Agenda CMDI Workshop

9.15	Welcome
9.30	Introduction to metadata and the CLARIN Metadata Infrastructure (CMDI)
10.15	Coffee
10.30	Use of ISOCat within CMDI
11.00	The CMDI Component Registry and CMDI Component Editor
11.45	ARBIL, the CMDI metadata editor
12.30	Lunch
13.00	Standard Metadata Components and Profiles available from the registry
13.30	Metadata creation scenarios and try it your self opportunity
15.00	Coffee
15.15	Metadata creation scenario's and try it your self opportunity, continued
16.00	Further hands on practice with guidance
17.00	End

CMDI

CLARIN Component Metadata Infrastructure

Daan Broeder et al.
Max-Planck Institute for Psycholinguistics

CLARIN metadata project background

- CLARIN EU WP2 since 2007 investigated and creates (prototypical) solutions for:
  - Common AAI infrastructure
  - Single system of persistent identifiers (PIDs) for resources
  - Common metadata domain - CMDI
  - …
- CMDI is being developed by CLARIN partners: Austrian Academy, IDS, MPI for Psyl, Sprakbanken Univ. Gothenborg,
- National CLARIN projects: CLARIN-NL, (D-SPIN) CLARIN-DE have committed resources to work with CMDI
  - CLARIN NL metadata project has been testing the CMDI basics

Metadata in General

- Data about Data
- Structured Data about Data
  - Not a prose description (although that can be a part)
  - … but keyword/value type of data:

    Name = “myresource”, Title = “mybook”
- *Internet*: Machine readable Data about Data
  - XML format.

Used for:

- Resource discovery / accessing
- Management
- …

Dublin Core (DC) Metadata Set

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

DC Example

Qualifiers either specify:

- encoding scheme
- refinement

DC.Title = “My first book”

DC.Title /Alternative = “My last book”

DC.Creator = “L. Smith”

DC.Subject /LCSH = “Building”

DC.Description/Abstract = “…….”

DC.Language/ ISO639-2 = “eng”

Metadata for Language Resources I

- Resource types:
  - Video, audio, pictures, annotations, primary texts, notes, grammars, lexica, …
- Different levels of description (granularity):
  - complete corpora e.g. Brown Corpus.
  - sub corpora or corpus components: e.g. all Flemish recordings in the Spoken Corpus Dutch with all the transcriptions
  - (recording) sessions: e.g. the recording of a dialogue (sound file + transcript)
  - individual resources: e.g. a text file

Metadata for Language Resources II

- Metadata was/is often embedded in annotations
  - CHAT format
  - TEI
- Advantage of splitting this:
  - Independent formats allowing combinations as IMDI metadata with CHAT annotations
  - Keep several versions for different tools
- … but danger of inconsistencies

Current Metadata Situation

Fragmented landscape

- Metadata sets, schema & infrastructures in our domain:
  - IMDI, OLAC/DCMI, TEI
- Problems with current solutions:
  - Inflexible: too many (IMDI) or too few (OLAC) metadata elements
  - Limited interoperability (both semantic and functional)
  - Problematic (unfamiliar) terminology for some sub-communities.
  - Limited support for LT tool & services descriptions

Common metadata domain

Why a common metadata domain:

- Finding and sharing resources housed at all archives & repositories participating in CLARIN
- Specify distributed heterogeneous collections of LRs and processing these collections
- In general, a common metadata domain helps bringing along a single domain of LRs

