

Working towards a Metadata Federation of CLARIN and DARIAH-DE

Thomas Eckart

Natural Language Processing Group
University of Leipzig
teckart@informatik.uni-leipzig.de



Tobias Gradl

Applied Computer Science Group
University of Bamberg
tobias.gradl@uni-bamberg.de



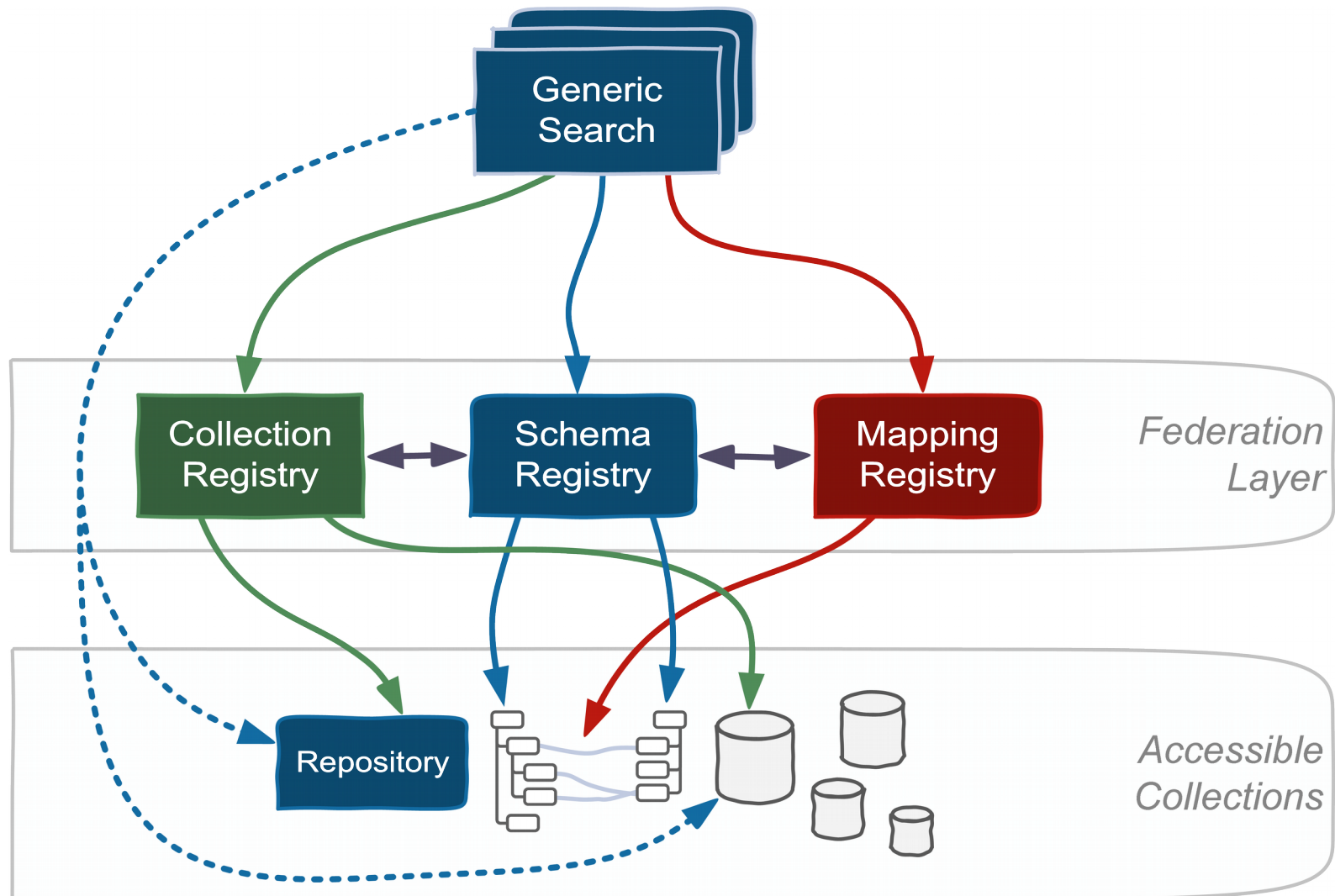
Agenda

- *Metadata Infrastructure in CLARIN*

(CMDI, Component Registry, OAI-PMH, VLO, Metadata Curation ...)

