Tobias Gradl, Universtität Bamberg, tobias.gradl@uni-bamberg.de
Timo Steyer, Forschungsverbund Marbach Weimar Wolfenbüttel, steyer@hab.de

A research-oriented and case-based data federation for the Humanities

A case study of DARIAH-DE and MWW

MWW stands for three institutions, united by more than 500 years of German and European ideas, literature and culture – and since 2013 by a research association.[1] The German Literature Archive Marbach, the Klassik Stiftung Weimar and the Herzog August Library Wolfenbüttel collect, preserve and provide access to sources that are crucial to the study of German literary and intellectual tradition. Thereby they have devoted themselves to the long-term transformation of cultural heritage in Germany. In all three institutions, digitization of their collections has been an important task to build digital collections and provide worldwide access to this cultural heritage. Furthermore, new approaches in the field of digital humanities have been adopted to strengthen the connection of digital collections with the humanities and cultural studies (Schreibmann et al. 2004). All of these goals can be reached easier by cooperation and the use of established tools and infrastructure, which will be demonstrated, based on the example of the implementation of a generic search over collections of all three institutions. For the implementation of the generic search the *Data Federation Architecture (DFA*) from DARIAH-DE is used.

Context and motivation

A significant share of the collections from the MWW institutions is already available in digitized form. For the optimized use of the collections, access and findability is a fundamental requirement. Furthermore the collections that are described in more detail by metadata have a greater potential for digital methods. The central argument is that specific research requires specific data. That is why the digitization of collections and their description with metadata is often accompanied by the development of an own web portals and databases. Some examples from the MWW institutions include:

- the research portal of the (historic) university of Helmstedt[2]
- the collection of letters from Johann Wolfgang von Goethe[3]
- the collection of funeral sermons[4]
- the compilation of members of the Fruchtbringende Gesellschaft[5]

In the context of funding research as well as technological development several problems arise. Most of the research projects are externally funded, which means that personnel and

[1] http://mww-forschung.de
[2] http://uni-helmstedt.hab.de
[3] https://ora-web.swkk.de/swk-db/goerep/index.html
[4] http://diglib.hab.de/edoc/ed000010/startx.htm

[5] http://www.die-fruchtbringende-gesellschaft.de/index.php?article_id=15

technical capacities are only available for a limited time. After the end of the project most of the researchers move on to new projects, often also to different institutions, and with them their expertise is lost. However, this expertise is needed to analyse and interpret the content of historical objects, documents or images. Even a well formulated documentation often cannot compensate this loss sufficiently. In the digital age, also the need for technical knowledge is crucial for the transformation of analogue information and the development of a research infrastructure for the presentation of data. But again, the resources are often only temporarily available. These two branches of knowledge and their dreaded loss culminate in the fact that after the establishment of the collection in a specific infrastructure and the end of a project a long term data curation cannot be always guaranteed.
Making matters even more complicated is the fast change of technology standards and applications. Some of the mentioned portals are older than ten years, the data is still relevant to scholarly research although the infrastructure is no longer state of the art. Because of the diverse backgrounds of their creation, their purposes and their (assumed) audiences, the collections presented in these portals are modelled in different ways, in different formats and database systems. Also, each institution has its own workflows, priorities and standards by which these projects are built and maintained. This makes the combination of diverse collections and their integration into one research system more difficult.

Data federation in DARIAH-DE

Despite their diverse disciplinary and organizational circumstances with particular purposes and focus domains, MWW collections show a particular intrinsic cohesion among their individual items. To provide unifying views of heterogeneous data, traditional approaches are often based on the harmonization of data within the constraints of a globally integrative data structure, such as a global schema or ontology. Following the established requirements of completeness, correctness, minimality and understandability (Batini et al. 1986), the contextual depth of such an integrative structure is reduced by an increasing contextual breadth.

