Working towards a Metadata Federation of CLARIN and DARIAH-DE

Thomas Eckart
Natural Language Processing Group
University of Leipzig, Germany
teckart@informatik.uni-leipzig.de

Tobias Gradl
Media Informatics Group
University of Bamberg, Germany
tobias.gradl@uni-bamberg.de

Abstract

Over the past years great effort went into the establishment of research infrastructures for the Humanities and Social Sciences. As a result of the diversity of their targeted research fields and communities, miscellaneous infrastructure projects have developed unique solutions for overlapping target groups. This especially holds for describing available resources by structured metadata and providing them to a wider audience in a user-friendly fashion. This paper focuses on recent work to overcome the gap between the metadata infrastructures of *CLARIN* and the German branch of the *DARIAH* project, focusing on design decisions made by both projects and preliminary work focusing on the synergetic evolution of both.

1 Introduction

Research infrastructures for the Humanities and Social Sciences like CLARIN or DARIAH have established specific environments for dealing with descriptive metadata. Naturally, design decisions taken reflect characteristics of the targeted research communities and contribute to specific project goals. As a consequence, implemented solutions differ in their architecture and functionality.

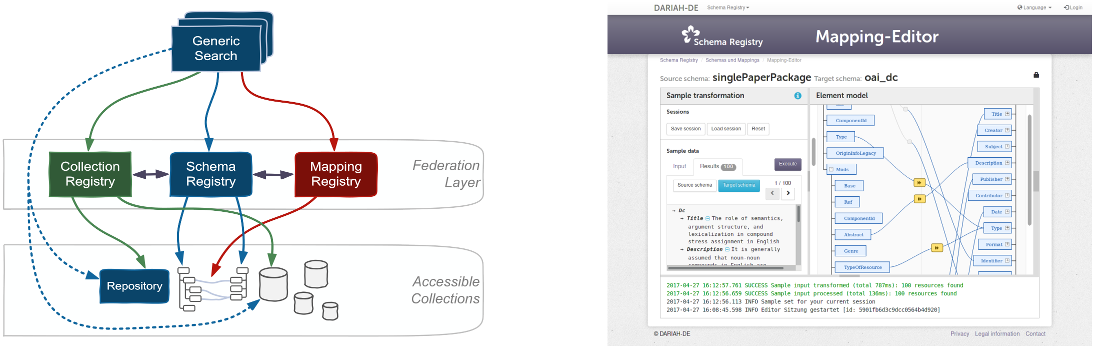
This paper shortly describes the different approaches taken in CLARIN and the German branch of the DARIAH project (DARIAH-DE) and how interoperability between both could be established. Basic assumption is that the existing metadata infrastructures are already optimized for their specific use cases and user groups. Hence, no demand is seen in creating another interoperability layer on top of both. Instead the focus lies on reusing existing software components and interfaces and (if appropriate) combining functionalities of both projects in a federated manner.

2 Overview of the CLARIN metadata infrastructure

CLARIN's metadata infrastructure[1] is tightly linked to the *Component MetaData Infrastructure* CMDI (Goosen et al., 2015) that aims to overcome semantic and structural heterogeneity by using a component-based approach of specifying metadata schemata in a federated environment. The key aim is that resource providers (typically certified centres registered in CLARIN's *Centre Registry*) create structured description "templates" for their resources according to this CMDI "meta standard" in a central registry (*Component Registry*) and provide their structured metadata instances via the standard interface OAI-PMH to the public. For popular metadata schemata or vocabularies (like DCMI, OLAC, etc.), transformation tools are provided for a seamless conversion to their CMDI-based counterparts.

Besides the Component Registry, several tools were developed to accompany resource providers and users in the process of metadata creation, distribution, aggregation and retrieval.

[1] https://www.clarin.eu/content/component-metadata



Source schema: singlePaperPackage Target schema: oai_dc

```
→ Dc
  → Title ⊟ The role of semantics,
    argument structure, and
    lexicalization in compound
    stress assignment in English
  → Description ⊟ It is generally
    assumed that noun-noun
```

```
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: 5901fb6d3c9dcc0564b4d920]
```

(a) Architectural overview of the DARIAH-DE DFA

(b) Mapping CMDI based schema to Dublin Core

Figure 1: DARIAH-DE Data federation architecture

As a central and user friendly search interface, the VLO (Goosen and Eckart, 2014) was developed and is constantly improved. Problems of metadata quality (like vocabulary mismatches or unfit schemata) are countered by an accompanying metadata curation taskforce creating a constant feedback loop between metadata providers and the processing instances in CLARIN (Ostojic et al., 2016).