- Metadata Infrastructure in DARIAH-DE
- General Thoughts
- Steps to a MD Federation
- Conclusion

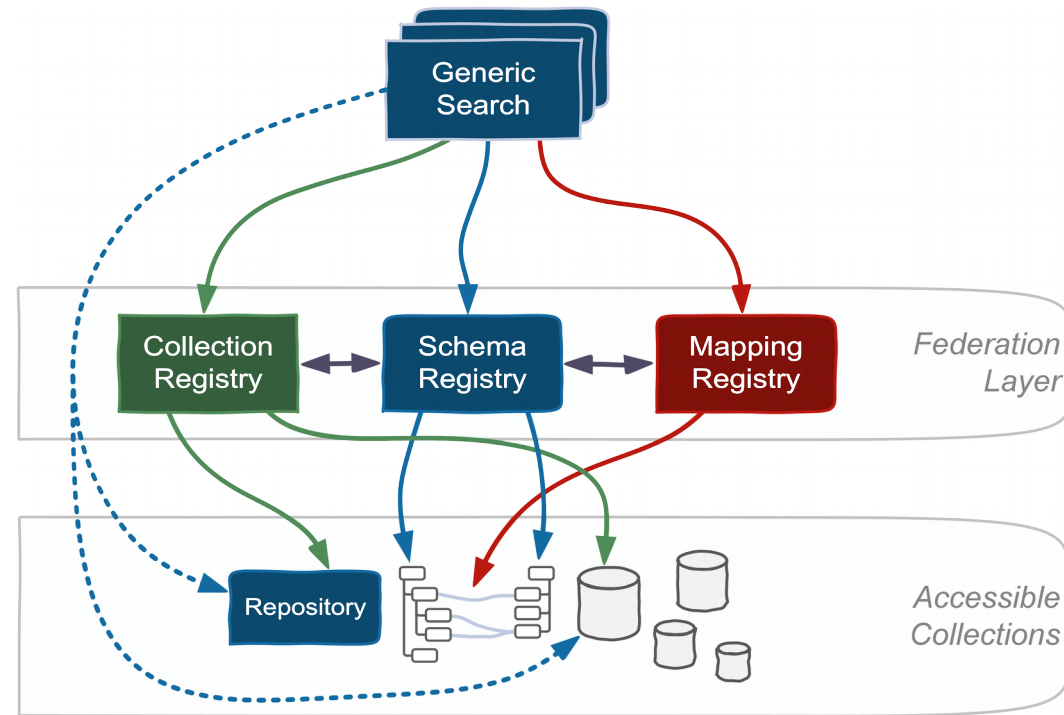
DARIAH-DE MD Infrastructure: The big picture



DARIAH-DE MD Infrastructure: Differences and Analogies

- No central integration model, but “semantic clusters”
- Integration models are generated according to needs of particular research questions
- Larger perspective: mapping to more generic schemata

DARIAH-DE	CLARIN
Collection	~ <i>Centre</i>
Collection Registry	Centre Registry
Schema Registry	CMDI Component Registry
Mapping Registry	<i>None, implicit via ConceptLinks</i>
Generic Search	VLO



DARIAH-DE MD Infrastructure: Mapping Registry

The screenshot displays the DARIAH-DE Mapping-Editor interface. At the top, the header includes the DARIAH-DE logo, 'Schema Registry', and navigation links for 'Language' and 'Login'. The main title 'Mapping-Editor' is prominently displayed. Below the header, the breadcrumb trail reads 'Schema Registry / Schemas und Mappings / Mapping-Editor'. The interface is divided into two main sections: 'Sample transformation' on the left and 'Element model' on the right.