Metadata Components

CLARIN chose for a component approach: CMDI

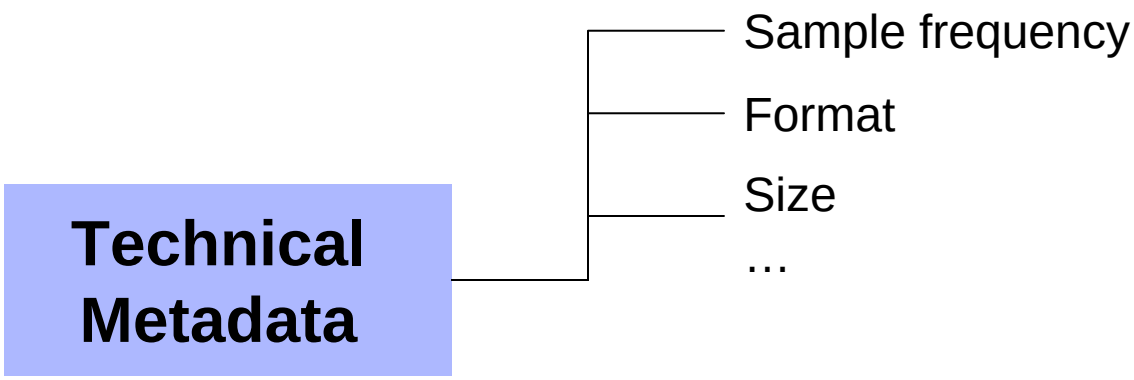
- NOT a single new metadata schema
- but rather allow coexistence of many (community/researcher) defined schemas
- with explicit semantics for interoperability

How does this work?

- Components are bundles of related metadata elements that describe an aspect of the resource
- A complete description of a resource may require several components.
- Components may use and contain other components
- Components should be designed for reusability

