CLARIN
Common Language Resources and Technology Infrastructure

CMDI: an overview

Daan Broeder and many others

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)

CMDI Origins

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

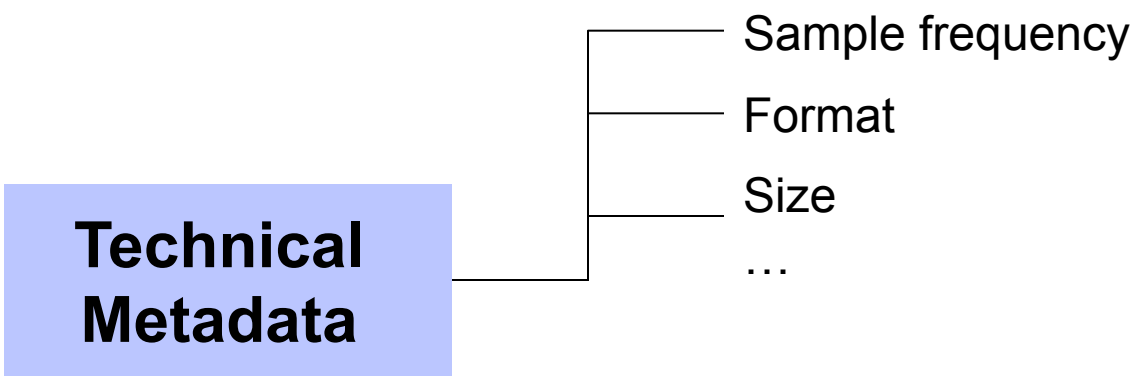
Metadata Components

Let's describe a speech recording

...