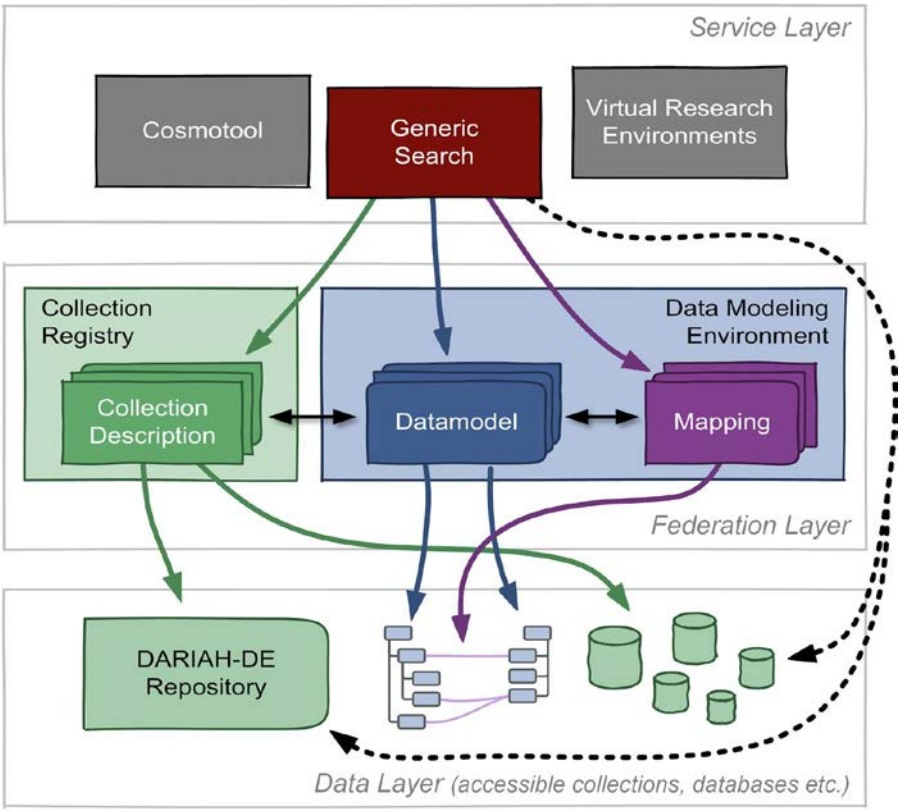
Inuse cases originating primarily from information systems in the economic domains, an extensive and often exclusive data analysis phase is a prerequisite/basic requirement for traditional data integration. This means that:

1. all relevant data sources and structures are identified, evaluated and specified
2. the global integration model is derived from a) the intersection of relevant concepts in the heterogeneous schemata and b) the information need of the focused use case or system
3. concepts of the target model are mapped from the source models
4. and data is transformed to conform to the constraints of the target model and thereby integrated.

Globally integrative scenarios are beneficial in cases that allow extensive a priori analyses executed by few experts of both the application domain and data integration. For use cases

in the arts and humanities the definition of globally unifying data models can be feasible if at least some context parameters are defined. Initiatives such as Europeana[6] or the German digital library[7] successfully follow a global model approach by defining a target system and focus group. This approach reduces the complexity of the required integrated view to which all local data models need to be mapped.

In contrast, data integration in the context of DARIAH-DE is not focused on a particular system or target audience, but instead employs the idea of a *research-oriented federation* (Gradl, Henrich 2016a) as a driving force for its *Data Federation Architecture (DFA)*[8]. The DFA consists of three layers (see below).



Opposed to traditional integration settings, fundamental theses for the conception of such a federation are:

Multiple domain experts are required to correctly analyse, interpret and describe the generative contexts of data. Due to the diversity of collections, i.e. collected items and organizational settings, a high degree of contextual knowledge is required.

Once local data has been modelled and enriched by domain experts, its transformation and integration requires in-depth knowledge of the application context. In the MWW initiative, this knowledge can be found at DH positions. These

[6] https://www.europeana.eu
[7] https://www.de.ddb.com
[8] https://de.dariah.eu/data-federation-architecture

DH-specialists conceptualize and design services such as the integrated search component by analysing target audiences and use cases that are facilitated by such services.

Separating technical and semantic tasks of data integration enables abstraction from repetitive and generic tasks – allowing domain experts to focus on conceptual tasks of data modelling and integration. Some tasks, such as accessing data through various interfaces, unarchiving compressed data, conversion of encodings and formats, the execution of processing pipelines (e. g. natural language processing to detect entities) etc., are often resolved on a per-project basis. The DFA implements such tasks in a generic fashion and hence separates contextual from technical aspects of data processing. Arts and humanities experts are left with tasks that require their attention and knowledge.