3 Overview of the DARIAH-DE metadata infrastructure

Infrastructural components for registering sources of metadata, as well as modeling and processing schemas, are encapsulated in the DARIAH-DE Data Federation Architecture (DFA)[2]. Figure 1a presents an overview of the DFA (Gradl and Henrich, 2016a)—highlighting its primary architectural components and access patterns. In general, the DFA employs the idea of community-driven regulation: holders of DFN-AAI[3] accessible user accounts can register, describe and enhance artefacts. Although the components implement concepts of drafting, limitation of write-access and versioning, a fundamental idea behind the DFA consists in the desired community-ownership of data.

- The *DARIAH-DE Collection Registry* contains collection descriptions. Any data source or physical collection of potential relevance to a scholar or research project is allowed to be registered.
- Within the *DARIAH-DE Schema & Mapping Registry*, data models can be imported from schema definitions like XML Schema, extended and mapped with other models. Particular focus is on the *explicability of context information* to protect and improve the interpretation of data in contexts outside of its original collection (Gradl and Henrich, 2016b).
- The *DARIAH-DE Generic Search* is one of the main consumers of the federation layer services. Due to the generic nature of the search, Dublin Core is widely used to create overall views of highly heterogeneous data. More sophisticated schemas can be utilized, however, to exploit features of a limited and cohesive set of collections.
- Aside from external collections that are accessible to DARIAH-DE components, the *DARIAH-DE Repository* serves as central component for storing and publishing research data.

With respect to metadata integration, the DFA avoids the definition of a central integration model, but instead employs the idea of self-developing so called *semantic clusters* (Gradl and Henrich, 2015). Scholars and researchers specify integration models according to the needs of particular research questions. Ideally, such models should be based on standards such as

[2] https://de.dariah.eu/data-federation-architecture

[3] https://www.aai.dfn.de/

CLARIN'S CMDI. They can however, also be developed in a generative fashion to support cases that require new perspectives. A semantic cluster in the exemplary context of a particular research project is developed on the basis of some relevant input data sources (typically OAI-PMH or XML/CSV files). In case there are multiple export models, they are assessed, selected and modeled with respect to their usability for the particular use-case. Mappings between the local models or between the local models and a selected/developed integration model compose a semantic cluster—solving the needs of a particular domain or research question. In combining multiple such semantic clusters, research data can be reused in other contexts by mapping the integration models of the connected clusters. Overall perspectives on data, e.g. within the DARIAH-DE Generic Search, finally utilize more generic schemata such as METS/MODS or Dublin Core.

4 Metadata integration

As a result of the diversity of research questions and domains addressed by both initiatives, CLARIN and DARIAH-DE have developed individual concepts and components as foundations of their respective metadata infrastructures. Initial evaluations and tests in both projects resulted in the key finding that—although disparate research contexts require the continuing focus on the individual concepts—both infrastructures can be considered complementary to each other as they focus on different aspects of the research data lifecycle.

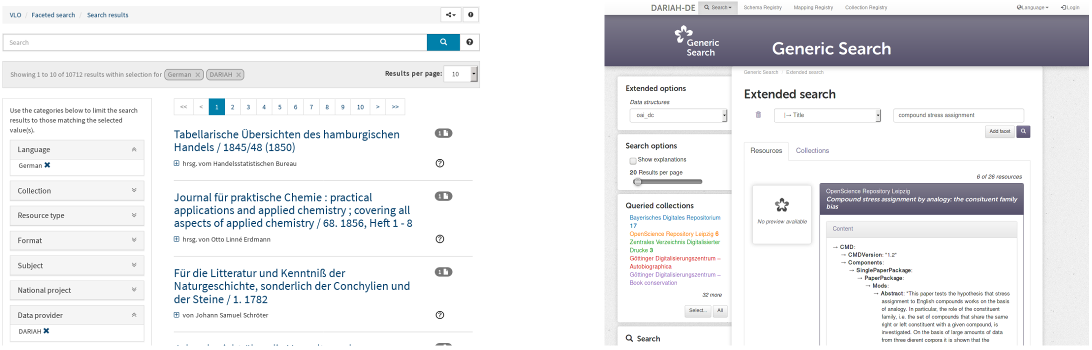
Importing CLARIN metadata in DARIAH-DE Employing DARIAH-DE's terminology, CLARIN centres and schemata form a *semantic cluster*, which—depending on particular users' perspective—can be decomposed in multiple partial clusters (e.g. corpora cluster, web services cluster). CMDI schemata represent collection- or domain-specific data models. By registering and associating data sources and schemata by means of the DFA, metadata become available in DARIAH-DE's Generic Search (see figure 2b). Extending or combining CLARIN's semantic clusters with external sources and data models, CLARIN metadata could be further contextualized by means of the DARIAH-DE infrastructure—without compromising the high data quality standards of CLARIN's metadata infrastructure.