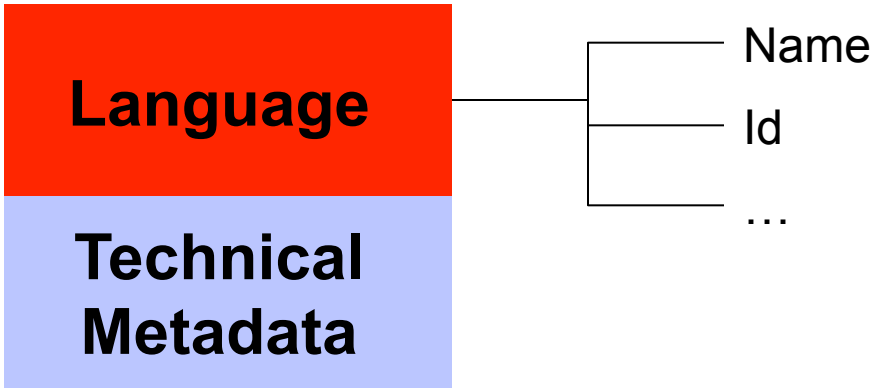
Metadata Components

Let's describe a speech recording



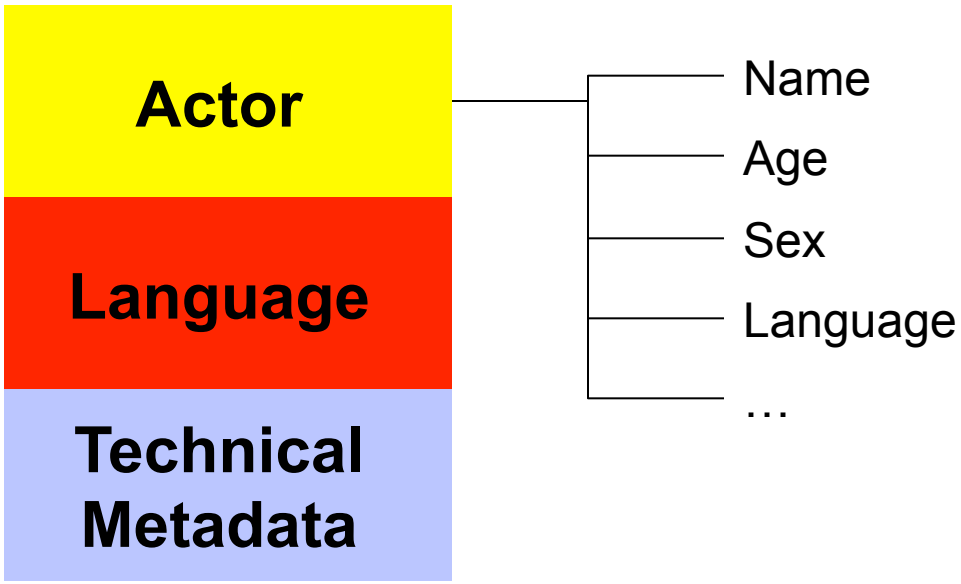
Metadata Components

Let's describe a speech recording

