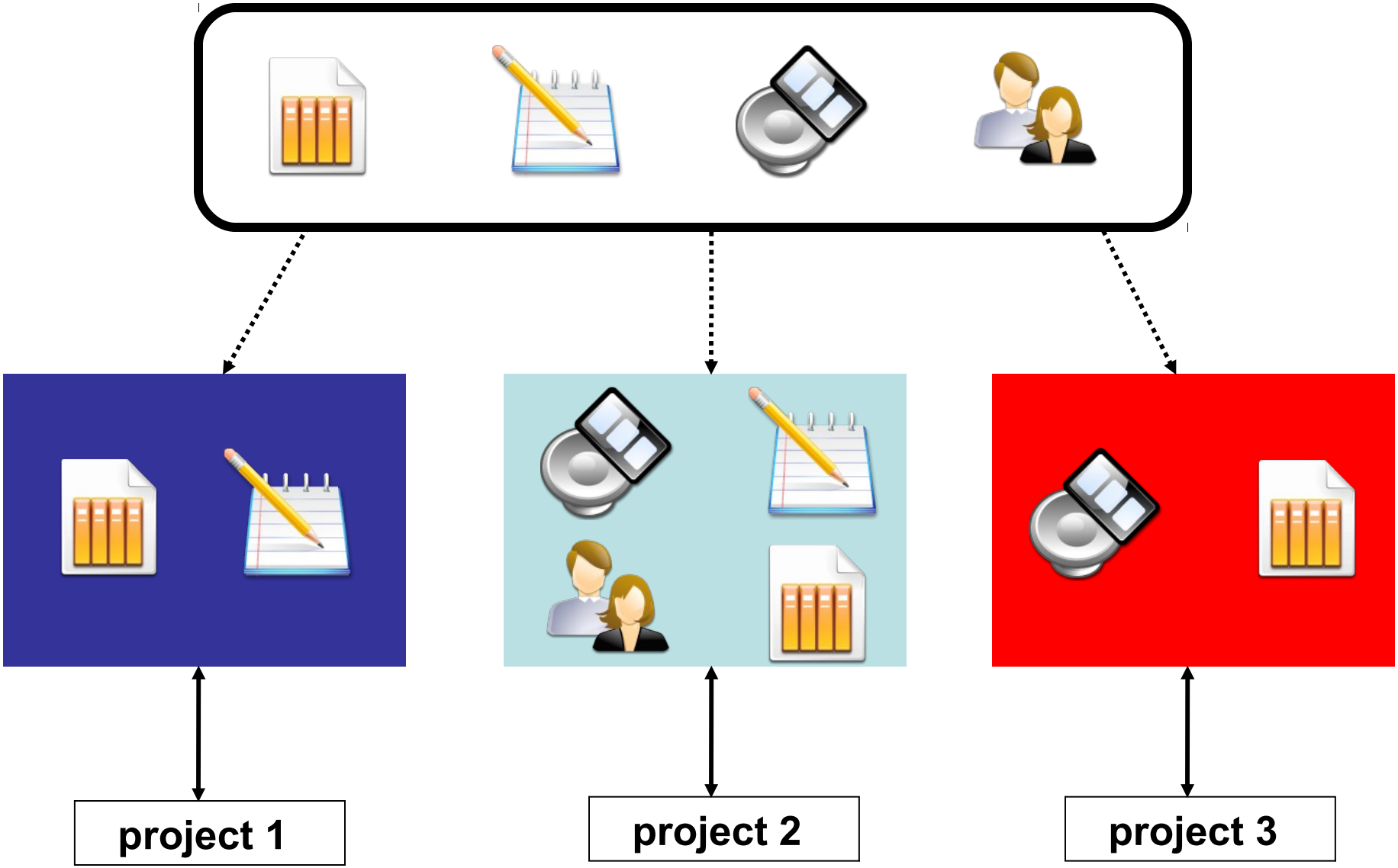
5) Other metadata model: **CMDI**

Metadata Schema - CMDI

CMDI = Component MetaData Infrastructure

- Based on a modular concept
- Integral part of the CLARIN (Common Language Resources and technology INfrastructure) initiative
- More information at http://www.clarin.eu/cmdi

Metadata Schema - CMDI



6) ARBIL





What is Arbil?



