CMDI - Component Metadata Infrastructure

CLARIN Conference 2013, Prague

Searching for suitable data & services

Basically 2 ways of resource discovery:

- Searching in content:
  - Google like search, works fine for (un)structured text
  - … smarter approach is using structure and semantics such as (is intended in) CLARIN federated search
- Searching in metadata:
  - Structured data that describes other data

Why CMDI?

Problems with existing LR metadata solutions:

- Inflexible: too many, too specific metadata (IMDI)
- … or too few (DC/OLAC) to general metadata
- Problematic & unfamiliar terminology for some communities
- Limited interoperability (both semantic and functional)

CMDI Origins

CLARIN goal:

- a joint metadata domain for LRs that allows (1) to build a single metadata catalog, (2) gather information on resources (statistics), …

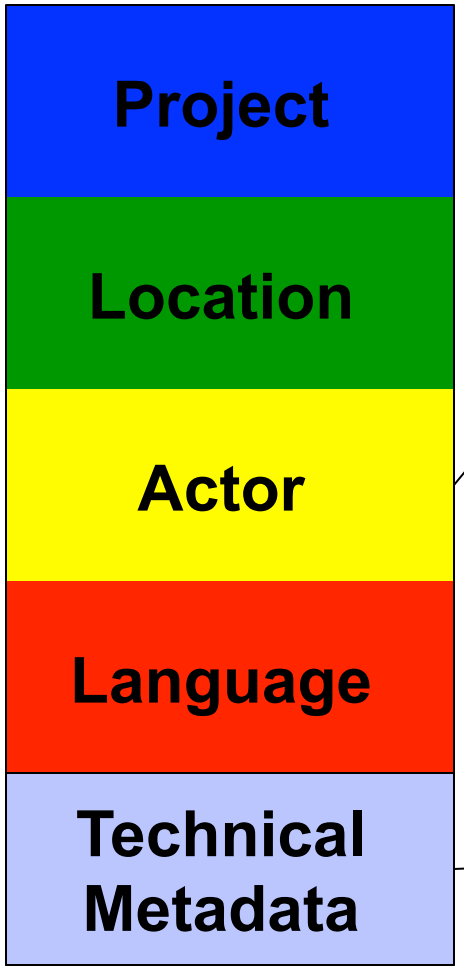
CMDI is

- not a new metadata schema that should supersede all others
- … but rather an environment supporting different metadata schema
- 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 shareable reusable Metadata Components from a central registry to build a metadata schema
- with a well defined syntax
- … and explicit semantics for the metadata elements

Metadata Components



Metadata profile

Profile definition XML

Component definition XML

Let's describe a speech recording

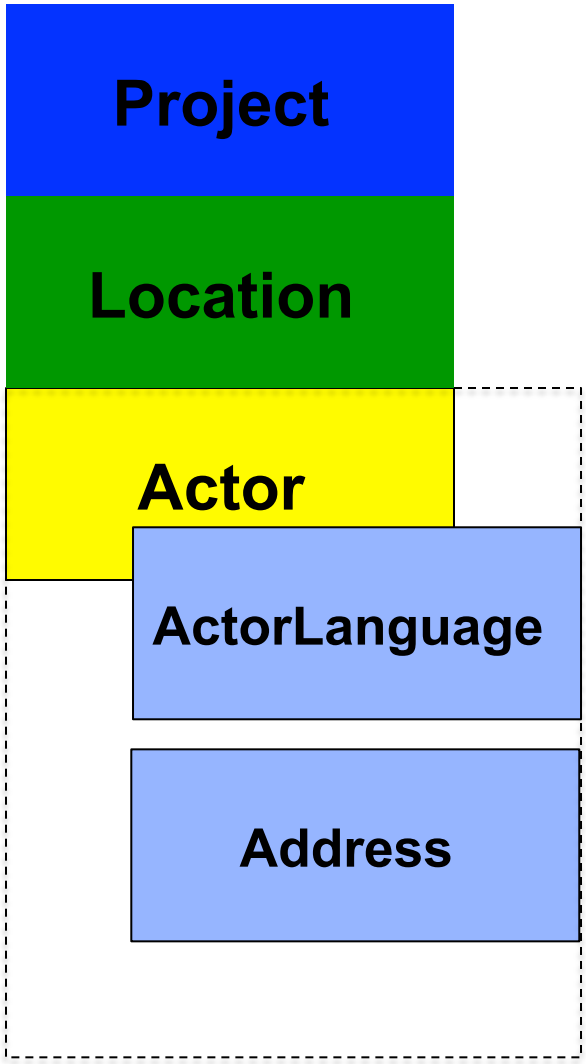
Metadata schema

W3C XML Schema

Metadata description

XML File

Recursive model



- Recursive Component model
- Components can contain other components
- Enhances reusability

