

Browsable Corpus

The Basics

Internal WS presentation

BC against chaos

- Growing mass of multi-media Language Resources
- Organization based on implicit meta-data by storing in directory structures
(All recordings of Tamil speakers in one directory)
- Only people that already know their way around the data can access the resources

Browsable Corpus Concept

Organize the LRs by:

- Tagging the LRs with meta-data
- Create browsable hierarchies of corpora to complement directory hierarchies of resources
- Hierarchies are based on meta-data
- Create tools that use these hierarchies for browsing, searching and accessing the LRs

Meta-Description Files (MDF)

- Session Meta-Description File; describes a session = bundle of connected LRs
(Transcript + media + annotation files)
- Corpus Meta-Description File; linked to a logical part of a corpus (e.g. all sessions with female speakers)

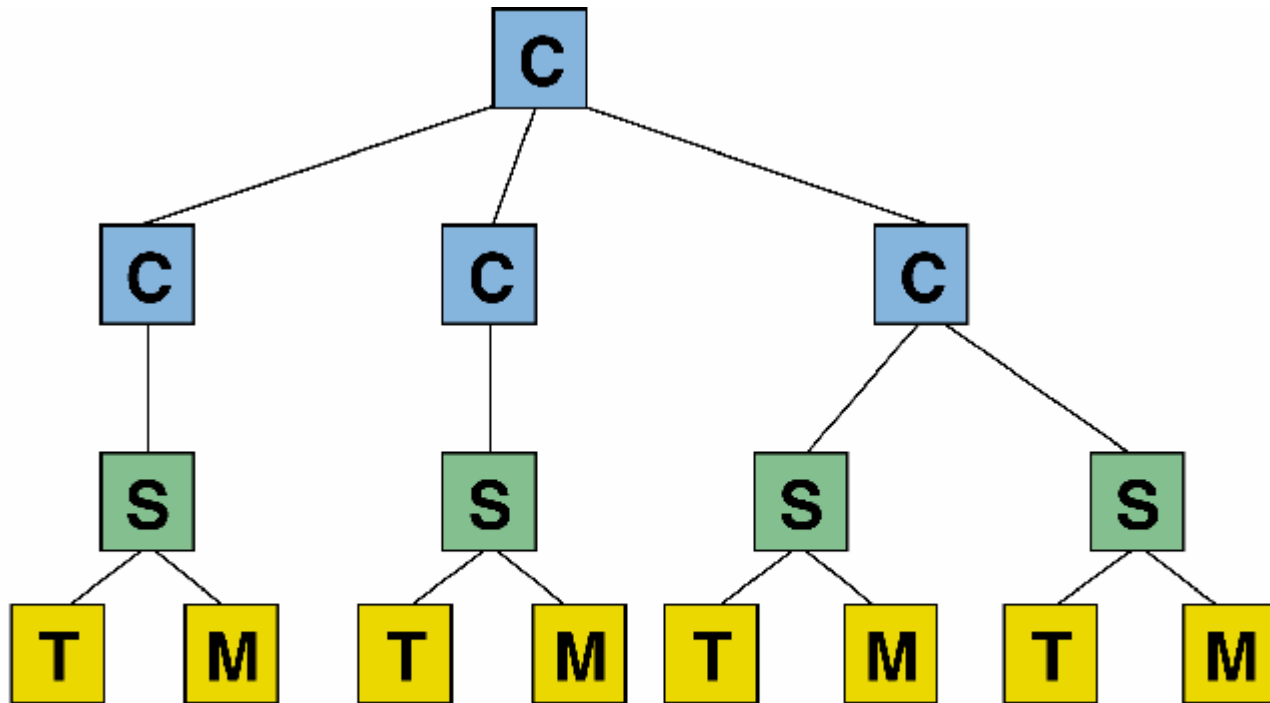
Session Meta-Description Files

- Session MDFs contain meta-data and tag LRs
- The meta-data characterizes the LR's content and form
- In the session MDF there are pointers to the LRs. References are URLs

Corpus Meta-Description Files

- Corpus MDF describe (a meaningful) part of a corpus
- A corpus MDF can contain meta-data that pertains to all LRs of the particular corpus.
- Corpus MDFs point to other Corpus MDFs or to Session MDFs to form a hierarchical structure. References are URLs