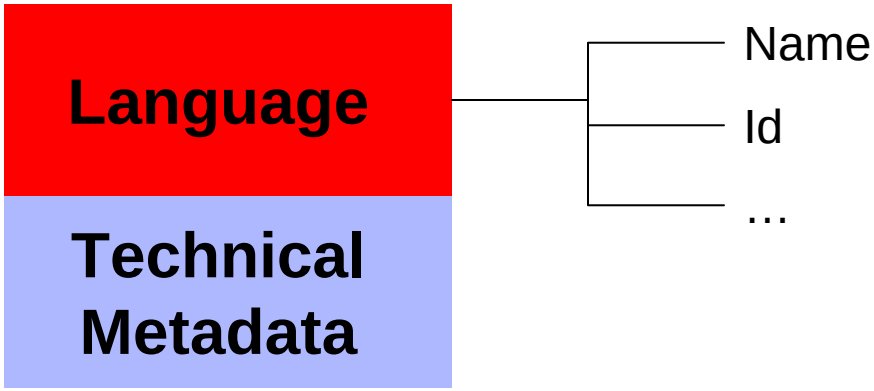
Metadata Components

Lets describe a speech recording



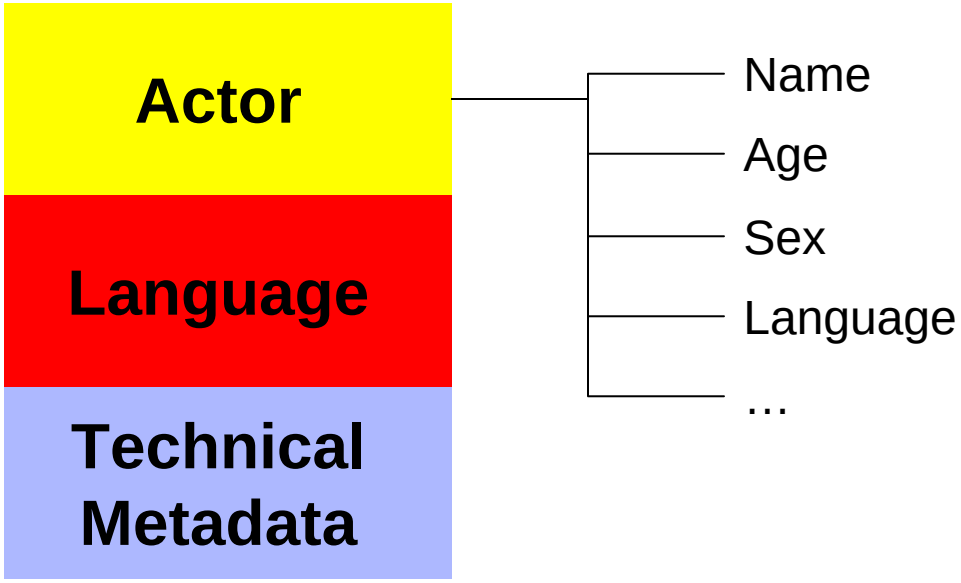
Metadata Components

Lets describe a speech recording

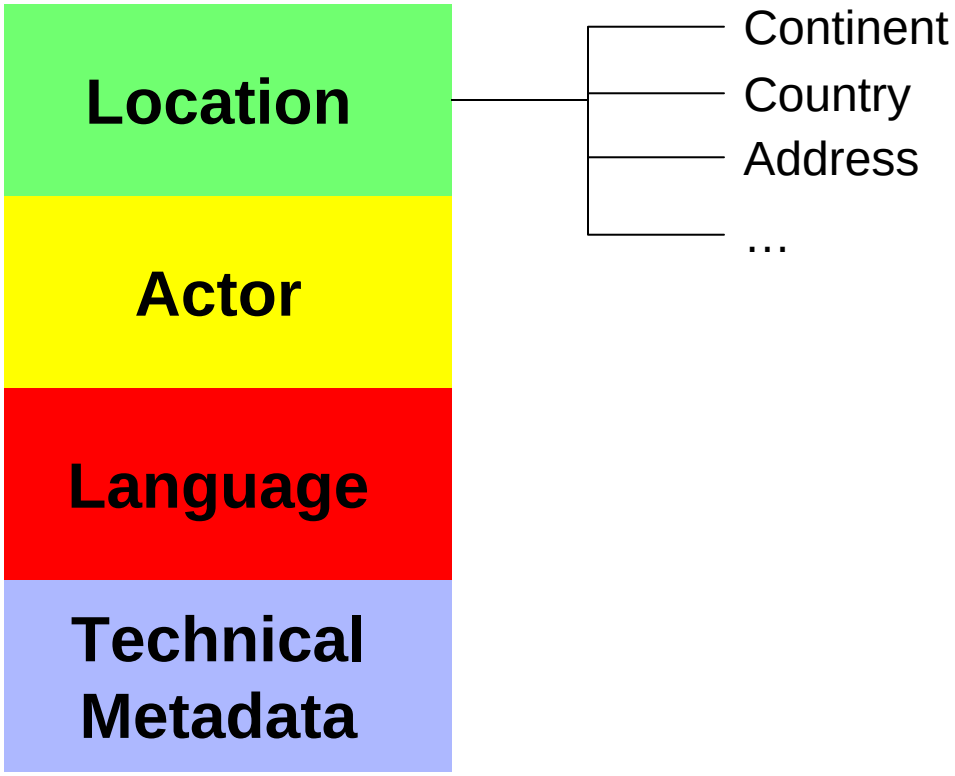


Metadata Components

Lets describe a speech recording

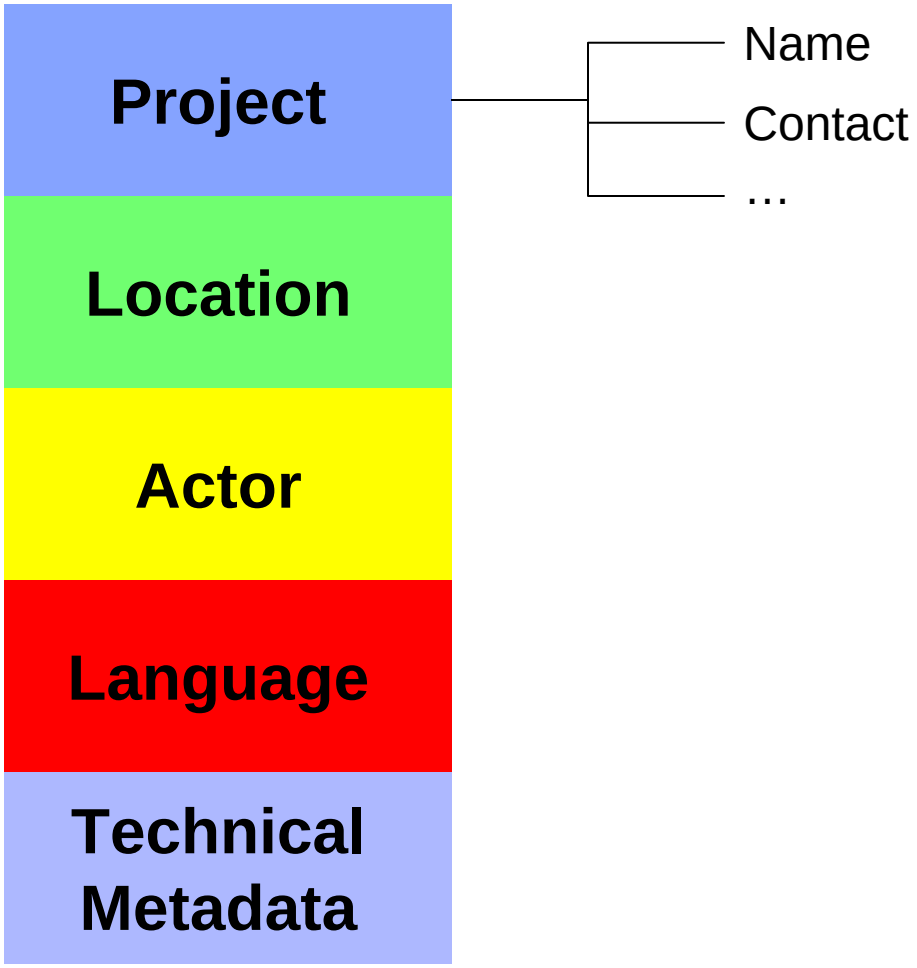


Metadata Components



Lets describe a speech recording

Metadata Components



Lets describe a speech recording

Metadata Components

Project

Location

Actor

Language

Technical Metadata

Metadata profile

Lets describe a speech recording

Metadata schema

Metadata Components