The prototypical integration of CLARIN metadata in DARIAH-DE is currently focused on mappings of CMDI to Dublin Core and prove the conceptual and technical compatibility of the infrastructures with this respect. More differentiated and potentially beneficial representations require an intense analysis of respective collections, contexts and research questions: efforts, which might require additional resources and could result in interesting cooperative proposals.

Importing DARIAH-DE metadata in CLARIN From the perspective of CLARIN every OAI-PMH endpoint participating in DARIAH, may be seen as an additional resource center providing their metadata via this standard interface. Hence, the integration in standard working procedures (regular crawling of metadata inventory, metadata postprocessing, import in applications) was tested for a subset of those endpoints—including TextGrid, the Göttingen Digitisation Centre (GDZ), the OpenEdition project and more—for Dublin Core-based metadata.

It could be shown that the established workflow was operational without major adaptions in tools like CLARIN's metadata harvester or the VLO importer. As expected, the quality of presentation in interfaces like the VLO depends in parts on postprocessing and normalization of imported data. In that respect however, these new endpoints resemble many CLARIN centres where metadata curation has proven to be a helpful procedure for enhancing visibility and usability of provided records (Ostojic et al., 2016).

The results of the described importing procedures are illustrated in Figure 2: DARIAH-DE resources included in a test version of CLARIN's metadata search engine VLO (left) and CLARIN resources accessible in DARIAH-DE's Generic Search portal (right).



Showing 1 to 10 of 10712 results within selection for German ✕ DARIAH ✕

Tabellarische Übersichten des hamburgischen Handels / 1845/48 (1850)
hrsg. vom Handelsstatistischen Bureau

Journal für praktische Chemie : practical applications and applied chemistry ; covering all aspects of applied chemistry / 68. 1856, Heft 1 - 8
hrsg. von Otto Linné Erdmann

Für die Litteratur und Kenntniß der Naturgeschichte, sonderlich der Conchylien und der Steine / 1. 1782
von Johann Samuel Schröter

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

compound stress assignment

6 of 26 resources

OpenScience Repository Leipzig
Compound stress assignment by analogy: the consituent family bias

```
→ 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 dierent corpora it is shown that the
```

(a) DARIAH MD records in a test instance of the VLO (b) CLARIN MD records in DARIAH's Generic Search

Figure 2: Metadata records in the opposite project's search interface

5 Summary and further work

Despite following different attempts to build powerful and still user friendly metadata infrastructures in both projects, it could be shown that those approaches are in general compatible and able to collaborate with each other on the basis of already established infrastructure modules. Even though the described solution is still not part of the productive systems, the process has already identified minor problems and led to bug fixes and adjustments in several associated applications. The described process for the integration of metadata in both directions will be subject to further improvements. This contains the creation of more schema descriptions and mappings in the *Schema Registry* or adaptations in CLARIN's metadata curation process for records originated in DARIAH.

In the long term, both projects may benefit from each other's software components. This comprises the usage of additional search interfaces with their distinctive features in both infrastructures or using the *Mapping Registry* in CLARIN to overcome semantic ambiguity when evaluating underspecified CMDI profiles. For a tighter integration and mutual benefits, several adaptations seem to be possible: a (semi-)automatic import of metadata schemata between both schema registries could be a promising starting point.

The most pressing problem, however, seems not to be potential deficiencies in technical compatibility, but outstanding decisions about which resources are actually relevant for the opposite infrastructure, and which resources should be contained in productive applications and hence be available to the end users.

References

[Goosen and Eckart2014] Twan Goosen and Thomas Eckart. 2014. *Virtual Language Observatory 3.0: What's New?*. CLARIN Annual Conference 2014, Soesterberg, NL.

[Goosen et al.2015] Twan Goosen, Menzo Windhouwer, Oddrun Ohren, Axel Herold, Thomas Eckart, Matej Ďurčo and Oliver Schonefeld. 2015. *CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure.* Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, NL.

[Gradl and Henrich2015] 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.

[Gradl and Henrich2016a] Tobias Gradl and Andreas Henrich. 2016. *Die DARIAH-DE-Föderationsarchitektur – Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen.* BIBLIOTHEK – Forschung und Praxis 2016; 40(2): 222–228.

[Gradl and Henrich2016b] Tobias Gradl and Andreas Henrich. 2016. *Data Integration for the Arts and Humanities: A Language Theoretical Concept.* 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5–9, 2016, Proceedings: 281-293.

[Ostojic et al.2016] Davor Ostojic, Go Sugimoto and Matej Ďurčo. 2016. *Curation module in action - preliminary findings on VLO metadata quality.* CLARIN Annual Conference 2016, Aix-en-Provence, France.