

# Developing a Standard for Meta-Descriptions of Multimedia Language Resources

Daan Broeder, Pirkko Suihkonen ,  
Peter Wittenburg

Max-Planck Institute for  
Psycholinguistics



Max Planck Institute for Psycholinguistics

---

# MPI Projects dealing with Metadata for LR's

- Browsible Corpus (BC)
- EAGLES/ISLE (metadata part)
- resulting in IMDI proposal.
- DOBES (Documentation Bedrohter Sprachen)
- Dutch Spoken Corpus



# Language Resources

- Video tape
- Photographs
- Digitised video file
- Digitised photographs
- Digitisations of the images used as stimuli
- Transcription file
- One or more analysis files
- Field notes and experiment descriptions



# Special requirements

- In the linguistic domain language resources are clustered.
- Clustering through **session** concept
- No such concept exists in traditional librarian inspired metadata work

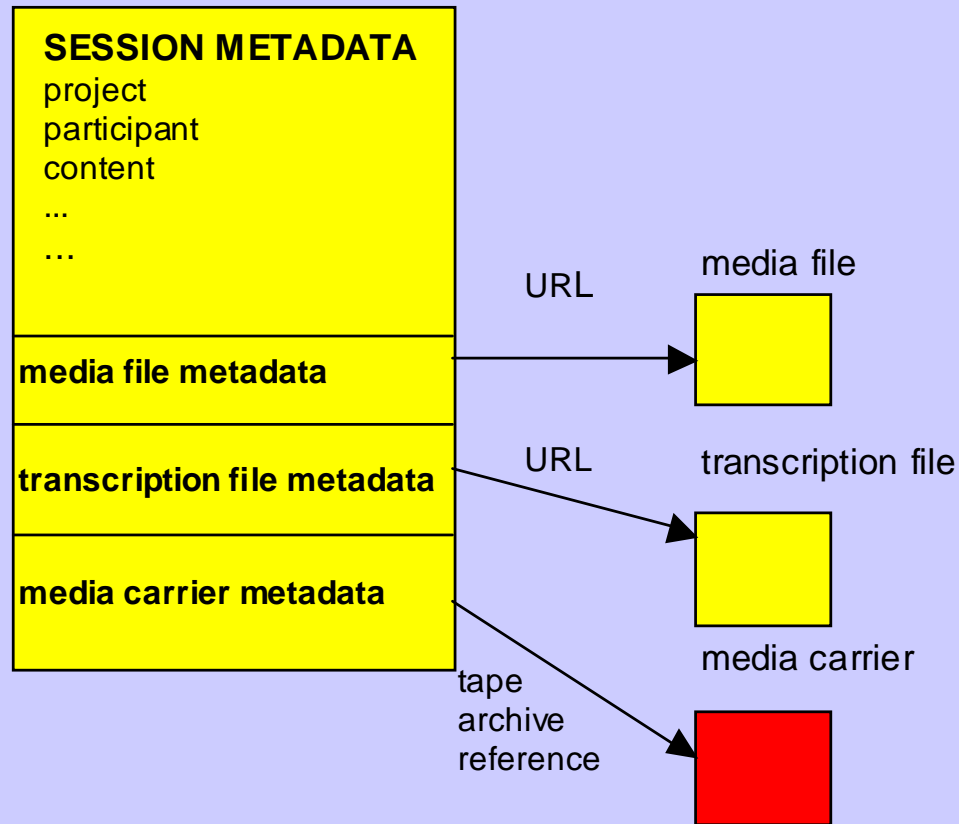


# Session Metadata Description

- Cluster of related Language Resources
- Media files, transcription + annotation files
- Metadata describing information connected with several resources like biographical info, content, genre etc.
- Metadata describing individual resources (file formats)



