

CLARIN

Common Language Resources and Technology Infrastructure



Flexible Component Metadata Frameworks - CLARIN Agenda

Daan Broeder

CLARIN / Max-Planck Institute for Psycholinguistics

NEERI, Helsinki Oct 1,2

Contents



CMDI - CLARIN Metadata Infrastructure

- CMDI Context
- CMDI Overview
- CMDI Software components
- Implementation Issues
- Progress



- Other Metadata Infrastructures in our domain
 - IMDI
 - Extensive metadata set
 - Simple harvesting model
 - Supported by tool & archive infrastructure
 - OLAC
 - Well defined limited metadata set
 - Harvesting system using OAI
 - TEI
 - Extensive metadata set
 - MD merged with annotations

CLARIN Project - CMDI



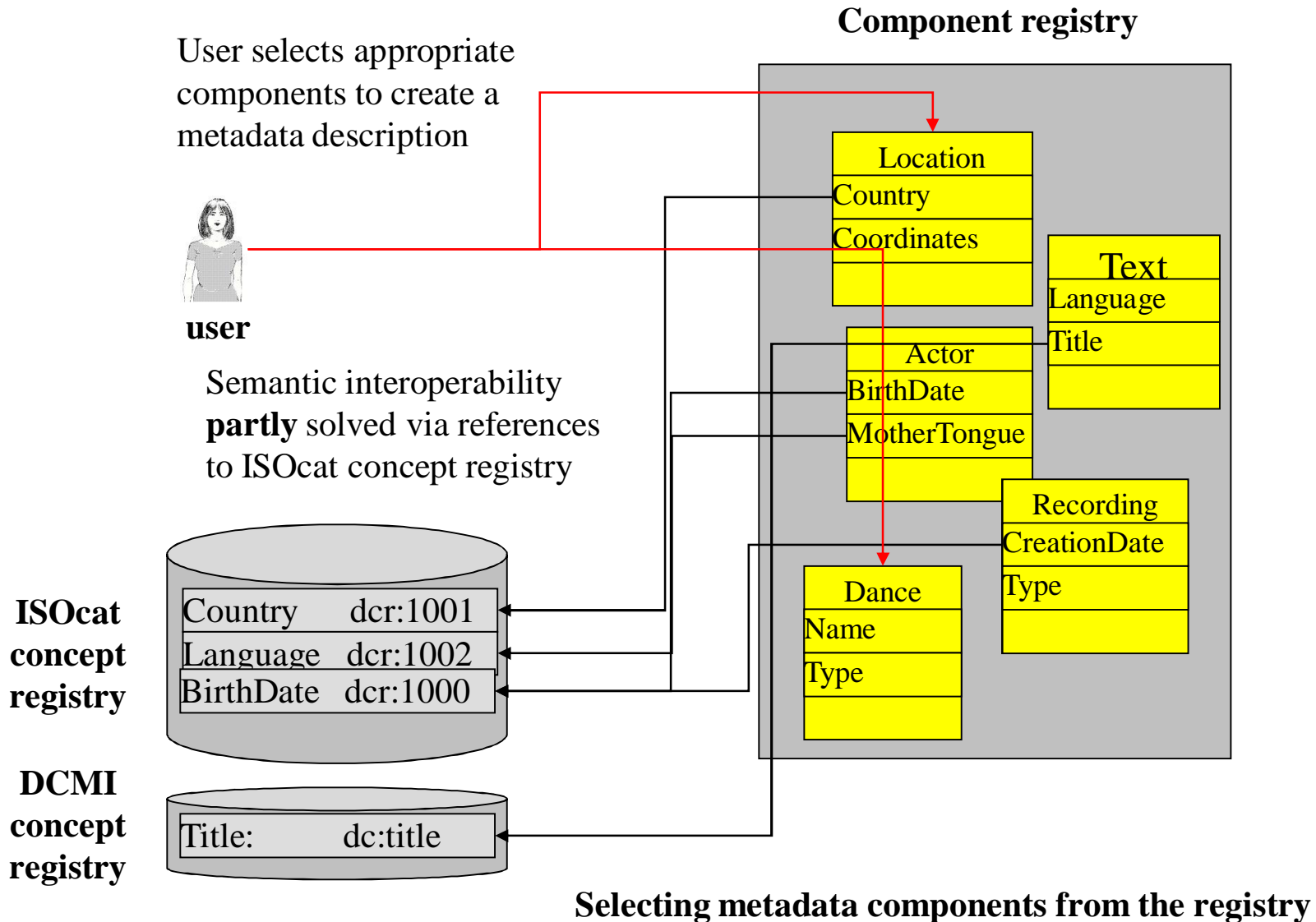
- Metadata infrastructure based on a “Component Metadata Model”
- Aims
 - Flexibility
 - Researcher should themselves decide what metadata fits their needs
 - Offer ready made metadata components
 - Allow them to create new metadata components if they want
 - Complete Infrastructure: software for metadata modeling, editing, harvesting, exploitation
 - Compatibility with existing frameworks: OLAC, IMDI, TEI

