

# Meta-Description Standard for Multi-Media Language Resources

P. Wittenburg, D. Broeder,

F. Offenga, D. Willems

Max-Planck-Institute for Psycholinguistics

[peter.wittenburg@mpi.nl](mailto:peter.wittenburg@mpi.nl)

# Problem Description

## MPI Nijmegen

- increasingly more multimedia language resources  
*corpora, lexicons, wordlists, grammar notes, ...*
- at MPI >30 researchers busy to create MMLR
- variety of access formats and tools
- only individuals know their way
- no clear archiving strategy even within an institute
- how to give notice about what is available (state)?
- how to guarantee re-usability by others?
- consequence: great waste of time & money
- needed urgent measures !!!

# Problem Description

## General Problem?

- YES - same situation at many places
- not only a technical problem

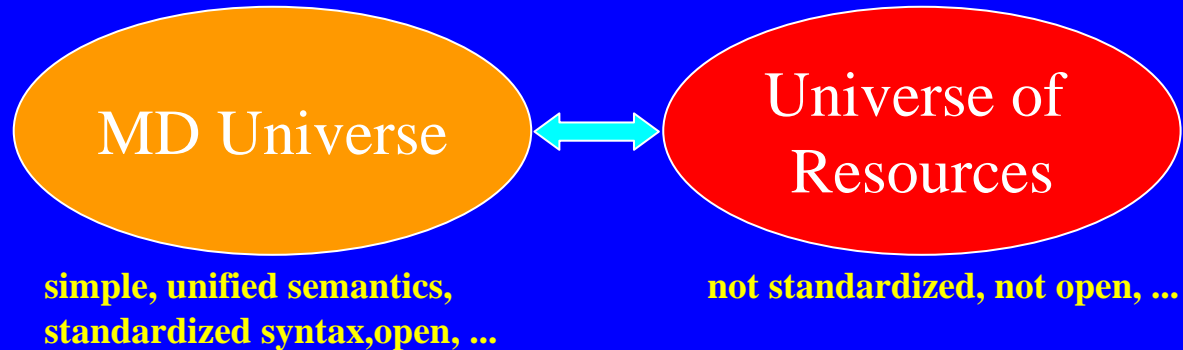
# Principle Idea

## Idea:

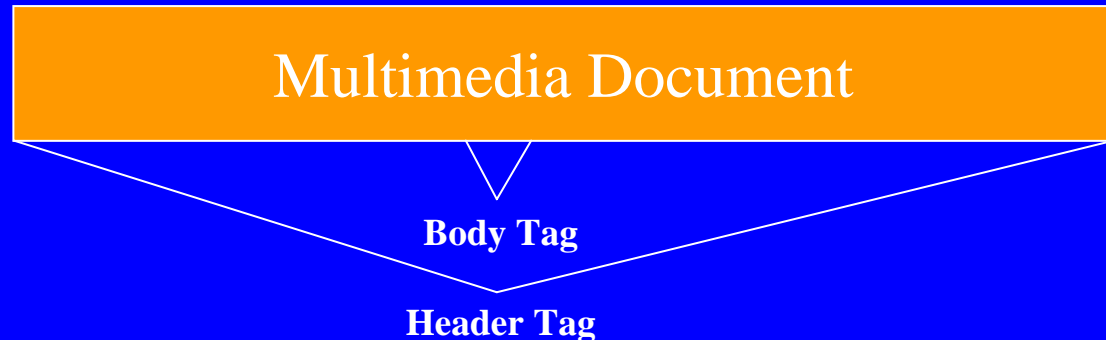
- Let's setup a distributed subspace of simple, linked MetaData Descriptions easily accessible by everyone.
  - Make it browsable & searchable
  - Allow the user to directly start a&a tools
- 
- HT: all data is metadata yes, but ...
  - MD concept comparable to DublinCore and not to MPEG7/TEI/CES

# Principle Idea

DC:

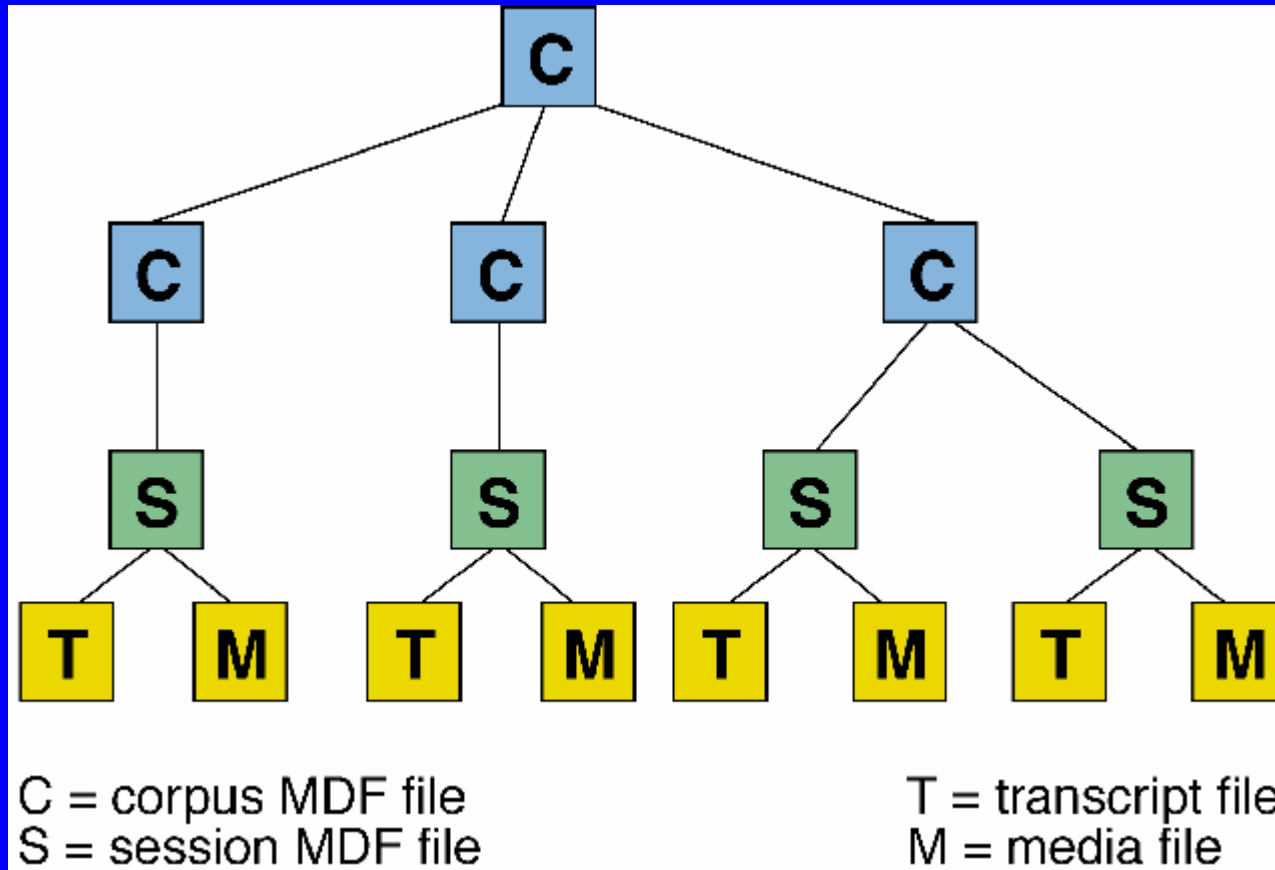


MPEG7/...