# Session Metadata Description File

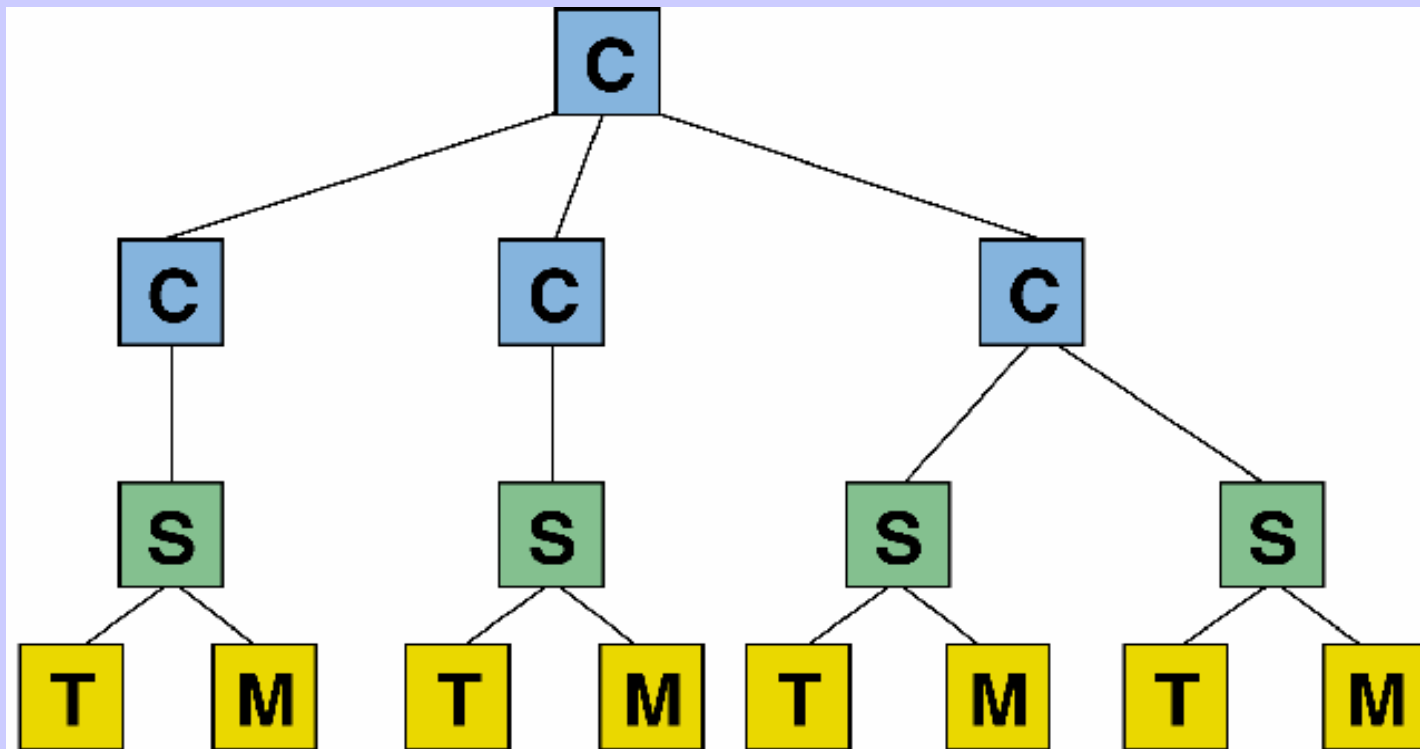


# Metadata Functionality

- Resource discovery:
  - Need a large set of keys with well defined value domains
- Use for corpus exploitation as well as cataloguing
- Browsing.
  - Need description fields and links to description files.
  - Structure the resources in interesting browsable hierarchies
  - Direct links to resources for access with tools



