

# CMDRSB

CLARIN **Metadata** Repository/**Service**/Browser

<http://clarin.aac.ac.at/MDService2/>

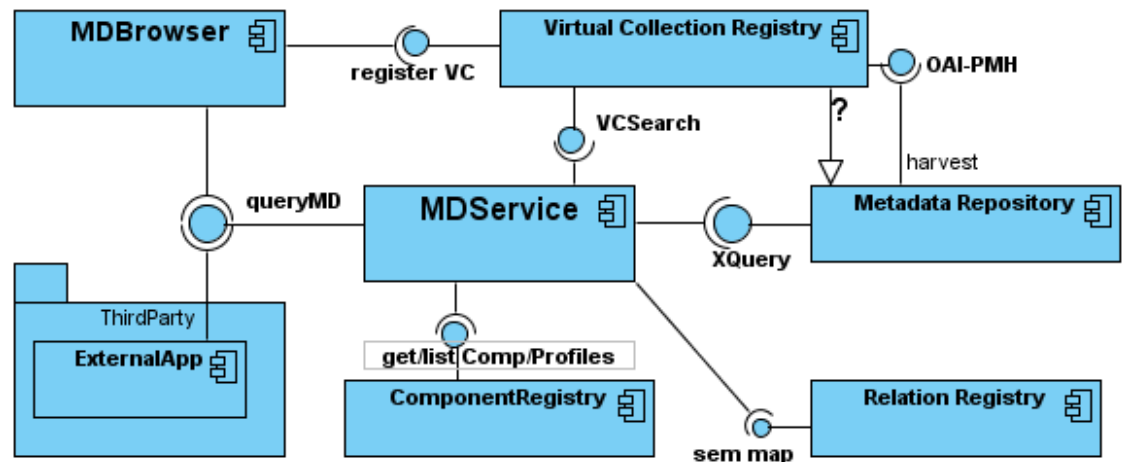
CMDI Workshop

2011-01-17 , MPI Nijmegen

Matej Ďurčo, ICLTT, Vienna;

Leif-Jöran Olsson, Sprakbanken, University of Gothenburg

- **MDSERVICE** accepts queries about metadata from **MetadataBrowser** (and external Applications)
- and passes them to the **Metadata Repository**(ies)
- and/or to the **Virtual Collection Registry**,
- optionally applying **Semantic Mapping** based on the information from **Component Registry**, **Data Category Registries** and **Relation Registry**
- receiving results and passing them (optionally formatted) back to the requesting node.

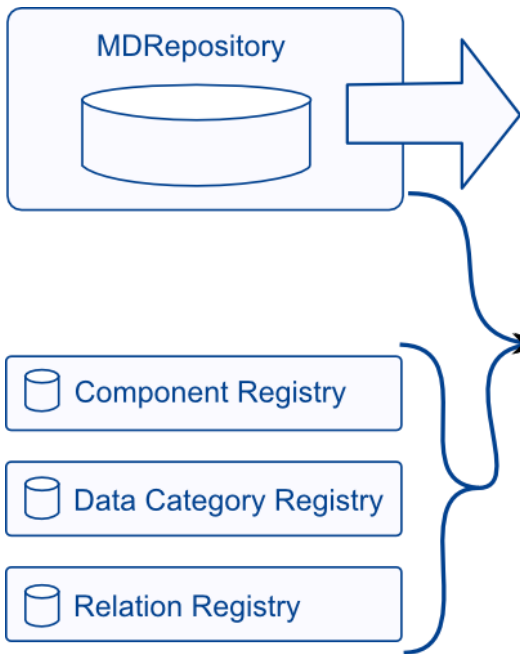


→ REST-interface ([MDService2 WADL](#), [HTML version](#))

- collections
  - list the „natural“ hierarchical collections-structure of the repository
- model
  - return xml-elems used in the repository (with usage statistics)
- terms
  - return terms/indices/xml-elems used in the repository enriched with
    - a) the usage statistics (count occurrences and distinct values)
    - b) the corresponding CMD-components and data categories
- values
  - list distinct values for given index (similar to facet functionality)
- recordset
  - retrieve a list of MDrecords based on a query [CQL]
- record
  - retrieve individual MDrecord based on the identifier