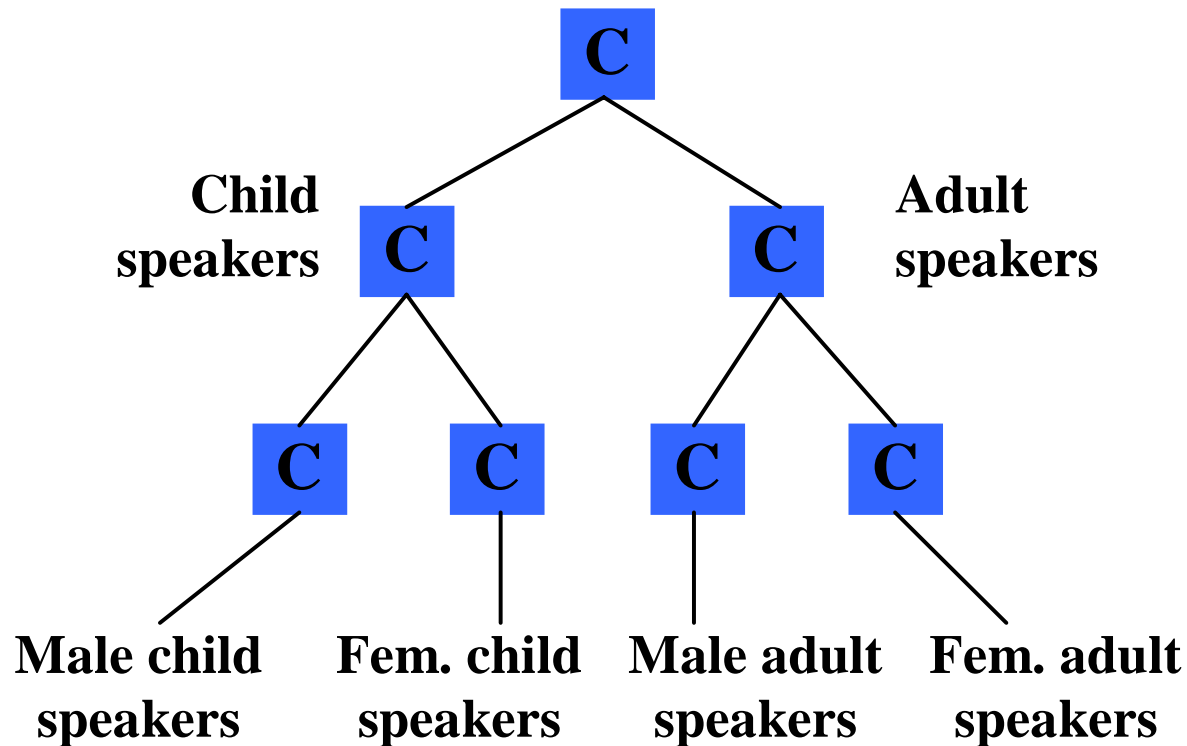
MDF Hierarchy



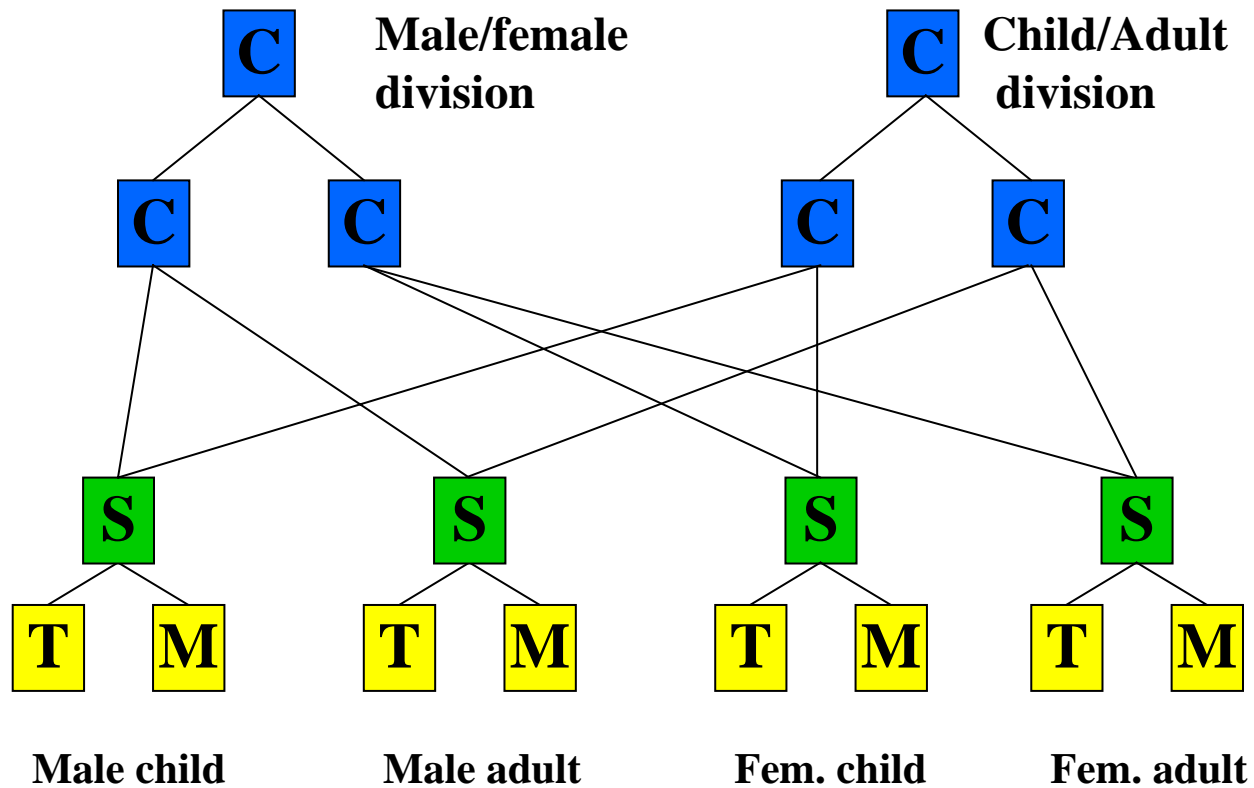
C = corpus MDF file
S = session MDF file

T = transcript file
M = media file

Trivial Corpus Hierarchy



Multiple Hierarchies for same LR



Meta-Description File Format

XML is used for the MDFs because:

- Its power to express structure
- Structure can be defined in DTD
- Use growing range of XML capable software
- Generally accepted as an open interchange format

Meta Description File Structure

Four main sub structures

- General (name, project controller, access,...)
- Content (language used, monologue/dial./)
- Participants (biographical info)
- Files/Resources (info on format and location of the LRs)

See Word Document
Lieran17k.1.xml

See Word Document
metatranscript.doc

From DTD to XML Schema

DTD offers insufficient power to define format of attributes.

- DATE specification
- Time specification
- File format specification

Need standard XML parser that supports XML Schema

To support the BC scheme we need Tools

- Editors to generate the MDFs
- Browsers to navigate the universe of linked MDFs
- *Search engines that search the BC hierarchy and find specified sessions.*
- *Tools to administrate MDFs; allow for changes in the MDF format etc.*

Meta Description Editor

- Offers a user friendly GUI for generating meta-description files
- Should offer the user guidelines for using the meta-data vocabulary
- *Should restrict the input of values where needed and check consistency if possible*

Failed Objectives BC Editor

- Self adapting to DTD
- Description of the editors of UI in a special specification language analogous to UIML for Motif

As a result the DTD and UI are hardwired in the editors Java code.

File Edit Help Structure

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/esp	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

BC Browser Tool

- Independent application although it will be able to run as an applet within Netscape
- Can use HTTP to access remote MDFs
- Coupled to tools that work with individual LRs. If the resources are remote, these tools must also use an appropriate protocol such as HTTP