CMDI history



- Berlin WP2 workshop, Oct. 2008
- Oxford WP2 workshop, Feb. 2009
- Documents:
 - Metadata Infrastructure for Language Resources and Technology v3 Dec 2008
 - Metadata Infra Work Document, Feb 2009
 - Requirements for Virtual Collections Mar 2009, limited circulation.
 - CLARIN developers wiki
- Nijmegen CMDI Developers Workshop, May 2009
- Dutch CLARIN Component creation project., September 2009
- NEERI CMDI Developers meeting

Metadata Components

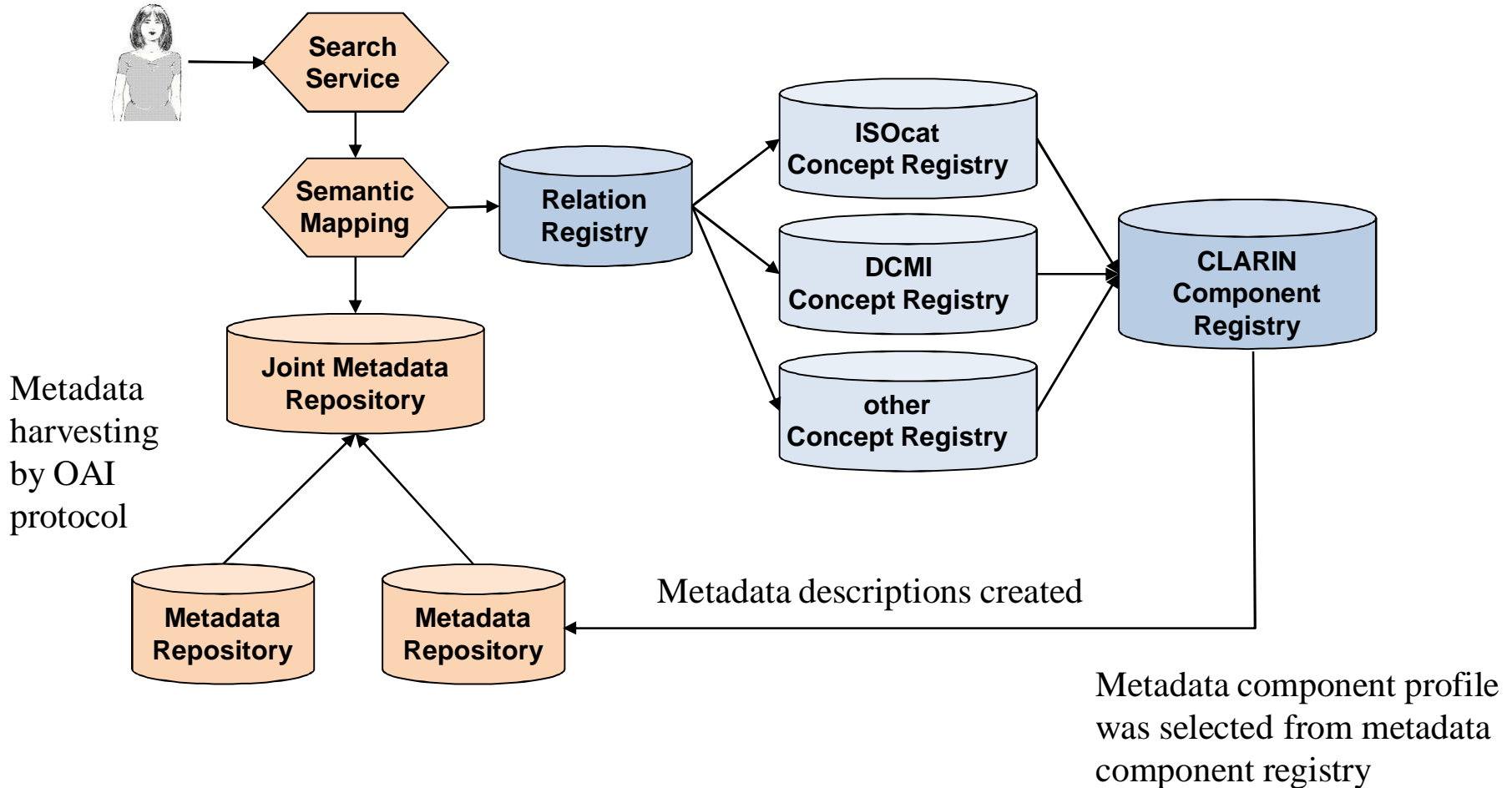


CLARIN MD Live-cycle

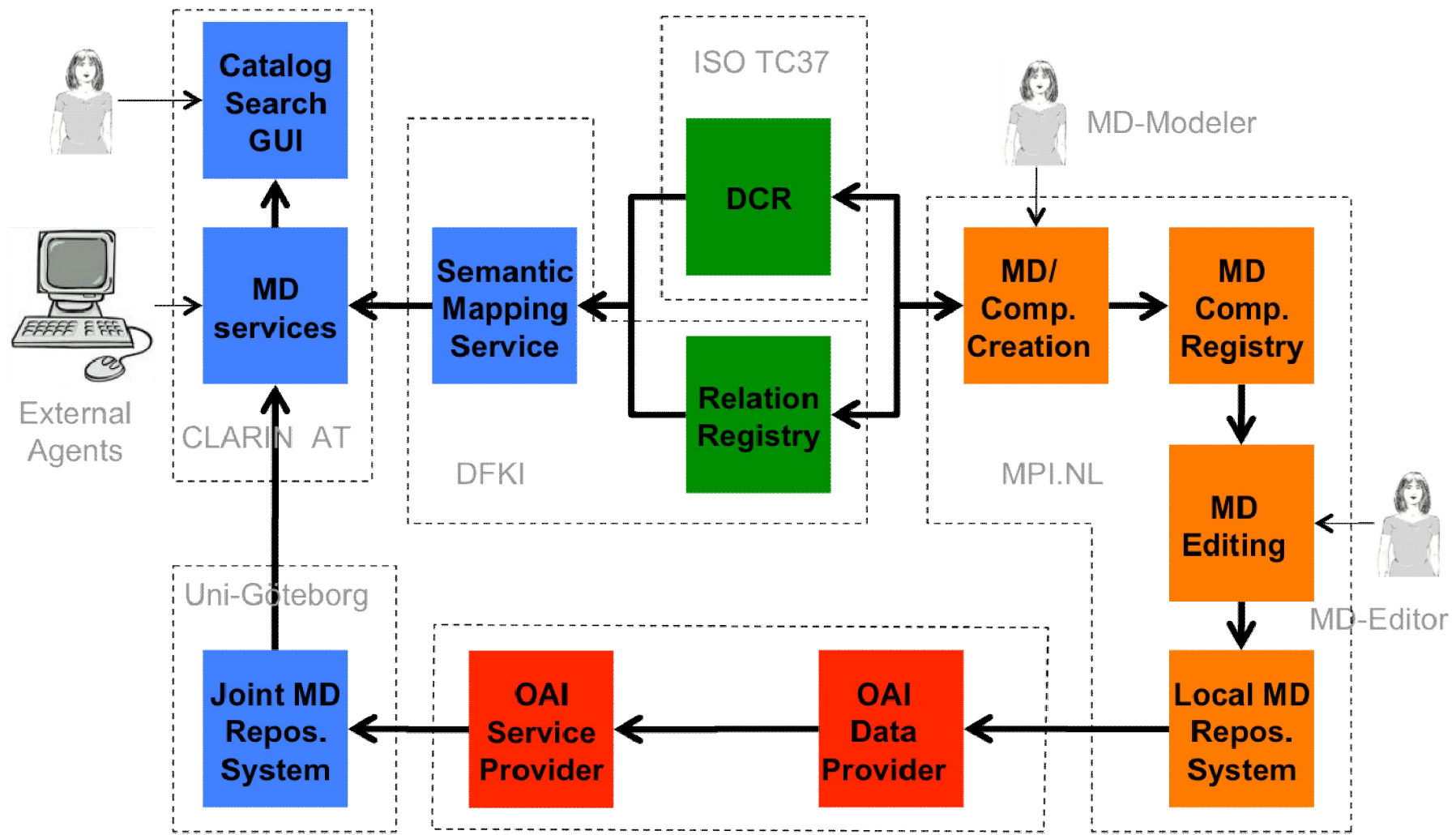


Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry

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



CMDI Software Components



SW Component Taxonomy



- Division into:
 - MD Producer components
 - MD Exploitation or consumer components
 - OAI components
 - Knowledge components: DCR, Relation Registry
- The CMDI takes a 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 can work in parallel on the same metadata
 - LT World (DFKI) is such a consumer