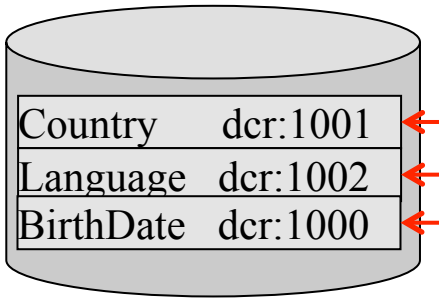
Reusability & Explicit Semantics

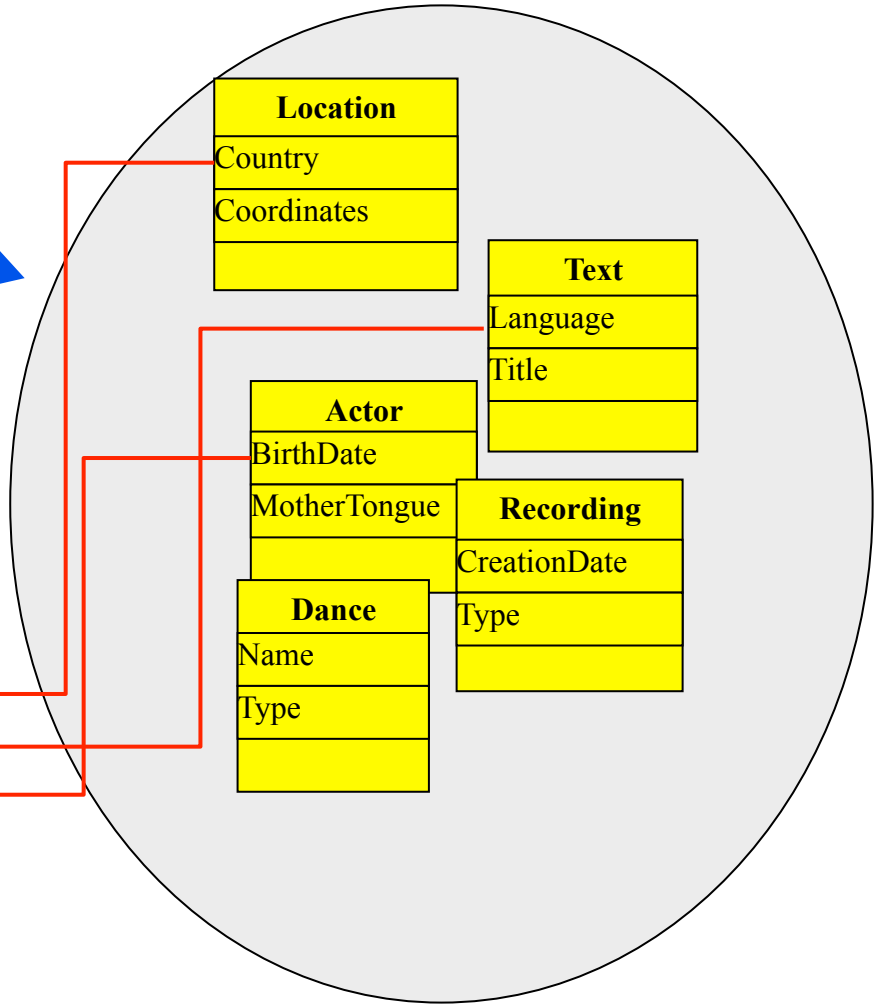
metadata modeler

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

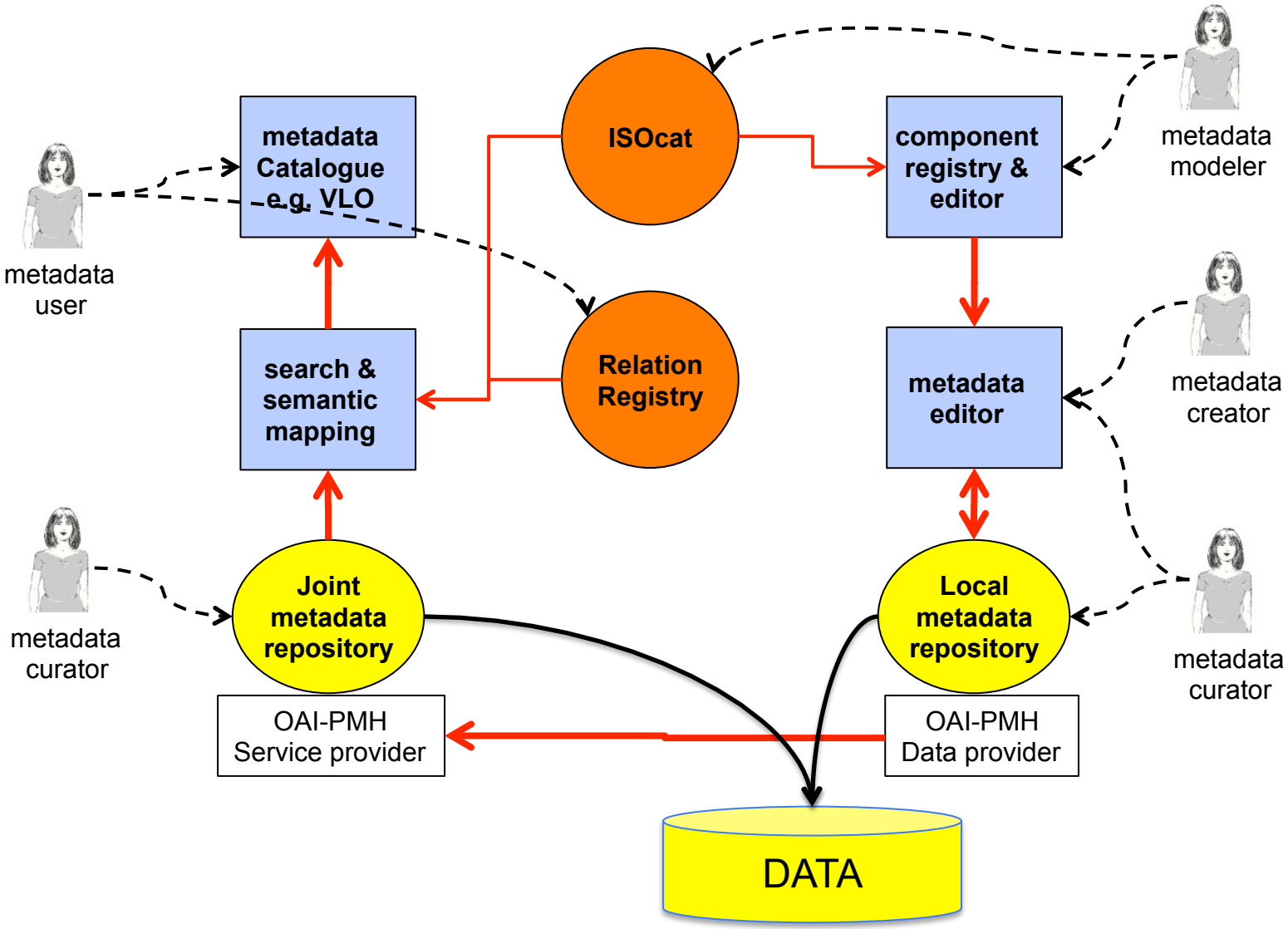
ISOcat concept registry

DCMI concept registry





CMDI Metadata life-cycle

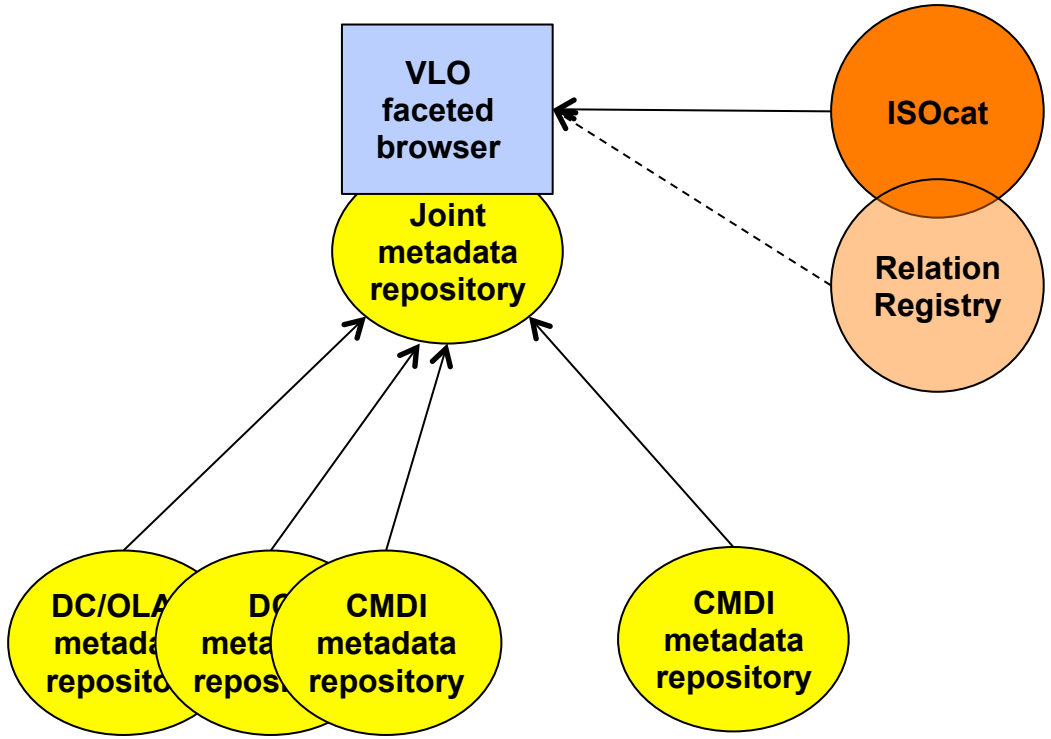


CMDI compatibility

- There is a ‘huge’ installed base of metadata records available for harvesting: OLAC, IMDI, DC
  - We want to embed this base in the CMDI infra
- CMDI component registry was seeded with:
  - IMDI profile
  - DC/OLAC profile
  - TEI header variant
- Specialist IMDI profiles for SignLanguage, Bilingualism, ... were developed within some CLARIN NL projects
- META-SHARE schema now available in CMDI