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

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
```

Root URL

List

Add

Remove

Remove All

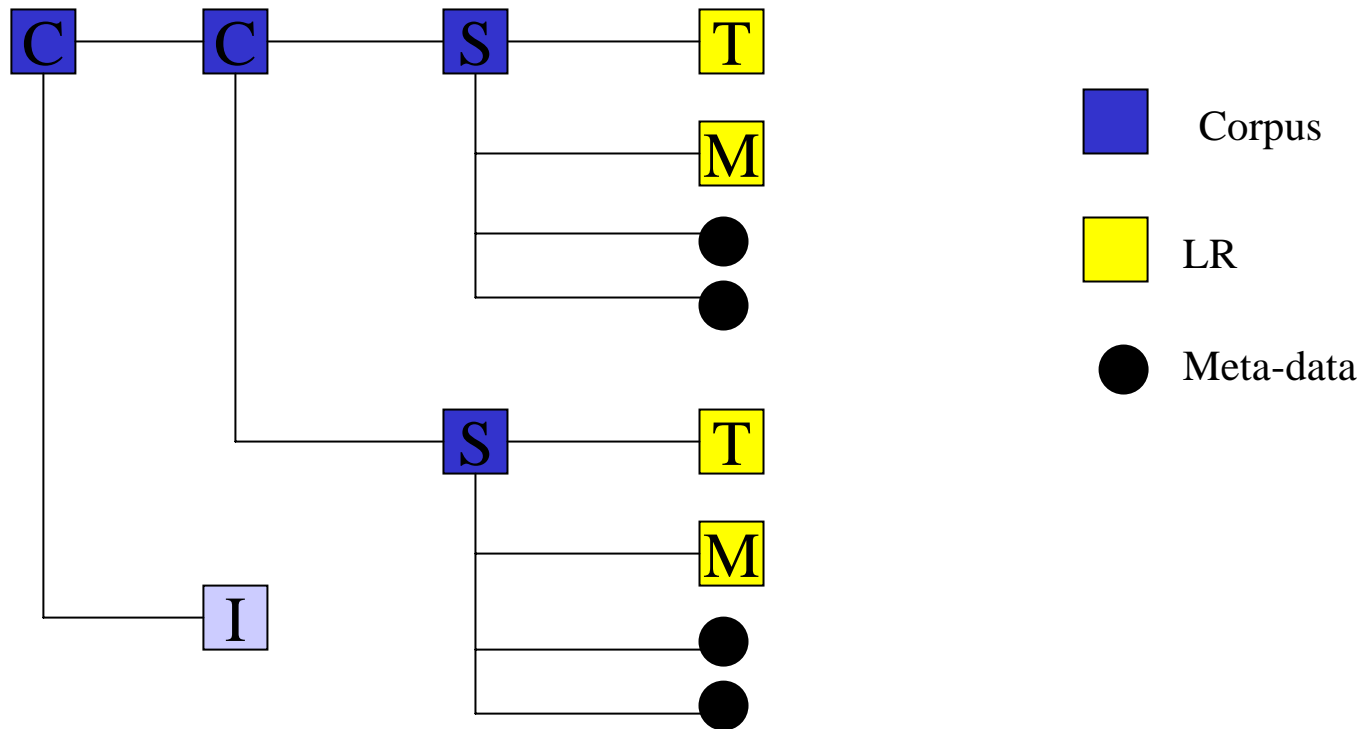
Browser Functioning

- Browser reads XML meta-description file and parses it into a DOM structure
- For every relevant Element in the DOM structure an appropriate member from the HierarchicalNode family is created.
- The HierarchicalNode family contains all relevant information.
- The HierarchicalNode is displayed in the browser window and reacts to user actions

HierarchicalNode family

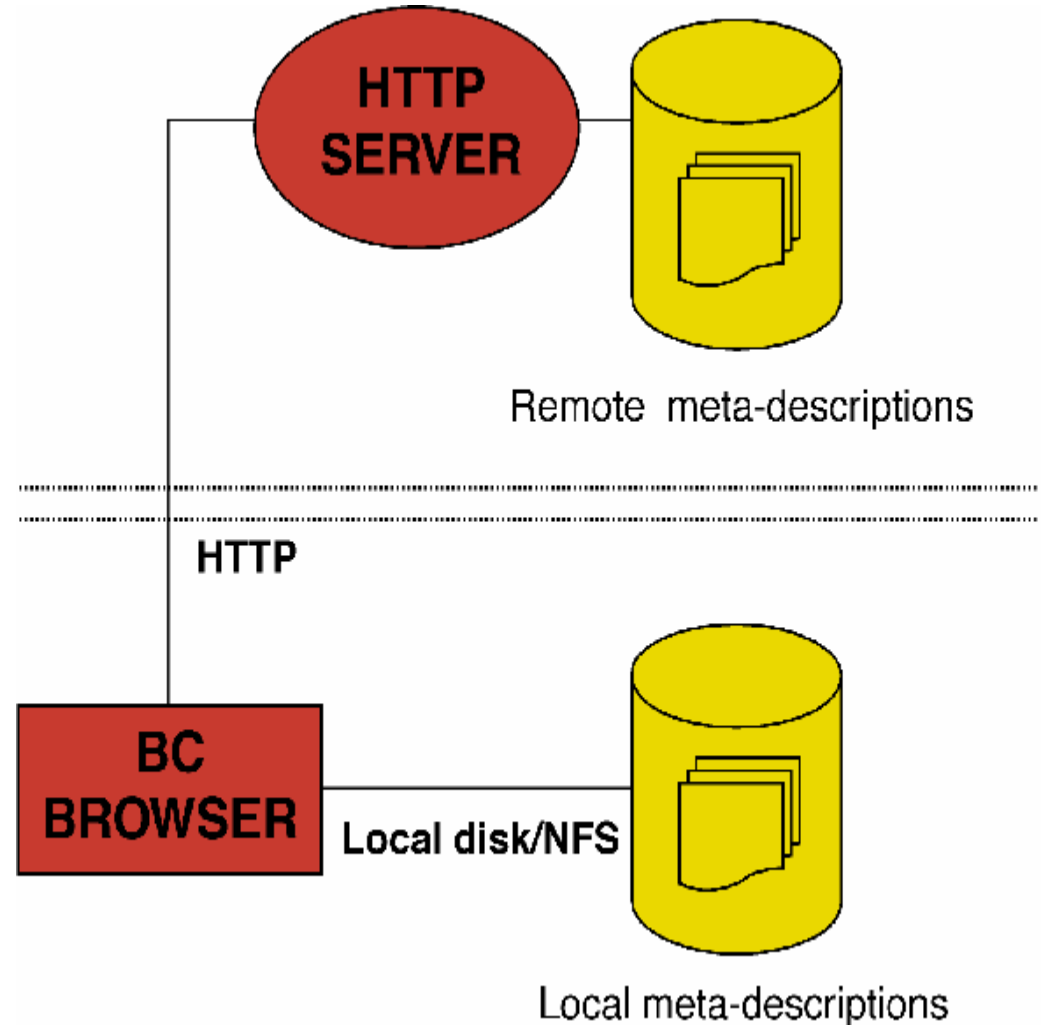
- CorpusNode sub family models Corpus & Session nodes
- DataNode sub family models meta-data ParticipantNode, ProjectControlerNode
- LRNode sub family models language resources
- InfoFile models legacy meta-data information files

HierarchicalNode Display



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



Challenges

- Integrate BC in the data processing standards used at the MPI
- MDF's should be central in MPI corpora data management
- MDF's should form the basic information unit for our MM archive.

Browsable Corpus & EUDICO

Just friends?

BC Universe is independent from EUDICO

- Is also usable for use on home PC without Internet access
