CMDI: a Component Metadata Infrastructure

Daan Broeder

TLA - Max-Planck Institute for Psycholinguistics
Why CMDI

Problems with existing solutions:

- Inflexible: too many (IMDI) or too few (OLAC) metadata elements
- Problematic & unfamiliar terminology for some communities
- Limited interoperability (both semantic and functional)
CLARIN project – one of its goals is a joint metadata domain for LRs

- Not a new metadata set that should supersede all others
- … but rather an environment supporting different metadata sets
- where new interoperable metadata schema can be created to describe new data types (or old data types for new purposes)

How to do this?

- Use reusable Metadata Components
- with well defined syntax
- … and explicit semantics
Let's describe a speech recording

Technical Metadata

- Sample frequency
- Format
- Size
  ...

Metadata Components
Let’s describe a speech recording
Let's describe a speech recording
Let’s describe a speech recording
Let’s describe a speech recording
Let’s describe a speech recording

Profile definition XML

Metadata schema
W3C XML Schema

Component definition XML

Metadata description
XML File
Recursive model

- Recursive Component model
- Components can contain other components
- Enhances reusability
Semantic interoperability partly solved via references to ISO DCR or other concept registry.

ISOcat or ISO DCR
- implementation of ISO-12620 standard for data categories
- under control of the linguistic community ISO TC37
- Metadata is just one of the seven “thematic domains”
Collections & metadata

Collection of tightly coupled resources

Collection of resources

Hierarchy of sub-collections
Metadata Actors & Entities

- **Metadata Users** use metadata to find or resources
  - Product: suitable resource
- **Metadata Creators** create metadata to describe resources
  - Product: metadata description of a resource
- **Metadata Curator** Updates metadata description for maintenance
  - Product: metadata description of a resource
- **Metadata modelers** create metadata schema and/or terminology
  - Product: metadata schema with explicit terminology
- **Metadata repository** facility that for managing metadata descriptions
- **Metadata catalogue** software that allows users to search & browse in metadata
CMDI Metadata life-cycle

- Metadata user
- Metadata curator
- Metadata creator
- Metadata modeler

- ISOcat
- Component registry & editor
- Metadata editor
- Local metadata repository
- Joint metadata repository
- Search & semantic mapping
- Metadata catalogue

- OAI-PMH Service provider
- OAI-PMH Data provider

- DATA

- Relation Registry

- ISOcat component registry & editor

- DATA
There is a ‘huge’ installed base of metadata records available for harvesting: OLAC, IMDI, DC

CMDI component registry was seeded with:
- IMDI profile
- DC/OLAC profile

Specialist IMDI profiles for SignLanguage, Bilingualism, ... will be developed within some CLARIN NL projects

Those communities used to these schemas can work

Others may need assistance to convert their metadata schema
CMDI Status

CMDI Usage

- Different national CLARIN projects: NL, D, DK, …; other national projects: NaLiDa
- Public components: 218, profiles: 49
- Metadata records: 180,000
- Component registry & editor 62 registered users (15 overall active)

Tools

- Production: Component registry & editor, ISOcat, ARBIL, VLO
- Prototypes: Relation Registry, complex metadata search
CMDI contributors

Collaboration on the CMDI implementation

CLARIN EU preparatory phase

- MPI for Psycholinguistics: metadata modeling and editing facilities
- Språkbanken, University of Gothenburg: Joint CLARIN metadata repository
- Austrian Academy: Metadata catalog, metadata & semantic mapping services
- IDS: Virtual Collection Registry

National CLARIN projects: CLARIN NL, CLARIN D,
Thank you for your attention

CLARIN has received funding from the European Community's Seventh Framework Programme under grant agreement n° 212230
Thank you for your attention