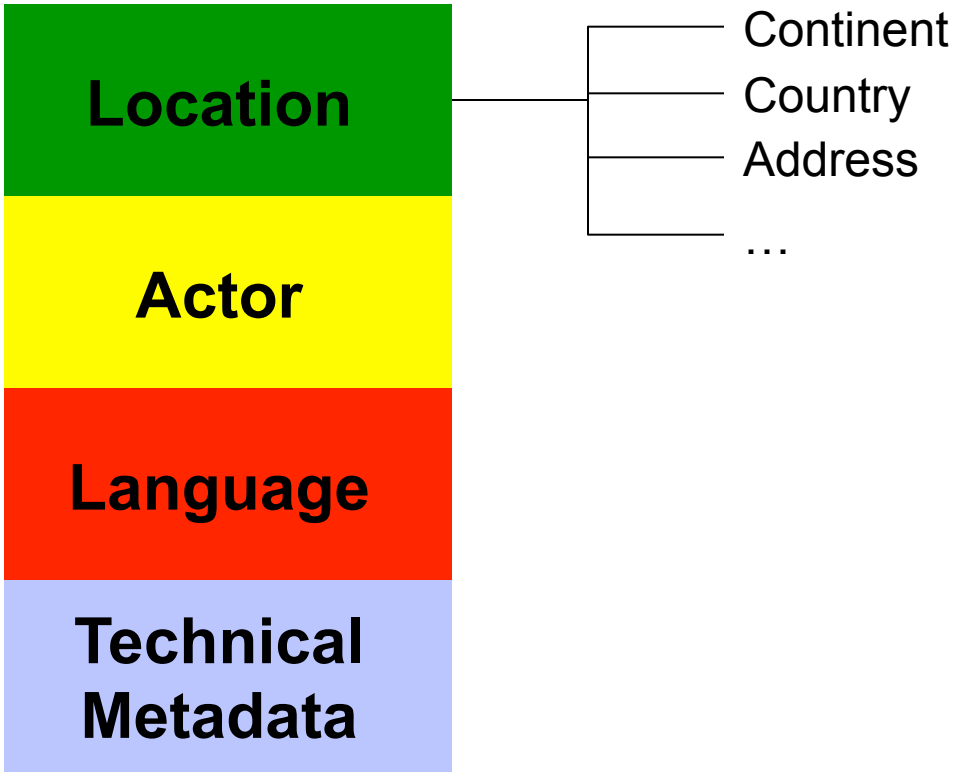


Metadata Components

Let's describe a speech recording

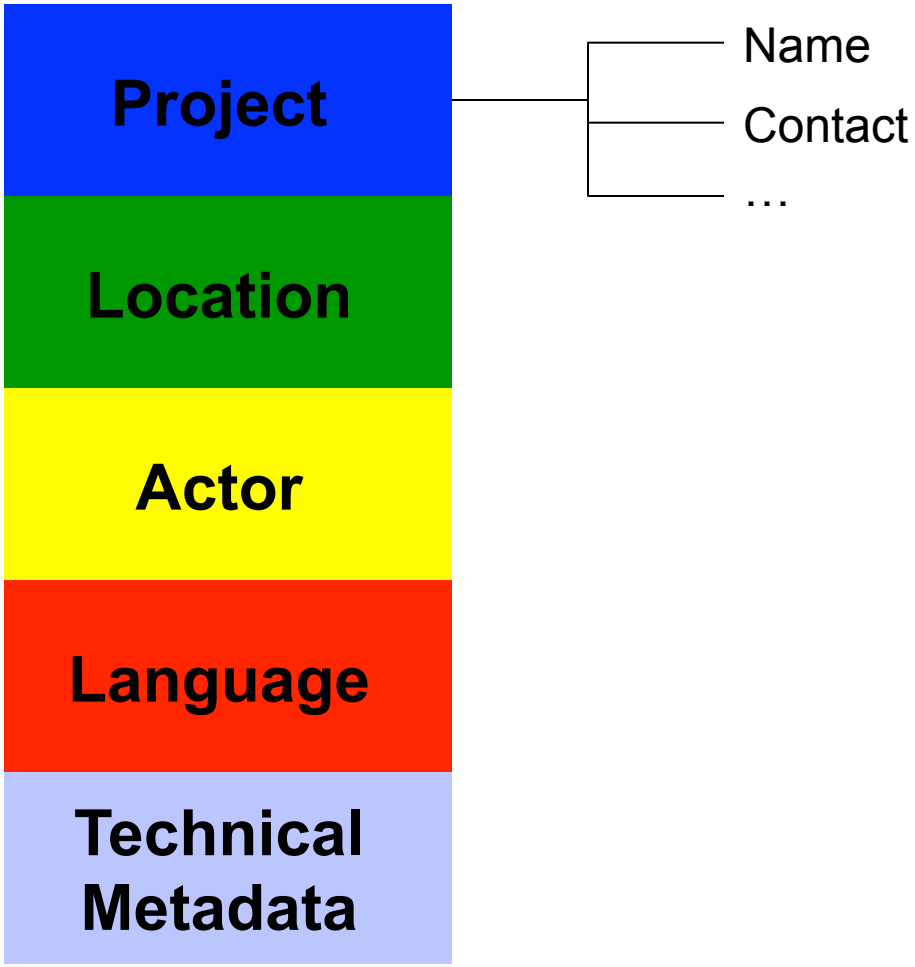


Metadata Components