- Does not require an EUDICO server to access LR's
- Offers local stand alone tools to access LR's (besides EUDICO)

Integration Aims

- Use the same Browser for EUDICO and BC; code sharing!
- Enable links from within BC corpora to Eudico references and vice versa
- Offer consistent view on both BC and corpora managed by an EUDICO server

Communalities

- Browsing linked MDF's <-> browsing
CLOM hierarchy
- HierarchicalNodes <-> Eudico
TreeViewables

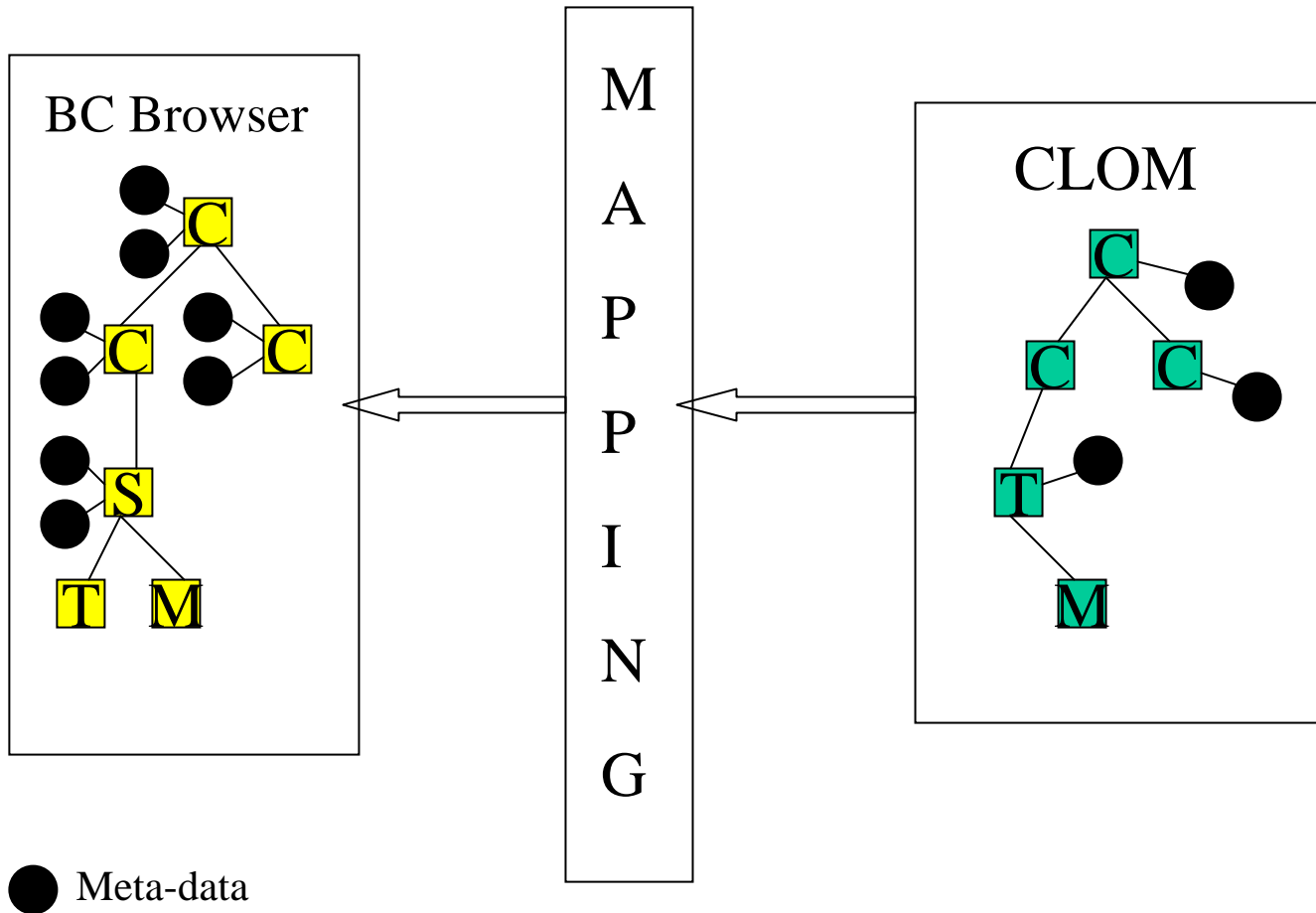
Exploit these to come to tight integration

Integration Procedure

- There is a mapping between XML elements and HierarchicalNode members
- Make a similar mapping between EUDICO TreeViewables and HierarchicalNode members where possible
- Create HierarchicalNode meta-data members on the basis of information from the EUDICO TreeViewables were needed.

Thus the same HierarchicalNode family can be used.

Mapping CLOM on BC



Enhancements of the CLOM

- Introduce Session concept in the CLOM
- Introduce Meta-Data concept in the CLOM?
- Introduce CLOM URL(ish) specifications:
 eudico://ourhost/Gesture_corpus/foo11
- Adopt the mime-type scheme to identify types of LR

Browsable Corpus Future

MPI

- Make the BC Editor and Browser accepted by our researchers
- Anchor the BC concept in the data processing procedures; digitisation, automatic archiving etc.
- Add access control mechanism for LR's
- make corpus level meta-description editor
- Integrate BC Browser with used existing tools like MED, MT, Shoebox, Praat

ISLE 1

Make applet versions of BC Editor & Browser

- Easy distribution & update
- HTML rendering function for BC Browser

Problem will be security

- Back-office solution for storing MDF's at the MPI

ISLE 2

- o Modify the tools or an ISLE version of the tools to keep up with the emerging standard

Problems:

- Lots of work
- Maintain different versions

The more reason to look for modular and configurable solutions

ISLE 3

Tried to make a configurable solution for the BC Editor. But:

- The UI is not a direct map of the DTD
- Need extra information about lay-out

Adapting will be labour intensive

ISLE 4

For the BC Browser more possibilities

No problem supporting several meta-description standards or extensions by one browser

- Should be well documented and conforming
- Should be same implementation mechanism

COREX

COREX 1

For the preliminary version

- Use tested version of BC / EUDICO integration
- Make indirect access mechanism for resources on removable disks
- Nice to have local tool functionality for Praat

COREX 2

- For final version use tight BC /EUDICO integration
- Make meta-data search