# Metadata Description File Hierarchy for Browsing

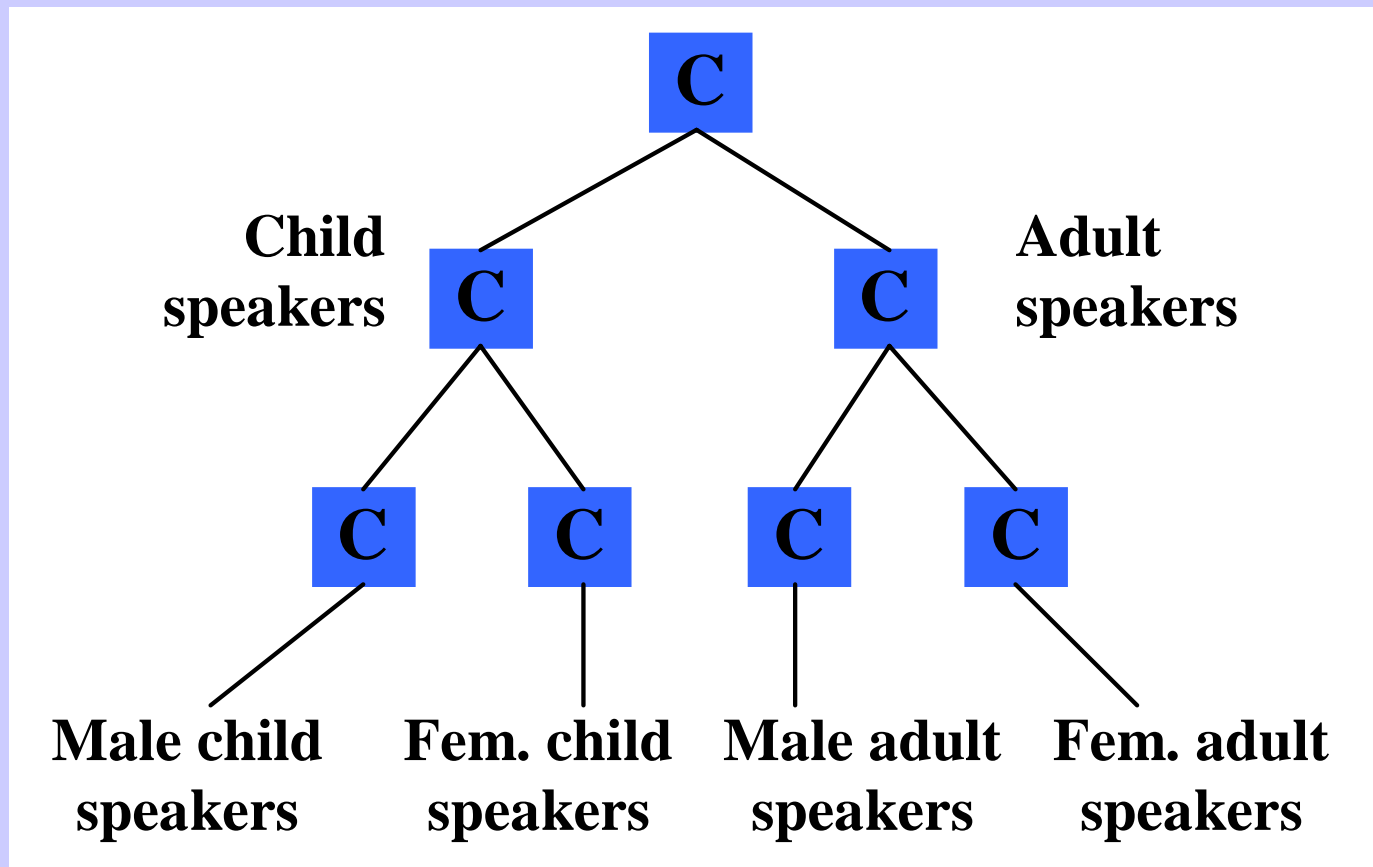


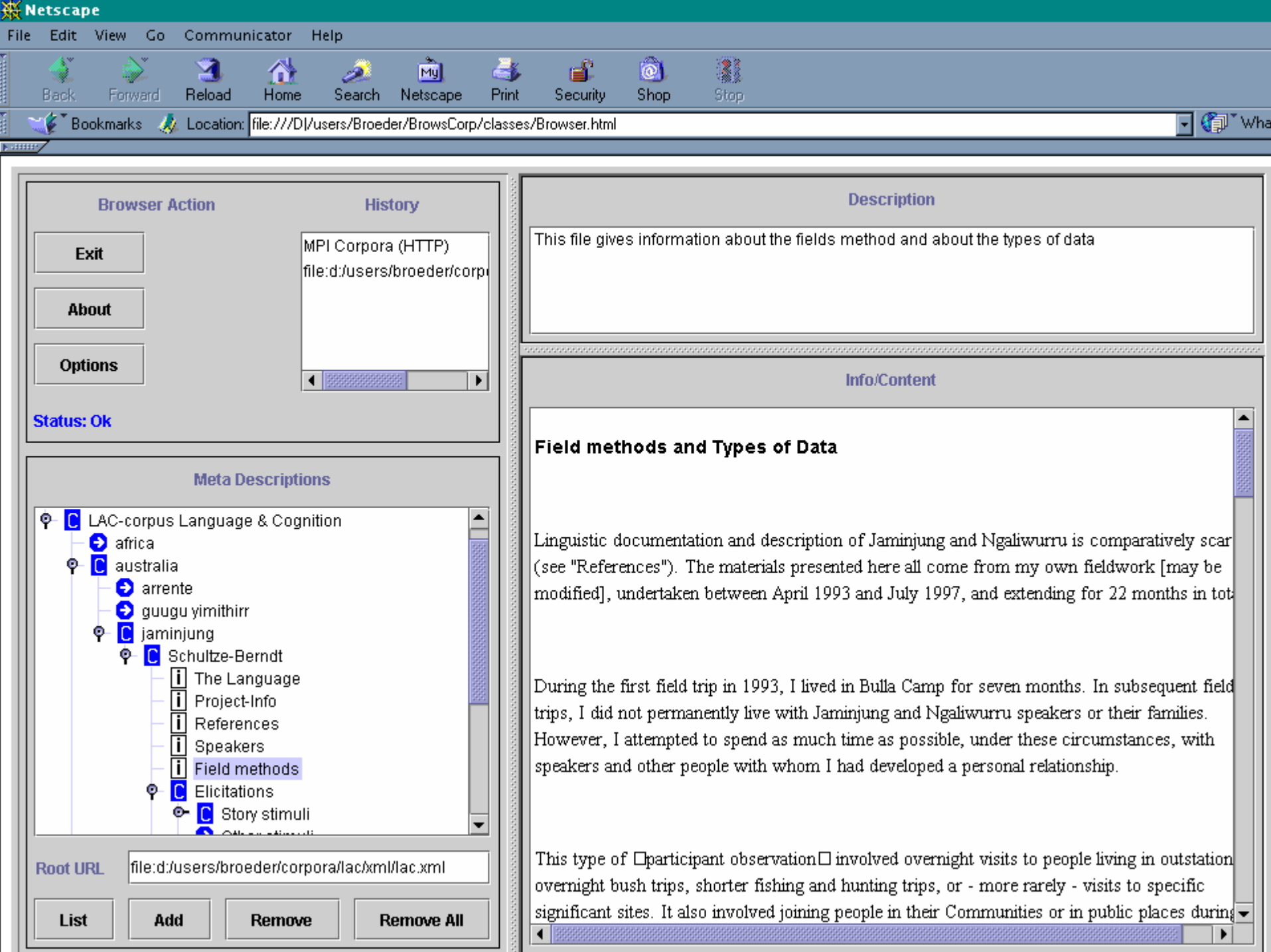
C = corpus MDF file  
S = session MDF file