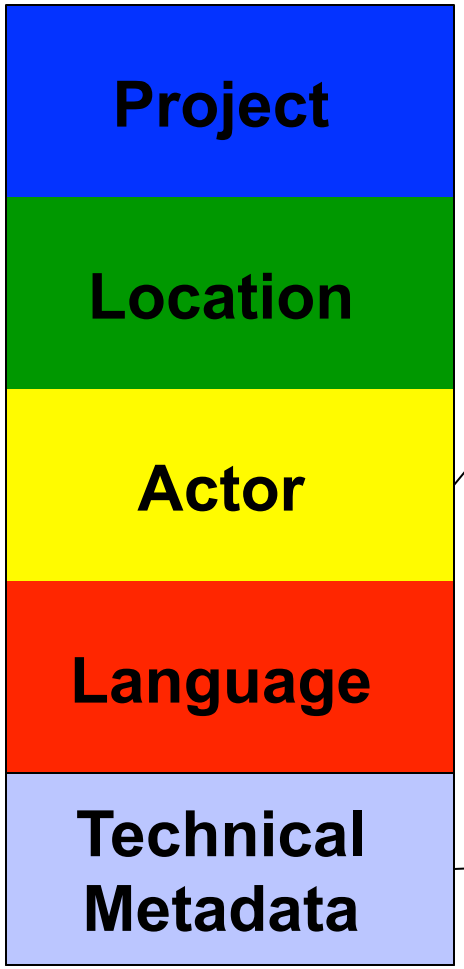
Let's describe a speech recording

Metadata Components



Let's describe a speech recording

Metadata Components



Metadata profile

Profile definition XML