Current CMDI development partners



- AAC, A Catalog, Metadata services
- DFKI, DE Semantic mapping, Relation Registry & LT World
- IDF, DE Virtual Collection specs & registry
- ISO/MPI DCR
- MPI-PL, NL MD provider components
- Språkbanken, SE Joint Metadata Repository

Implementation Issues



- The agreed programming language is Java we can accept a need to import modules written in other languages
- All used libraries and SW components from outside must be public domain (as will be our own code)
- The APIs will be:
 - JAVA APIs where appropriate
 - SOAP and REST

Related activities



- Started small project funded by Dutch CLARIN branch
- Test CMDI XML toolkit
 - XML schema for component specifications
 - Set of existing component specifications based on existing metadata sets: IMDI, OLAC, DC
 - XSLT transforms to create XML schemas from component specifications
- See if this covers the resources at two important Dutch LR providers: INL, Meertens Institute
- See if the current concepts in the ISOcat are sufficient
- Resulting components will seed the CMDI component registry

Progress



- Started with implementation of the Component Registry
- Modifying new IMDI metadata editor for use of metadata components.
- Testing performance of the foreseen metadata repository solution
- Further refining the existing design
- Expect first (first complete version) 3'th quarter – 2010
- Probably earlier prototypes for specific purposes
 - European CLARIN use-case combined metadata/content search

Risks



Too many components are created and semantics are too diffuse to be useful.

- Discourage creating new metadata components
- Limit the set of available metadata components and profiles
- At worst the unified catalog is in fact subdivided into separate domains, but still useful where semantics can be mapped.

Provide guidance by seeding the repository with balanced components

Create a threshold creating own components.

CLARIN

Common Language Resources and Technology Infrastructure



Thank you for your attention

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