Project

Location

Actor

Language

Technical Metadata

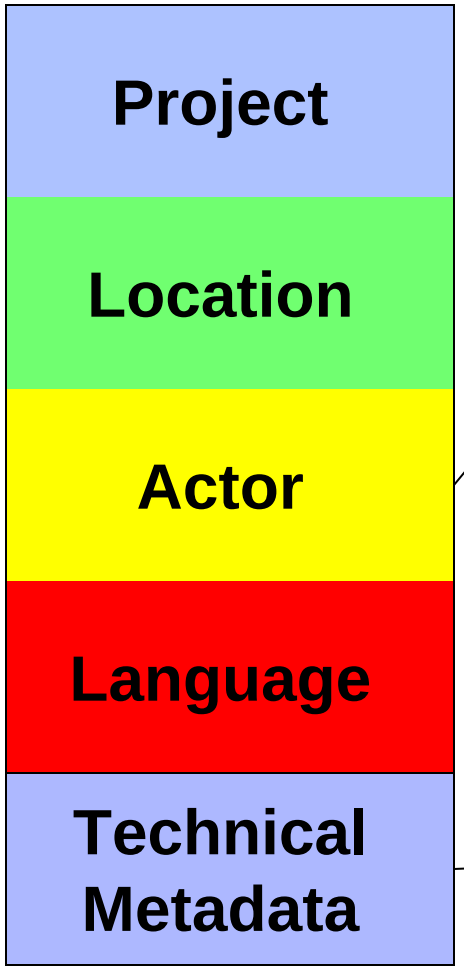
Metadata profile

Lets describe a speech recording

Metadata schema

Metadata description

Metadata Components



Metadata profile

Profile definition XML

Component definition XML

Lets describe a speech recording

Metadata schema

W3C XML Schema

Metadata description

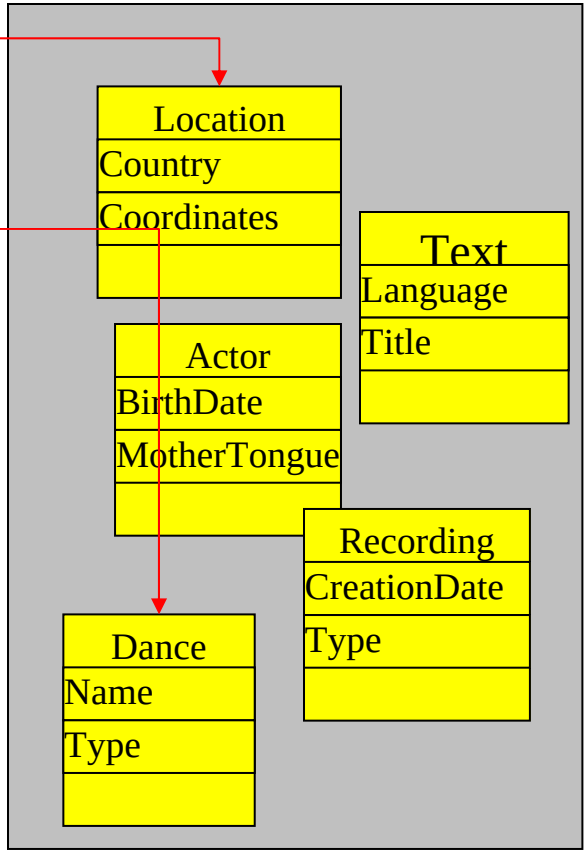
XML File

CMDI Component Reuse

User selects appropriate components to create a new metadata profile or an existing profile

user

Component registry



Selecting metadata components from the registry

Concept registries

- Basically a list with concepts and their descriptions where every concept has a unique identifier.
- Some have a complicated structure and are associated with elaborate (administrative) processes to determine the status and acceptation of concepts in the registry. e.g. ISO-DCR.
- others are static and simple lists of concepts and descriptions e.g. DCTERMS

CMDI Explicit Semantics

Component registry

User selects appropriate components to create a new metadata profile or an existing profile

user

Semantic interoperability **partly** solved via references to ISO DCR or other registry