Component definition XML

Let's describe a speech recording

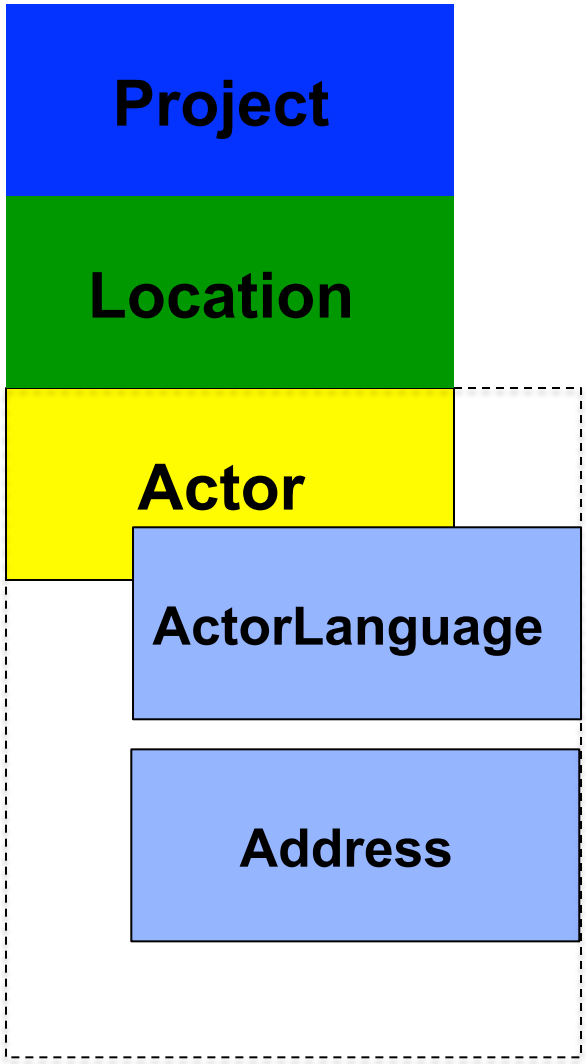
Metadata schema

W3C XML Schema

Metadata description

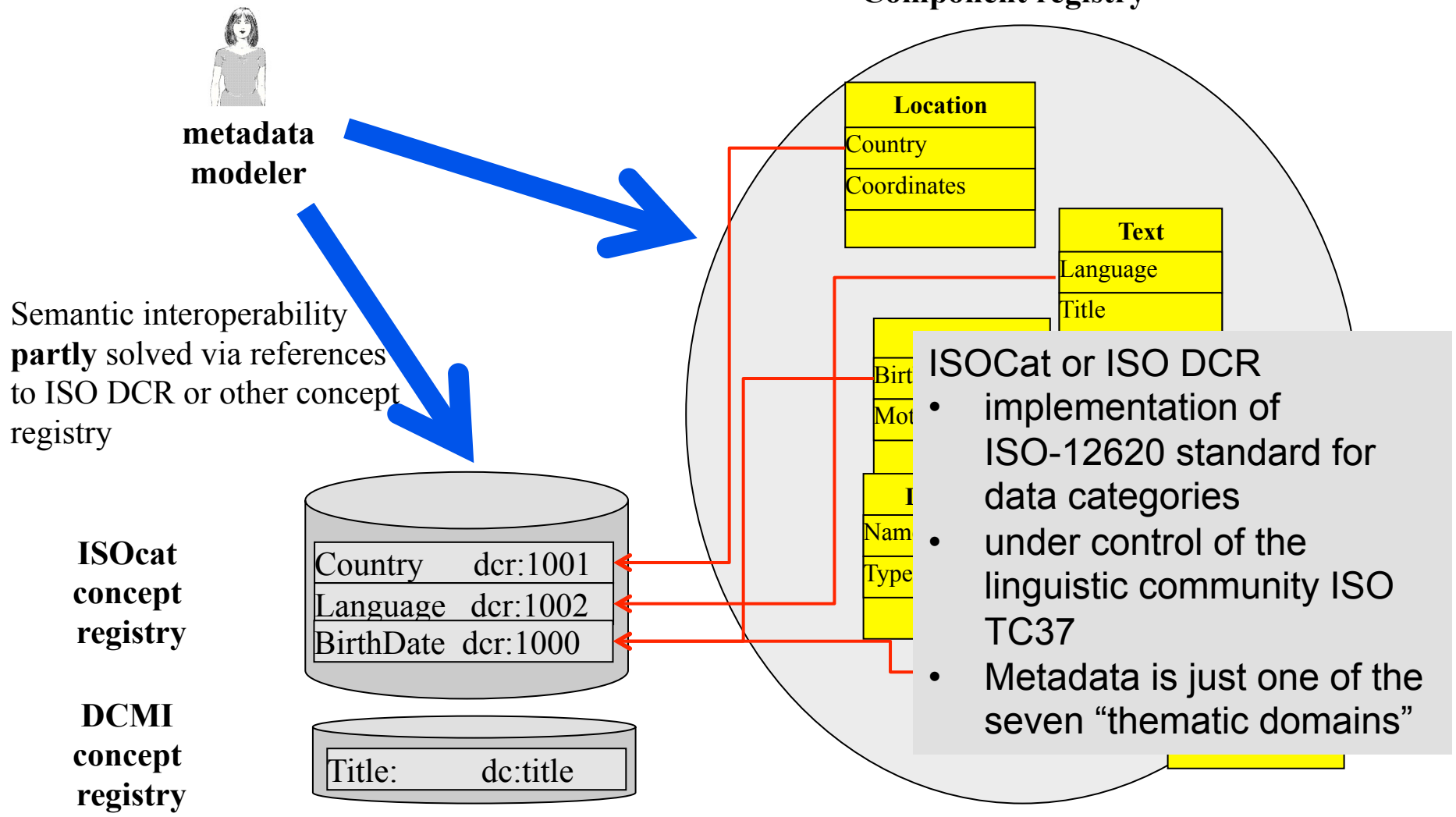