<http://clarin.aac.ac.at/MDService2/docs/htmlpage/info>

- Dynamic Repositories
- Collections browsing
- Terms/Values browsing
- Query Input
  - Simple full-text query
  - Complex queries (CQL-searchclauses, boolean op)
  - Index auto-completion
- Queryset/Resultset
  - work with multiple results in parallel
    - Paging
    - Variable views (select columns, auto-columns)
- Workspace (storing queries, bookmarks)
- „Linkable“ Queries
- (Semantic Mapping)



- collections

the „natural“ hierarchical collections structure of the repository (OLAC, IMDI, ... and subcollections)

- terms

+ information on the XML-structure of the records in the Repository  
= the profile usage statistics (occurrences and distinct values)

+ thanks to: CMD Profiles -> Schemas -> MD-Instances  
it is possible to track back from the instances to the **profiles**  
and from there to link to the **data categories**

```
MDRecord
<CMD><Header><MdProfile>{profileID}</MdProfile>
  <Components><{profileName}>
    <{component}>
      <{element}>
```

```
CMD-Profile-Specification
<CMD_ComponentSpec>
  <Header><ID>{profileID}</ID>...</Header>
  <CMD_Component name="{profileName}">
    <CMD_Component name="{component}">
      <CMD_Element name="{element}"
        ConceptLink="{datcat-uri}">
```

```
Data Category Registry
<dcif:dataCategorySelection>
  <dcif:dataCategory
    pid="{datcat-uri}" >
    {detail-information}
```

- simple full text queries: <http://clarin.aac.ac.at/MDSservice2/docs/htmlpage/queries>
  - simple term [system](#)
  - any of multiple terms (OR) [child | acquisition](#)
  - all of multiple terms (AND - default) [child acquisition](#)
  - Phrase a sequence of terms ["longitudinal study"](#)
- Bookmarks search/remember individual records (by handle/identifier)
  - [clarin-at:aac-test-corpus:C4:158](#)
- Complex search queries [SRU/CQL]
  - basic search clause (index relation term) : [author any Adler](#)
  - boolean [title contains a and imprint.date between 1910 and 1920](#)
- Combine simple query and search clauses
  - [university and \(title any system\)](#)
- Restriction by Collections: [aac-test-corpus](#)
- Search in a profile: [LrtInventoryResource](#)
- Search via DatCats ("Semantic Search"): [isocat:creationDate contains 191](#)

- Basic Idea

query:

```
Actor.Name any Peter
```

+ relations:  
(#DatCat)

```
#sameAs (#Actor, #Person)  
#sameAs (#Name, #FullName)
```

= expanded query:

```
Actor.Name any Peter  
OR Actor.FullName any Peter  
OR Person.Name any Peter  
OR Person.FullName any Peter
```

- Levels

1. just mapping based on the **ConceptLink** resolvable via **ComponentRegistry**
2. use equivalence relation between **DatCats** from **Relation Registry**
3. use equivalence relation also between **Component DatCats** (yet to come)
4. use also **other relations** in Relation Registry (`subClassOf`, `synonymy?`, ...)

- Shall allow for customization
  - save Perspectives/Views
- Currently supporting
  - Storing Queries
  - Bookmarks
- But User-management (join the federation) not solved yet  
Thus there is only one common shared space
- Public Space meant for News, as Dashboard, options...  
obviously function not clear yet



- Built on **CMD**
- Reading Data Category Registries
  - **isoCAT**, **dublincore**, ... (open for further DCRs)
- „Inspired by“ the standard-protocol **SRU/CQL**
  - Started opportunistically, but working towards conformance
  - Supported:
    - The query language CQL (parsing)
    - The format of the result <searchRetrieveResponse>
  - Main differences:
    - The current interface has to be mapped onto the protocol
    - mapping of collections - not solved (explain?)
    - Result-format: scan, explain
    - Diagnostics

<i>SRU/CQL</i>	<i>MDSservice</i>	<i>MDRepo</i>
(explain, Zeerex)?	collections	getCollections
explain	model, terms	queryModel
scan	values	scanIndex
searchRetrieve (CQL)	recordset, record	searchRetrieve (but XPath!)