Location
Country
Coordinates

Text
Language
Title

Actor
BirthDate
MotherTongue

Recording
CreationDate
Type

Dance
Name
Type

ISOcat concept registry

Country dcr:1001
Language dcr:1002
BirthDate dcr:1000

DCMI concept registry

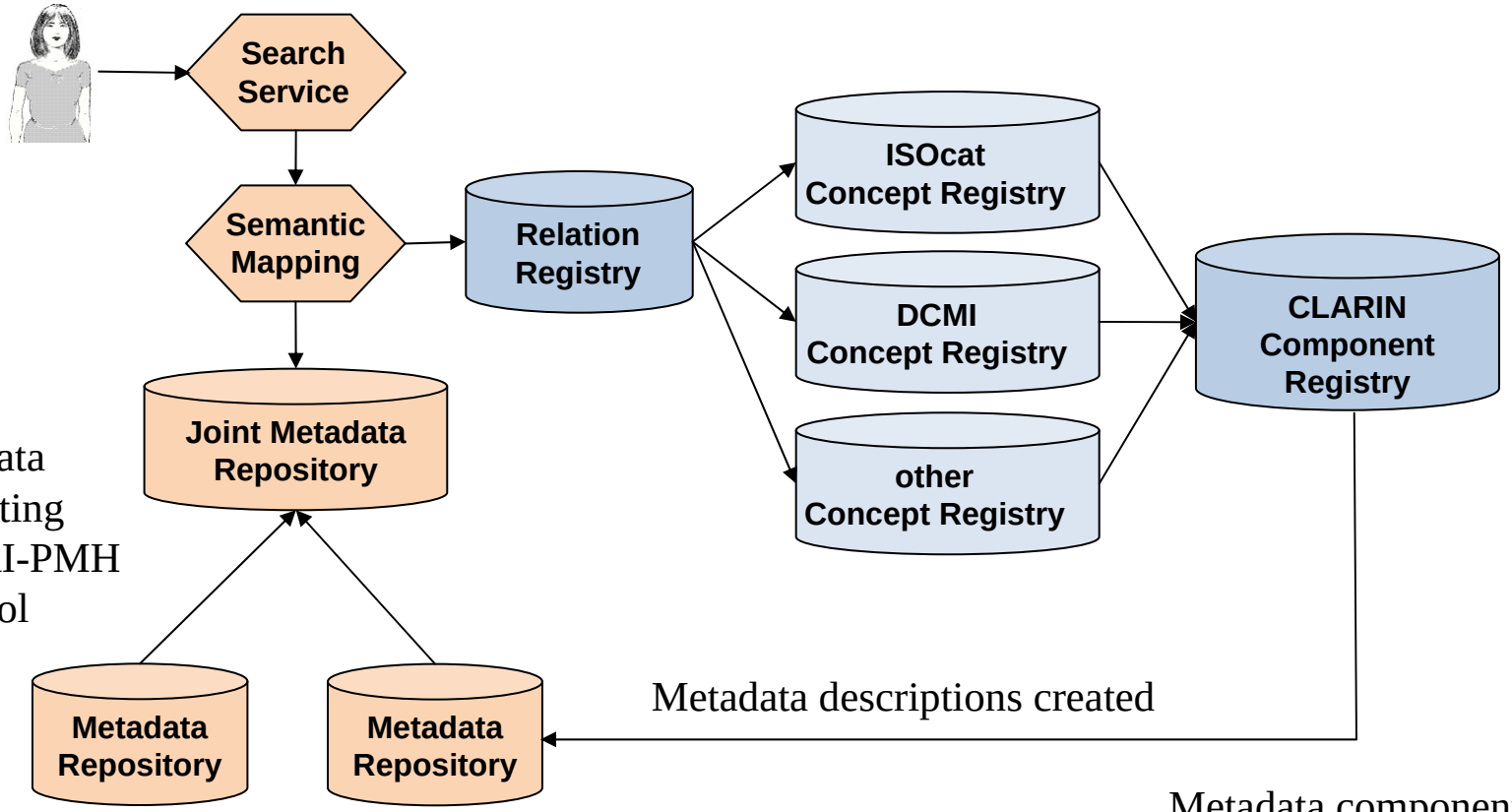
Title: dc:title

Selecting metadata components from the registry

CMDI Metadata Live-cycle

Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry

Create metadata schema from selection of existing components. Allow creation of new components if they have references to ISOcat



CMDI Architecture I

- The CMDI takes an archivist or “production” first viewpoint
  - Prioritize that the metadata can be of good quality: consistent, coherent, correctly linked to the concept registries
  - The consumer side can be more “experimental” and diverse.
  - Many MD exploitation “stacks” or consumers applications can work in parallel on the same metadata

CMDI Architecture II

ISO
TDG

user

MD
Catalog

Virtual
Collection
Registry

Metadata
modeler

ISO-Cat
DCR

MD
Services

Semantic
mapping
Services

MD Comp.
Editor

MD Comp.
Registry

External
agents

Relation
Registry

MD Editor.

MD
Creator

CLARIN
Joint MD
Repository

OAI-PMH
Service
Provider

OAI-PMH
Data
provider

Local MD
Repository

Current CMDI status I

- ISO-DCR: 218 metadata concepts
- CMDI component registry: 135 components, 19 profiles

Produced & inspired by:

- Deconstructing existing metadata schema IMDI, OLAC, TEI
- Considering requirements of other CLARIN activities like profile matching
- CLARIN NL metadata project tested the CMDI model and delivered components and profiles for the resources in two major Dutch Language Resource centers

Current CMDI status II

Operational or test phase:

- ISOCat DCR
- Component registry & editor
- ARBIL metadata editor

Still working on:

- *Joint Metadata Repository, Metadata Catalog, Semantic Mapping, Relation Registry*

Expect a usable first version in third quarter 2010

Thank you for your attention

CLARIN has received funding from
the European Community's Seventh Framework Programme
under grant agreement n° 212230