XML File

Recursive model

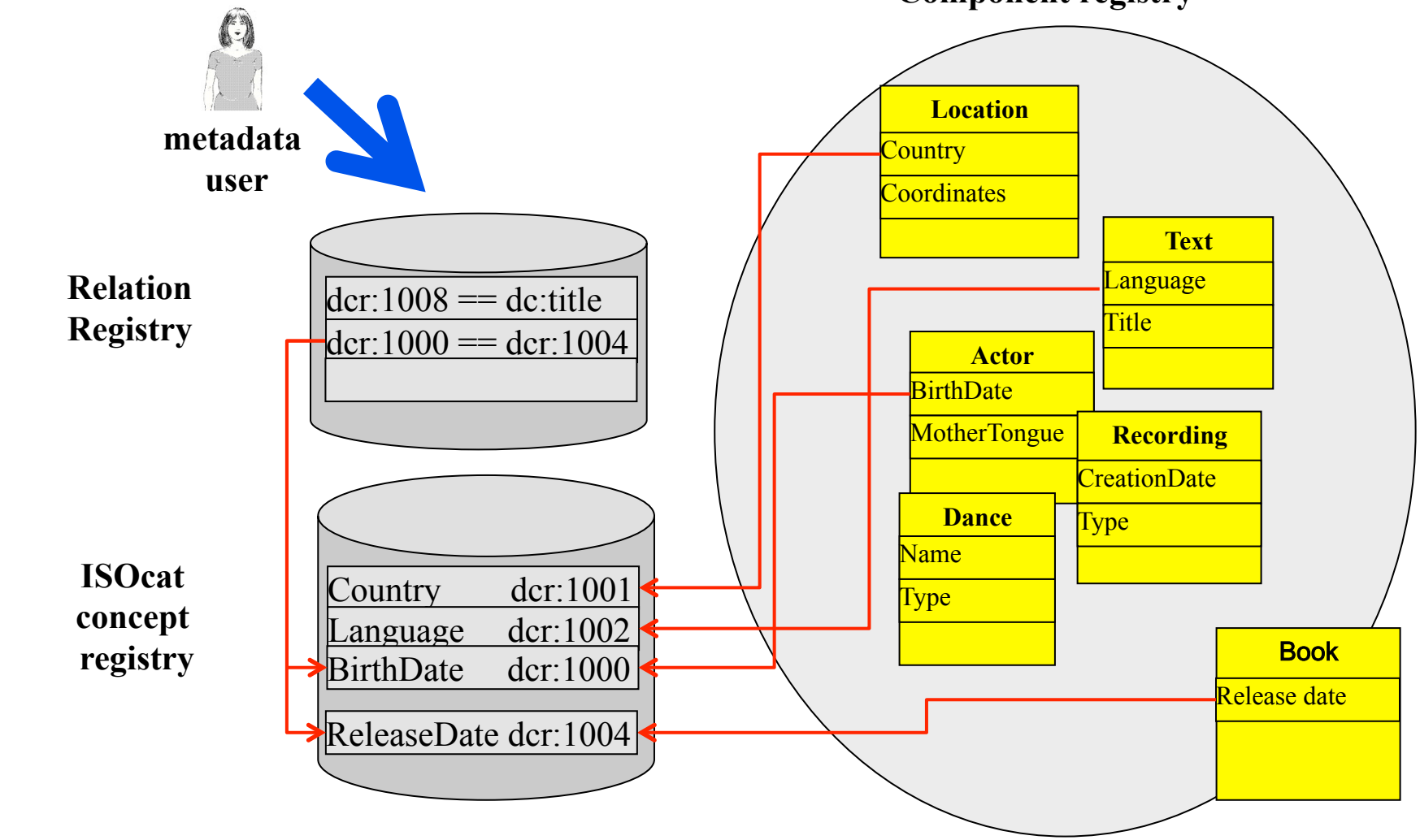


- Recursive Component model
- Components can contain other components
- Enhances reusability