Arbil

(short for “Archive Builder”)

is an application for organizing research data and associated metadata into a format appropriate for archiving.



Arbil Features



- There are many features in Arbil that enable users to view and edit their data.
- The data can be viewed side by side in tables and bulk edited in the same table.
- Arbil is designed so that it can be used ***offline*** in remote locations.

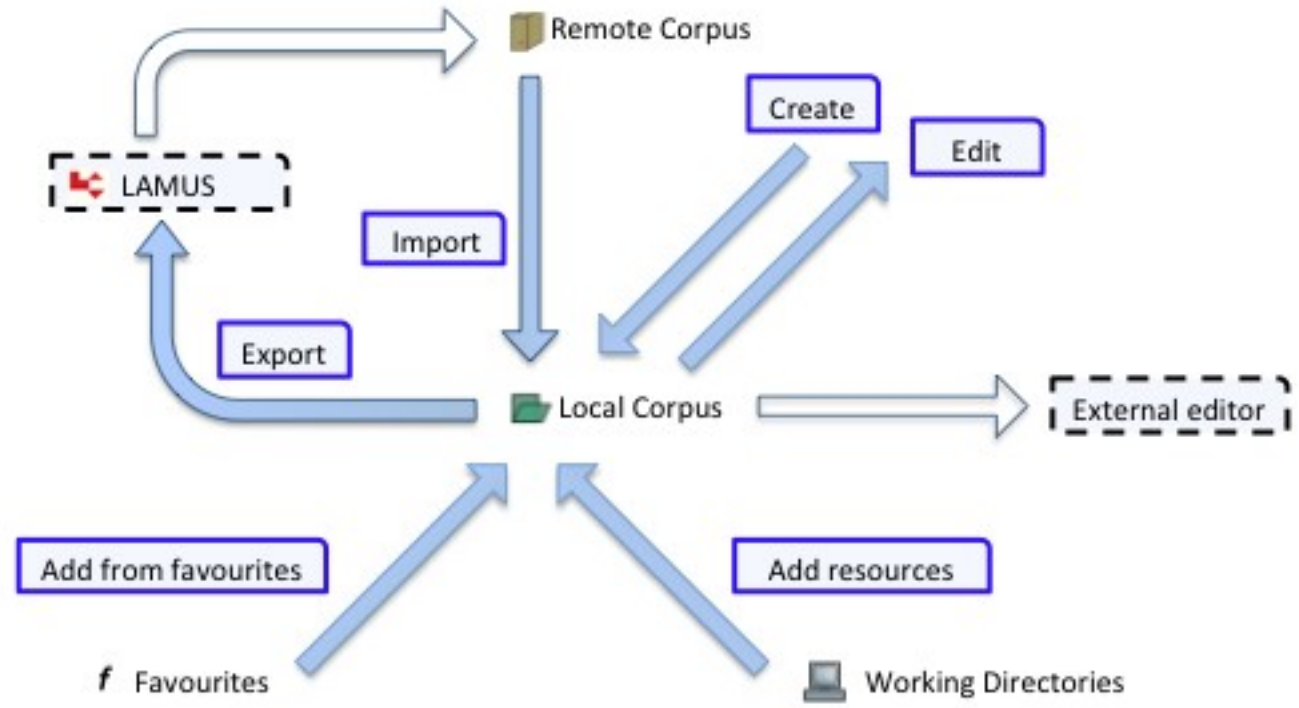
Main Arbil Principles





- Workflow focused
- Table views
- Drag and drop
- Bulk copy and paste
- Multiple undo and redo

Arbil Workflow



Get Arbil

There is a link to Arbil on the MPI website

http://tla.mpi.nl/tools/tla-tools/arbil/

Providing you already have Java installed the webstart version is the fastest way to start

Alternately there are installers for

- Windows,
- Mac
- and Ubuntu (Debian).