- Interaction with Virtual Collection Registry
  - Stored query → Intensional VC
  - Result of a (stored) query (Recordset) → Extensional VC
  - Collection of Bookmark → Extensional VC
- produce a SRU/CQL-protocol conformant REST-interface
- Custom Termsets
- Custom Views
- Queries Sorting
- Result Export (especially also only selected fields (not full records))
- translate UI, BUT mainly also the search-indices
  - based on DCR language-sections
- Commenting – collaborative curation
  - Allow to „annotate“ / comment / make notes on the MDRecords
  - just get email. via POSTing trac.tickets?

- The MDRepository currently contains around 109.000 records, mainly from the datasets: OLAC and IMDI ([data statistics](#))
- Currently there are three instances of the MDRepository running providing similar but not identical datasets:
  - University of Gothenburg (main)
  - ICLTT, Vienna
  - MPI Psycholing, Nijmegen

(they can be accessed by the same MDService, by switching the target repository in the UI)
- A first version of the MDService and Browser is online:
  - [clarin.aac.ac.at/MDService2](http://clarin.aac.ac.at/MDService2) (this address may change)
  - Although the repository and interface already provide a lot of information and functionality, it is demo-quality and cannot yet be seen as reliable service.
  - Lot of work is still needed both on the data quality and user interface:
    - rework of the UI - based on feedback at CMDI-Workshop, Nijmegen20110117
    - continuous integration of new datasets (provided for harvest by the centres)
- Nevertheless we invite you to try it out and look forward to any critical remarks

## Overview of functionality (not completely uptodate but largely correct)

**CLARIN Metadata Service** login  
update-uri:/MDSservice2/compprofile/htmllist/p\_1274880881885

*simple search*

**Query**

tel:biblStruct.monogr.title = rat sc0-0

tel:monogr.author any Peter sc0-1

tel:imprint.date < 1910 sc1-0

**collections** C4\_transl **columns** titleStm.title.sourceDesc.biblStruct.monogr.author.s  
ourceDesc.biblStruct.monogr.imprint.publisher.sour  
ceDesc.biblStruct.monogr.imprint.pubPlace.source

q2: C4\_transl hits: 29; from: 1 max: 10

q1: sourceDesc.biblStruct.monogr.title any rat hits: 21; from: 1 max: 10

pos	title	author	publisher	pubPlace	date
1	Die Prostitution in Wien in historischer, administrativer und hygienischer Beziehung: a machine-readable transcription	Schrank, Josef	Im Selbstverlage des Verfassers	Wien	1886
2	Führer durch die Schautellungen des Wiener Thiergartens (k. k. Prater, am Schüttel) und des Wiener Vivariums (k. k. Prater, Hauptallee 1): a machine-readable transcription			Wien	1895
3	Sammlung von Zeitungs=Artikeln betreffend die Baum-Frage.: a machine-readable transcription			Wien	1871
4	Der kategorische Imperativ: a machine-readable transcription	Bauernfeld, Eduard		Wien	1851
5	Laurenz Hailers Praterfahrt: a machine-readable transcription	Auernheimer, Raoul	S. Fischer	Berlin	1926

*view full record-detail*

**Public Space**  
Personal Workspace  
Querysets sample queries  
name sample queries  
Test | aac-test-corpus  
Schau |  
sourceDesc.biblStruct.monogr..|  
monogr any Leben |

**Collections**  
aac-test-corpus [467]  
OLAC: olac-root [5375]

**Terms**  
teiHeader [467]  
teiHeader [467/0/0]  
fileDesc [467/0/0]  
profileDesc [467/0/0]

**Profiles/Components**  
imdi-corpus  
ID: clarin.eu:cr1:p\_1274880881885  
Name: imdi-corpus  
Description: IMDI corpus profile  
imdi-corpus

**Annotations:**

- stored queries: - Public Space, - Personal Workspace
- browse collections hierarchy
- explore and search by "terms" i.e. elements from used profiles, referenced data categories and relations between them
- complex search: AND/OR combined SRU/CQL search clauses: 'index relation term'
- dynamic columns (auto-columns)
- paging
- working with queryset: multiple queries/results in parallel
- link to the resource (provide Resource Viewers)

**Resource Viewers:** Content ProviderX, Content ProviderY, ProviderZ