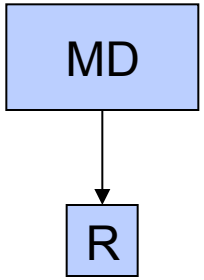
Reusability & Explicit Semantics

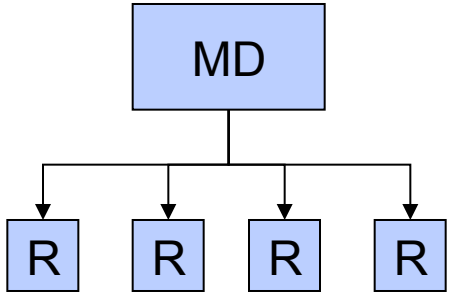


Reusability & Explicit Semantics

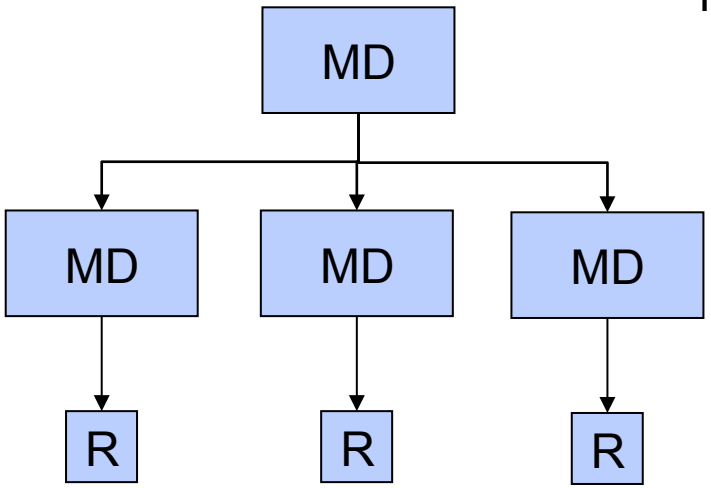


Collections & metadata

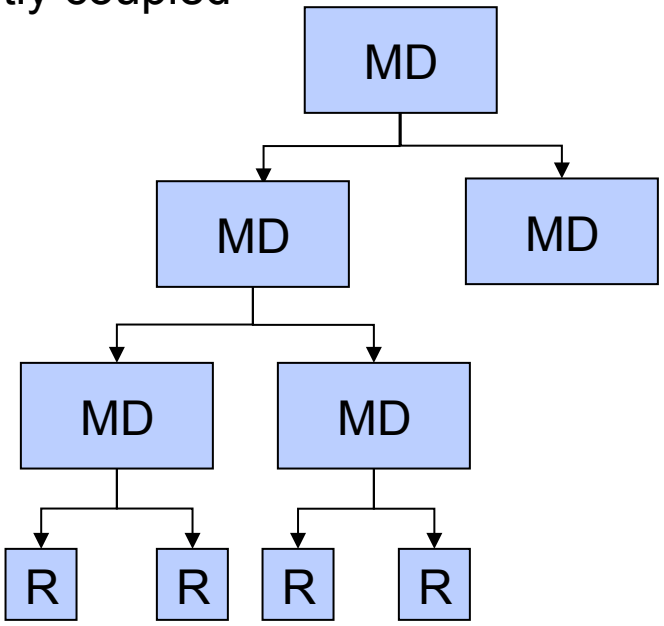




Collection of tightly coupled resources



Collection of resources



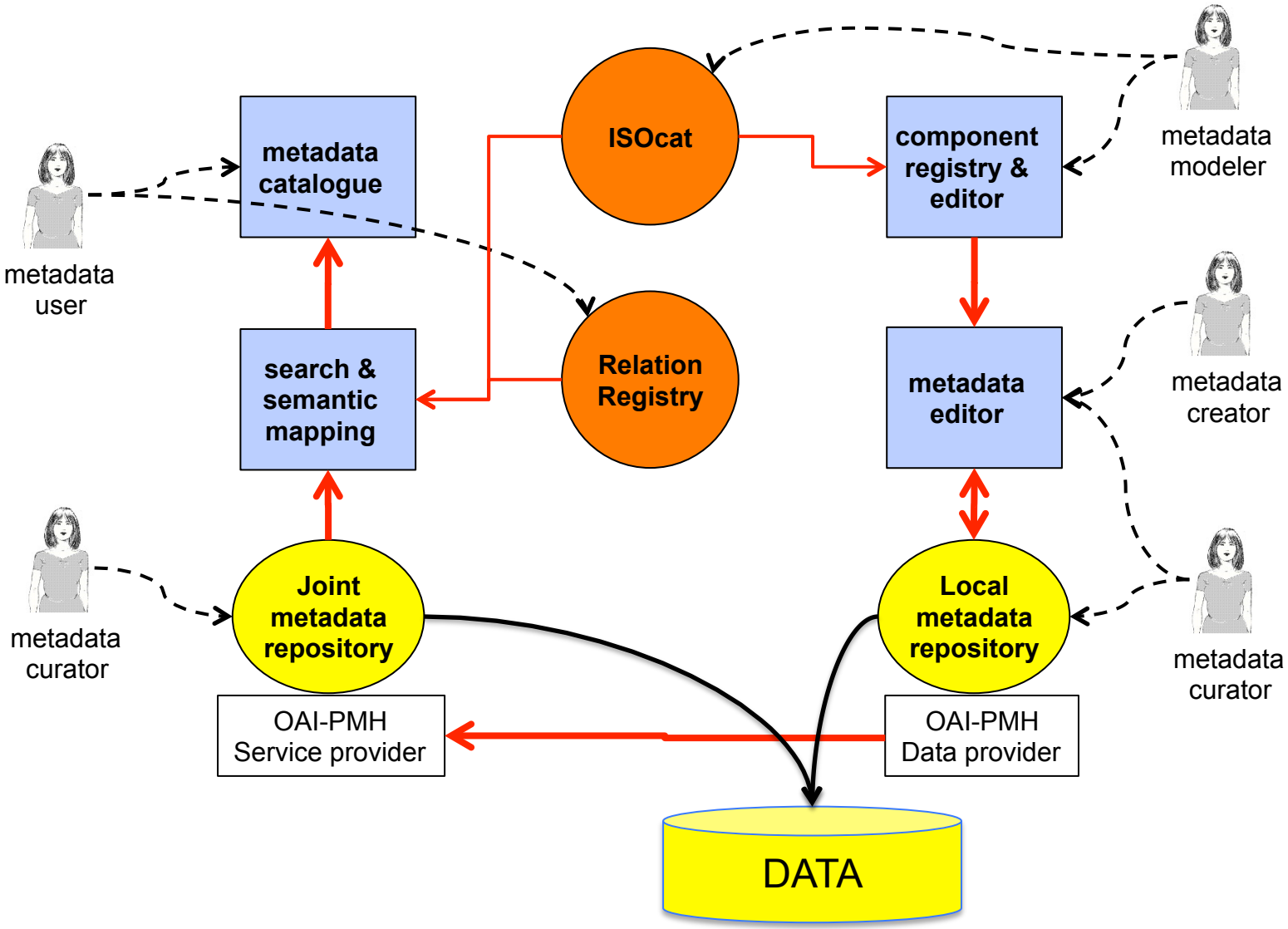
hierarchy of sub-collections

Metadata Actors & Entities

- **Metadata Users** use metadata to find or resources
  - result: suitable resource
- **Metadata Creators** create metadata to describe resources
  - result: metadata description of a resource
- **Metadata Curator** Updates metadata description for maintenance
  - result: metadata description of a resource
- **Metadata modelers** create metadata schema and/or terminology
  - result: 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



CMDI backward compatibility

- 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, CLARIN-AT

National projects: NaLiDa

CLARIN
Common Language Resources and Technology Infrastructure

META SHARE

It is worthwhile to separate the standardization work in a number of separate tasks so that the work can divided over multiple people and that different stakeholder projects can have a share in the responsibility