Sample transformation section includes:

- Sessions:** Buttons for 'Save session', 'Load session', and 'Reset'.
- Sample data:** A section with 'Input' and 'Results' tabs. The 'Results' tab shows '100' resources found. An 'Execute' button is present.
- Source schema:** A dropdown menu showing 'singlePaperPackage'.
- Target schema:** A dropdown menu showing 'oai_dc'.
- 1 / 100:** A pagination indicator.
- DC:** A list of elements with their descriptions:
 - Title:** The role of semantics, argument structure, and lexicalization in compound stress assignment in English
 - Description:** It is generally assumed that noun-noun compounds in English are

Element model section shows a diagram mapping elements from the source schema to the target schema. The source schema elements on the left include 'ComponentId', 'Type', 'OriginInfoLegacy', 'Mods', 'Base', 'Ref', 'ComponentId', 'Abstract', 'Genre', and 'TypeOfResource'. The target schema elements on the right include 'Title', 'Creator', 'Subject', 'Description', 'Publisher', 'Contributor', 'Date', 'Type', 'Format', and 'Identifier'. Blue lines with yellow double arrows indicate the mappings between these elements.

At the bottom of the interface, a log shows the following messages:

- 2017-04-27 16:12:57.761 SUCCESS Sample input transformed (total 787ms): 100 resources found
- 2017-04-27 16:12:56.659 SUCCESS Sample input processed (total 136ms): 100 resources found
- 2017-04-27 16:12:56.113 INFO Sample set for your current session
- 2017-04-27 16:08:45.598 INFO Editor Sitzung gestartet [id: 5901fb6d3c9dccc0564b4d920]

The footer of the interface includes '© DARIAH-DE' on the left and 'Privacy Legal information Contact' on the right.

URL: <http://schereg.de.dariah.eu/>

General Thoughts

- Design decision reflect characteristics of the targeted research communities
- As a consequence, implemented solutions differ in their architecture and functionality
- Basic assumption: both metadata infrastructures are *optimized* for their specific use cases and user groups
 - **No need for another interoperability layer, reusing and combining functionality instead**

Steps towards a MD Federation: DARIAH-DE → CLARIN

- Central interface: Dublin Core via OAI-PMH
- CLARIN MD Harvester already capable of transformation to CMDI; test harvests
- Acceptable (but still suboptimal) results in VLO
- Minor changes in VLO vocabulary mappings
- Problematic:
 - (Missing) License information
 - Resource links usually “Landing pages”

Steps towards a MD Federation: DARIAH-DE → CLARIN

The screenshot shows the Virtual Language Observatory (VLO) search results page. The header includes the VLO logo, search, and help links, along with the CLARIN logo. The breadcrumb trail indicates the path: VLO / Faceted search / Search results. A search bar is present with a search icon and a help icon. Below the search bar, it shows the number of results (21 to 30 of 2466) and the selected filters: German, DARIAH-DE, and Monograph. The results per page are set to 10. On the left, there are faceted search categories: Language (German), Collection, Resource type (Monograph), Format, Subject, and National project (DARIAH-DE). The main results area displays a list of search results, including titles like 'Dresdner Briefsteller zum Gebrauch für Stadt- und Landschulen' and 'An Ihre Hochwohlgebohrnen Gnaden, die Frau Majorin von Brettin'. Each result has a small icon indicating the number of documents (1) and a question mark icon.

Virtual Language Observatory Search Help CLARIN

VLO / Faceted search / Search results

Search

Showing 21 to 30 of 2466 results within selection for German DARIAH-DE Monograph Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language German

Collection

Resource type Monograph

Format

Subject

National project DARIAH-DE

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Dresdner Briefsteller zum Gebrauch für Stadt- und Landschulen : Nebst einer Anweisung zur Orthographie, einem Titularbuche und einem Verzeichnisse französischer im gemeinen Leben oft vorkommender Wörter 1 ?