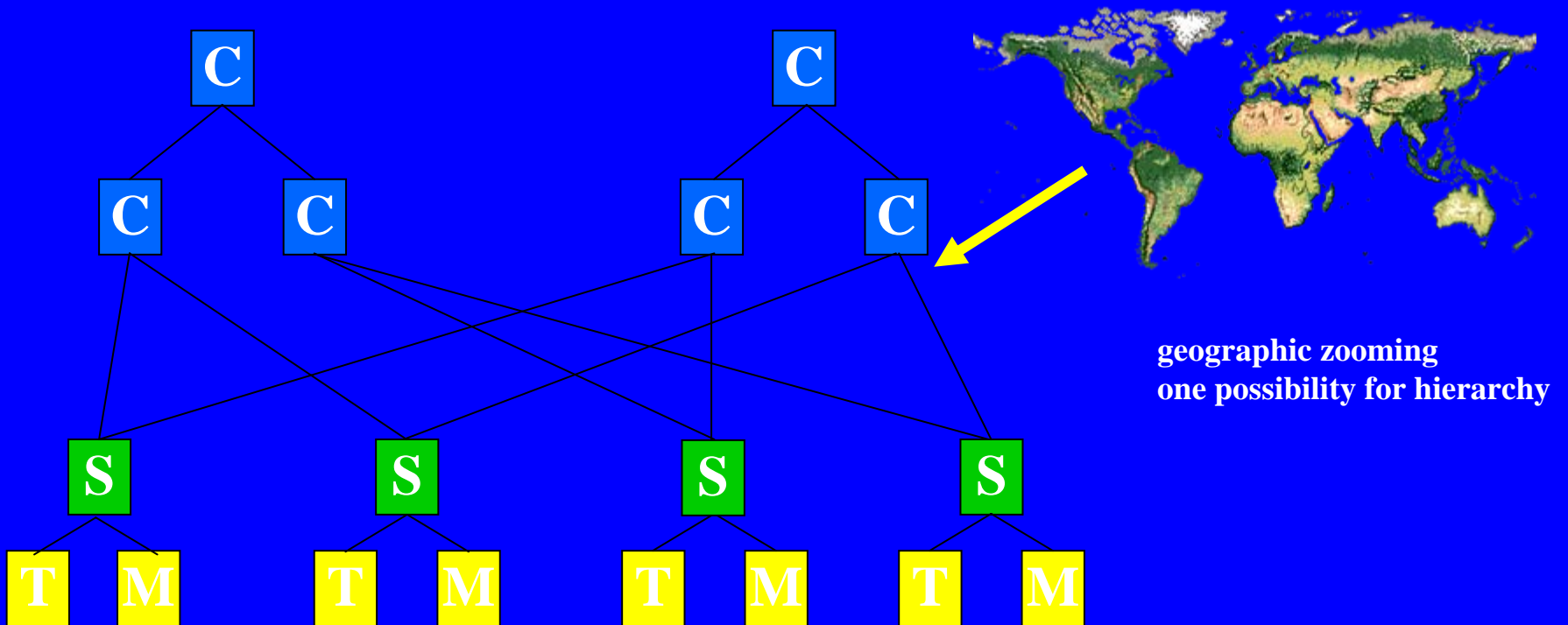


- no principal contradiction - but problem of consistency
- in MMLR much data is not open, not standardized, not well described, ...
- Dan Slobin: *“MD is our old dream to easily find the resources we need”*

# Principle Idea



# Principle Idea



# Broad Interest

MPI Nijmegen & Leipzig, Helsinki, Lund, ...:

- define MetaData Core set asap & apply it
- 1st MPI core set in use ([MPI MD set](#))
- [Editor & Browser Applets](#) are ready
- use MD as an entry point to start A&A Tools for (MM)  
LR to increase efficiency/easiness of access
- use MD as a mechanism to organise (internal) chaos
- for MPI&DOBES clear Workflow Scheme incl. MD  
*MD already in the field - otherwise no support!*

# Follow-Up: EAGLES/ISLE Project

## Idea:

- Let's see whether the idea can be applied to a larger community - therefore ask for EC money
- Follow the modular Warwick approach  
model MMLR world, use existing semantics & same syntax

## ISLE Mission Statement:

- Propose a Standard for Simple MetaData  
Descriptions of (MM) Language Resources
- Develop a Showcase for a Browsible &  
Searchable MetaData Universe

# EAGLES/ISLE Project

- funded by EC and NSF since February 00, 2 years
- **MPI Nijmegen**, SDU, IMS are responsible
- May 00: “US partners” were not yet interested
- May 00: **White Paper & LREC Workshop**
- since then:
  - set up of boards (SB, TB, AB)
  - meta data overview
- interested people should join the AB & comment
- **info under: [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)**

# Problems to be solved I

- Goal
  - **official**: Proposal & Showcase
  - **in principle**: agreement of the community & attractiveness
- Influences
  - expectations of the LR community
  - meta initiatives (DC, ...) and RDF
- Scope of LR and Community
  - type of LR to be included (restrictive or exhaustive)
  - sub communities to be approached (LE to anthropology)

# Problems to be solved II

- **Structure of Meta-Descriptions**
  - XML as basic syntax
  - mandatory core set, optional core set, extensions
  - is RDF a suitable framework?
- **Scope of Meta Data**
  - which characteristics to be described?
  - links to corpus files, tools to be directly started, ...
- **Meta-Element Vocabulary**
  - which meta elements to choose?
  - semantics to be defined

# Problems to be solved III

- **Element protection**
  - subject names have to be protected
- **Re-usage of other meta definitions**
  - re-usage of DC/CARD/... definitions
- **Requirements for Tools**
  - features of editors, browser, search tools
- **Practicable Scenario**
  - where to store, registry, linking, quality check, ...