STANDARDIZATION ROADMAP

Towards Convergence?

- CLARIN, NaLiDa, META-SHARE share many participants
- Metadata creators and modelers are not prepared to do things twice.
- Although implementation is different, the basic model: components & explicit semantics via ISOcat are mutual
- Cooperating in standardization work within ISO TC37/SC4 (Language Resource Management) seems a good base for convergence

Standardization Roadmap

- Standardization of metadata DCs in the ISO-DCR
  - Metadata TDG, chair Peter Wittenburg

- Defining Requirements for a Metadata Component Model and standardizing the Model itself
  - Project leader: Daan Broeder, CLARIN, NEN

- Standardizing a Component Specification Language
  - Project leader: Thorsten Trippel, NaLiDa, DIN

- Design/Specify a number of recommended components for specific data types and usages.
  - Project leader: Maria Gavrilidou, META-SHARE, ELOT

Thank you for your attention

History / Acknowledgement

- CLARIN preparatory phase WP2 workshop Oxford 2009
- MPI-PL (CLARIN NL),
- OAW (CLARIN-AT),
- UGOT,
- IDS (CLARIN-D),
- NaLiDa

- Matej Durco
- Thorsten Trippel
- Oliver Schoenfeld
- Leif-Joran Olsen
- Dieter van Uytvanck
- Menzo Windhouwer
- Peter Withers
- Twan Goossen