An Ihre Hochwohlgebohrnen Gnaden, die Frau Majorin von Brettin bey Ihrer Ankunft in Westphalen im October 1770 1 ?

Verzeichnuß des von nachstehenden Kayserlich-Königl. und Römischen Reichs-Regimentern bey der unterm 21. Sept. 1759. unweit Meissen vorgewesenen Affaire erlittenen Verlustes 1 ?

URL: <http://clariah.informatik.uni-leipzig.de/vlo/?fq=nationalProject:DARIAH-DE>

Steps towards a MD Federation: CLARIN → DARIAH-DE

- Import of CMDI schema in Schema Registry
- Manual creation of mappings CMDI ↔ Dublin Core
- Creation of test collections in Collection Registry for CLARIN OAI-PMH endpoints
- Import in Generic Search

Steps towards a MD Federation: CLARIN → DARIAH-DE

The image displays two screenshots of the DARIAH-DE web interface, illustrating the steps towards a Metadata Federation.

Left Screenshot: Mapping-Editor

The interface shows the 'Mapping-Editor' for mapping the 'singlePaperPackage' source schema to the 'oai_dc' target schema. The 'Sample transformation' section includes buttons for 'Save session', 'Load session', and 'Reset'. The 'Sample data' section shows 'Input' and 'Results' (100) with an 'Execute' button. The 'Element model' section displays a tree structure of elements: ComponentId, Type, OriginInfoLegacy, Mods, Base, Ref, ComponentId, Abstract, Genre, and TypeOfResource. The 'Sample data' section shows a list of sample data entries, including 'Title' and 'Description'.

Right Screenshot: Generic Search

The interface shows the 'Generic Search' results for the query 'compound stress assignment'. The 'Extended options' section includes a dropdown for 'Data structures' (oai_dc) and a 'Search options' section with checkboxes for 'Show explanations' and '20 Results per page'. The 'Queried collections' section lists the following collections and their counts:

- Bayerisches Digitales Repositorium 17
- OpenScience Repository Leipzig 6
- Zentrales Verzeichnis Digitalisierter Drucke 3
- Göttinger Digitalisierungszentrum – Autobiographica
- Göttinger Digitalisierungszentrum – Book conservation

32 more collections are available. The 'Search' button is at the bottom.

The 'Extended search' section shows the search results for 'compound stress assignment'. The 'Resources' tab is selected, showing 6 of 26 resources. The first resource is 'OpenScience Repository Leipzig' with the title 'Compound stress assignment by analogy: the constituent family bias'. The 'Content' section shows the following details:

- CMD:
- CMDVersion: "1.2"
- Components:
- SinglePaperPackage:
- PaperPackage:
- Mods:
- Abstract: "This paper tests the hypothesis that stress assignment to English compounds works on the basis of analogy. In particular, the role of the constituent family, i.e. the set of compounds that share the same right or left constituent with a given compound, is investigated. On the basis of large amounts of data from three diert corpora it is shown that the

URL: <https://search.de.dariah.eu/>

20.09.2017

CLARIN Annual Conference 2017, Budapest

10

Conclusion

- No major obstacles (but open work on details); still prototypical implementation
- Potential of mutual re-use of infrastructure components
 - Mapping registry: making relations between (CMDI) schemas explicit?
 - (Semi-)automatic import of schemas between CR and SR?
 - Metadata curation workflow for DARIAH-DE resources?
- Main problem: outstanding decision about relevant resources that should be imported in production systems

Thank you!

Questions? Remarks?

More information about MD in DARIAH-DE?

- Tobias Gradl and Andreas Henrich. 2015. *A novel approach for a reusable federation of research data within the arts and humanities*. Digital Humanities 2014: Book of Abstracts. Lausanne, CH: 382–384.