Current CLARIN MD infra

±500k records



±80 endpoints
16 pure CMDI

Current CMDI Status

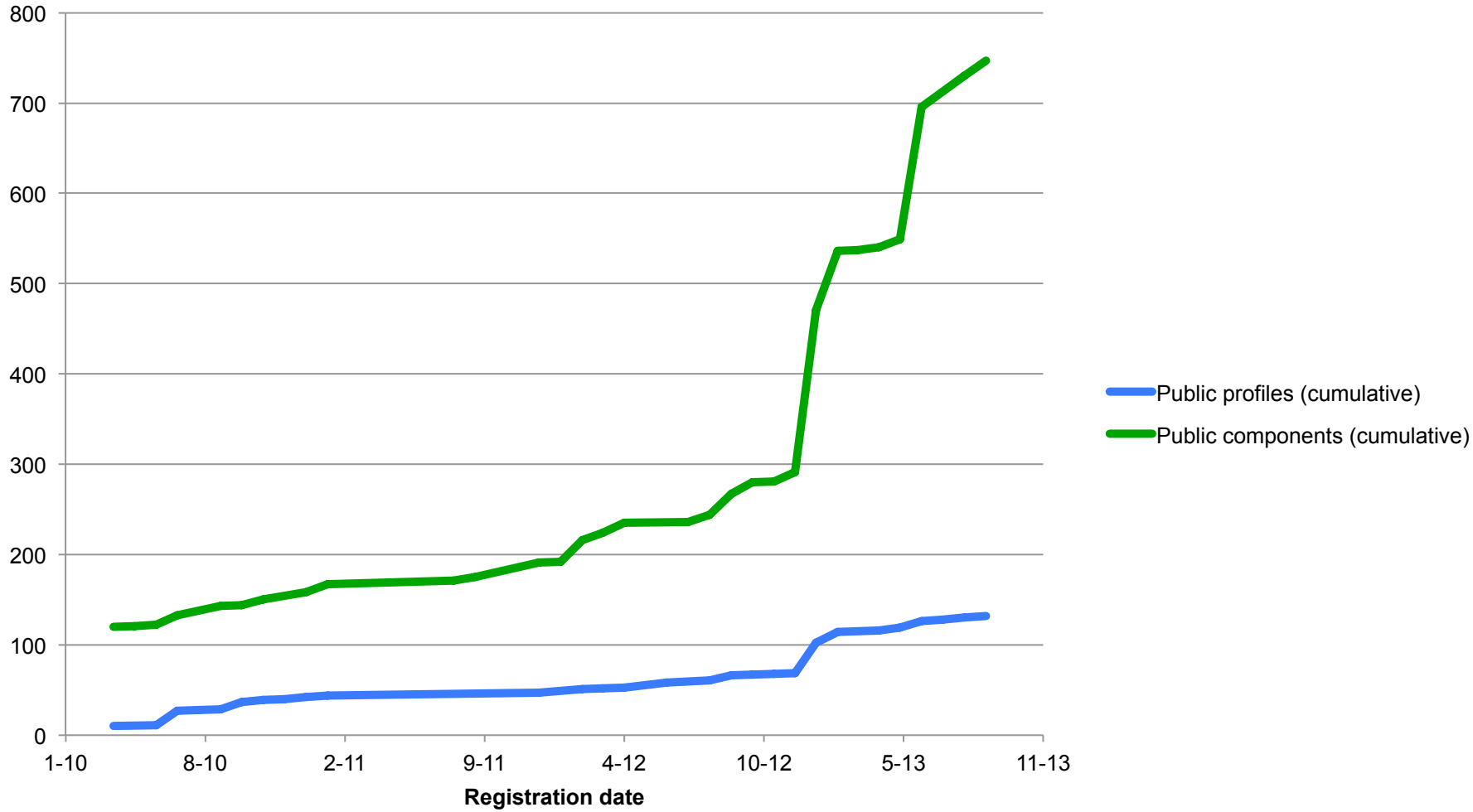
CMDI Usage

- Different national CLARIN projects: NL, D, DK, …; other projects: NaLiDa, RZG proj.
- It is a CLARIN center requirement!
- Public components: ±800, profiles: ±150
- Metadata records: ±500k
- Component registry & editor ±60 registered users (25 overall active)

Tools

- Production: Component registry & editor, ISOcat, ARBIL, VLO, MI CMDI search, SMC Browser
- Prototypes: Relation Registry, YAAS

Registered components and profiles



CMDI challenges

- CMDI Profile and Component proliferation, current ± 150 profiles, 800 components and not all of them very different and unique
  - Why do people insist creating their own?
  - Lack of guidance
    - Too little guidance information in the profiles & components
    - Too difficult to search for suitable profiles & components
  - Too much freedom
    - Better and more central CMDI content management, gentle push to ‘recommended’ profiles
- Granularity: currently collections and individual resources not always distinguishable
- Metadata quality
  - combination of the tool used to browse & search and the content of the harvested metadata records
  - Who is responsible, who can curate/repair, which tools can help us

Organizational & Planning aspects

- CMDI ISO standardization track
  - ISO DIS24622-1 Component Metadata Model submitted on September 9
  - CMD-2 Component Specification Language is now officially a work-item
- CMDI Interoperability workshop June 2013
  - Consultation with META-SHARE colleagues
  - META-SHARE schema available in CMDI
  - Repository of interoperability scripts & tools
- CMDI future workshop, Oct 2013
  - CMDI 1.2 schema, future schema changes
  - How to address component/profile proliferation
  - CMDI tools roadmap
  - Organizing the development work
- CLARIN ERIC committees & Task-forces
  - National metadata quality coordinators
  - CMDI task-force
  - Metadata curation task-force, Standards committee

CMDI progress needs …

- More people and coordination:
  - Work on the tools and share the burden -> CMDI TF
  - Check the metadata quality ->

    National md quality coordinators, Metadata curation TF
  - Better concept definitions, National ISOcat content managers, standardization TF
- Better tool guidance for selecting & sharing profiles and components
  - Versioning of profiles & components (underway)
- Mandatory components: collection, license, SRU ep, …
- Feedback system for metadata quality
  - Partly available in VLO
  - Metadata responsible contacts per center

Thank you