# Meta Workshop Results

- High quality to be assured
- urgent need to overcome chaos
- no ontology available - need several attempts
- link annotation - meta data (changes)
- many native (non-XML) formats
- flexibility / dynamics important
  
- keep in touch with DC, MPEG7, W3C, ....

# Meta-Project - next steps

- make overview about existing meta-descriptions
- describe a list of meta-data categories  
(content, formal, rights, type of annotations, ...)
- describe useful meta-elements per category
- DOBES meeting: define a minimal core set asap

Main Data

**Name:** liela17k.1

**Description:** This is the ESF subcorpus TL English, SL Italian, subject Lavinia, cycle 1, sequence 7

**Access:** Free

**Project Controller:** ESF

**Keys:**

Name:	Value:
SOUND_LINKING	??

Content Participants Files

Transcription Files

Name:	Format:	Remark:	Browse:
1. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1.cha	text/chat	Was generated from ESF transcript	Browse
2. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1tr	text/esf	Original ESF transcript	Browse
3.	unknown		Browse

Media Files

Name:	Format:	Start:	Duration:	Quality:	Remark:	Browse:
1. /data/corpora/esf_conv/data/english/longitudinal/italian/liela/liela17k.1.sd	audio/eps	unknown	unknown			Browse
2.	unknown					Browse
3.	unknown					Browse

Label Files

Name:	Format:	Start:	Duration:	Remark:	Browse:
1.	unknown				Browse
2.	unknown				Browse
3.	unknown				Browse

Add a transcription file...

Add a media file...

Add a label file...

### Browser Action

Exit

About

Options

Status: Ok

### History

MPI corpora (UNIX)  
MPI Corpora (HTTP)

### Meta Descriptions

- Xml mpi-unix.xml
  - ⊖ MPI corpora
    - ESF
    - ⊖ LAC-corpus Language & Cognition
      - africa
      - ⊖ australia
      - fareast
      - ⊖ mesoamerica
        - ⊖ lac-mesoamerica-yucatec.xml
        - ⊖ tzeltal
          - ⊖ Penny Brown
            - ⓘ Brown-TZELLG.html
            - ⓘ **Brown-DATATYPE.html**
            - ⓘ Brown-REFERENC.html
            - ⊖ Tzeltal Dictionaries
              - Tzeltal Grammar
              - 1971-73 fieldtrip
              - 1980 fieldtrip
              - 1990-2000 fieldtrips
          - zapotec
          - ⊖ lac-mesoamerica-olutec.xml
        - mideast
        - northamerica
        - oceania

Root URL

List

Add

Remove

Remove All

### Description

This file gives information about references

### Info/Content

```
<HTML>
<HEAD>
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;
charset=windows-1252">
<META NAME="Generator" CONTENT="Microsoft Word 97">
<TITLE>datatype</TITLE>
</HEAD>
<BODY>

<B><FONT SIZE=4><P>Info file describing types of data in the
database:</P>
</B></FONT><FONT SIZE=2><P>&nbsp;</P>
<P>The Tzeltal data in this corpus was collected during three major
periods of fieldwork. The first was P.Brown's PhD dissertation
fieldwork in 1971-73; data consist of audiorecordings, transcripts,
and fieldnotes. The second was a summer of fieldwork in 1980, by
P.Brown and S.C. Levinson, in preparation for (and funded by) a
research project at the Australian National University. The data were
both audiorecorded and (some) 8mm film recorded (and later copied
to VHS video). The third period of research is sponsored by the Max
Planck Institute; this started in 1990 and continues through the
present; part of this is in connection with the MPI Space project and
was conducted in collaboration with S.C. Levinson.</P>
<P>&nbsp;</P>
<P>Data collected during the first two periods consists of (i)
naturally-occurring Tzeltal conversation; (ii) Tzeltal speech in public
situations (political speeches, speech at markets, fiestas, church
sermons, court cases), (iii) linguistic elicitation, (iv) elicited songs and
narratives, (v) fieldnotes. The data collected during the decade
1990-2000 includes some of the same kinds of data as above, but
focusses in addition on (vi) spatial language, including
naturally-occurring Tzeltal spatial descriptions in everyday contexts,
in the household, on the trails, in the fields, as well as examples
systematically elicited from both adults and children in response to
...
</P></B></FONT></HTML>
```

# Accessing Meta-Description Files

- Does not require, but can use, a remote server
- Transparent access to remote MDFs via HTTP
- For direct tool access to remote LRs, tools should use suitable protocol