T = transcript file  
M = media file



# Trivial Corpus Hierarchy





Browser Action

History

Exit About Options

MPI Corpora (HTTP) file:d:/users/broeder/corpo

Status: Ok

Meta Descriptions

Tree view of meta-descriptions: LAC-corpus Language & Cognition, africa, australia, arrente, guugu yimithirr, jaminjung, Schulze-Berndt, The Language, Project-Info, References, Speakers, Field methods, Elicitations, Story stimuli

Root URL file:d:/users/broeder/corpora/lac/xml/lac.xml

List Add Remove Remove All

Description

This file gives information about the fields method and about the types of data

Info/Content

Field methods and Types of Data

Linguistic documentation and description of Jaminjung and Ngaliwuru is comparatively scar (see "References"). The materials presented here all come from my own fieldwork [may be modified], undertaken between April 1993 and July 1997, and extending for 22 months in tot

During the first field trip in 1993, I lived in Bulla Camp for seven months. In subsequent field trips, I did not permanently live with Jaminjung and Ngaliwuru speakers or their families. However, I attempted to spend as much time as possible, under these circumstances, with speakers and other people with whom I had developed a personal relationship.

This type of participant observation involved overnight visits to people living in outstation overnight bush trips, shorter fishing and hunting trips, or - more rarely - visits to specific significant sites. It also involved joining people in their Communities or in public places during

# How to Describe Resources ?

- Minimalist approach: Use few but “general” understandable metadata elements e.g. Dublin-Core with 15 elements
- Exhaustive approach: Use a large extensive set of specialised metadata elements.  
MPEG7



# Minimalist Approach

- + Little typing
- + Better interoperability, other communities using the same set of elements
- limited usefulness for answering domain specific questions. E.g. “Give me all transcriptions of female native jaminjung speakers younger than 18”



# Exhaustive Approach

- + Domain specific questions can be answered
- More specification work
- Limited interoperability with other communities (and tools)



# Compromise

- Metadata set should be specialised and large enough to answer our specific questions.
- Right tools can minimise typing effort by reusing existing descriptions
- Interoperability issue can be (partly) solved



# Metadata Description Editor

- Editor can use existing descriptions to generate new ones.
- Editor can support controlled vocabularies



## General Content Data

Description:

EPISODE: 06 Stage Direction; the ashtray experiment ;

Keys:

Name	Value
KEYWORDS	EXPERIMENT, ASHTRAY/ LEXICON, REFERENCE TO SPACE/

Profile:

ladhk23.prt

## Languages

Name:

Dutch

Description:

Target Langage

Infofile:

Name:

Arabic

Description:

Source Langage

Infofile:

Name:

# Parallel Metadata Sets

- Pair an exhaustive set to a general minimal set such as DC.
- General tool may then answer “limited” questions for the “general public”
- Special domain specific tools can profit from the elaborate specific metadata set
- Need to provide mapping between the fields that overlap !!



# IMDI Proposal

Tries to describe a session in a structured way with a sufficiently rich metadata set

- General – (creator, project, location ...)
- Content – (language, genre, modality ...)
- Participants – ( biographic information )
- Resources – ( URL, format, accessibility ..)



# Flexibility

- User defined keyword/value pairs offers flexibility for specific projects and sub-communities
- Version numbering for the metadata standard allows tools to work also with very specialised resources such as lexica



# Access Control

Separation between metadata and resources allows a scheme where:

- Free access to (anonymous) metadata
- Restricted access to resources
- The mapping information between codes and real names is again a protected resource



# IMDI Open Issues

- Is the proposed IMDI metadata set sufficient?  
Genre & modality seem not specific enough
- Infrastructure issues:
  - where to put the metadata description hierarchies.
  - Metadata standard registry & maintenance.
  - Maintenance of metadata search/browser and editing tools