Applicability

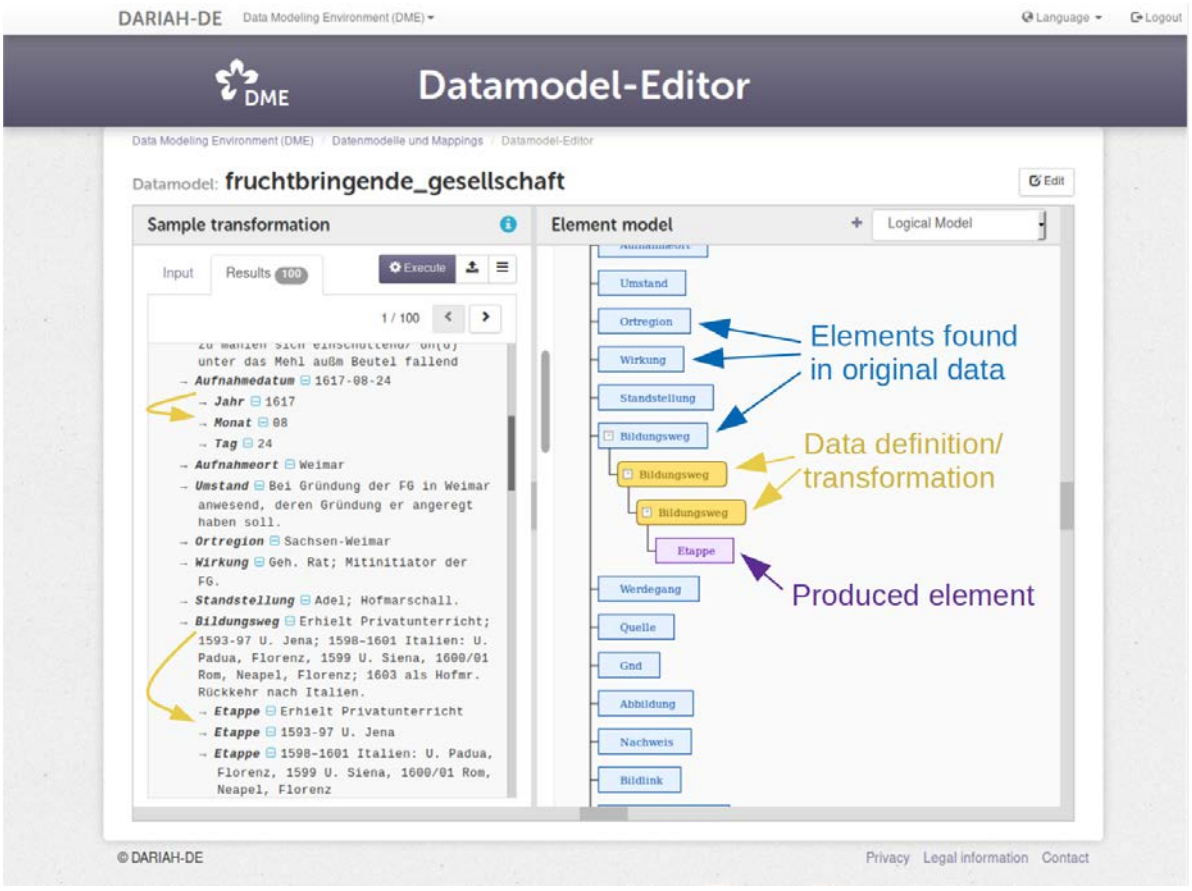
As part of our demonstration, we make use of the dataset *members of the Fruchtbringende Gesellschaft (FG)* of MWW, which provides XML-based data for members of the society. The exemplary entry below indicates the types of data that are provided by the feed (Ball et al. 2016).

```
<fg_mitglied>
   <nr>001</nr>
   <name>Teutleben</name>
   ...
   <aufnahmedatum>1617-08-24</aufnahmedatum>
   <aufnahmeort>Weimar</aufnahmeort>
   <umstand>Bei Gründung der FG in Weimar anwesend,
      deren Gründung er angeregt haben soll.</umstand>
   ...
   <bildungsweg>Erhielt Privatunterricht; 1593-97 U. Jena;
      1598–1601 Italien: U. Padua, Florenz, 1599 U. Siena,
      1600/01 Rom, Neapel, Florenz; 1603 als Hofmr.
      Rückkehr nach Italien.</bildungsweg>
   <werdegang>1608 Hofmr. der sachs.-weimar. Prinzen Friedrich
      (FG 4) und Wilhelm (FG 5), daneben seit 1611 Hofgerichtsassessor
      in Jena; 1613/14 Reisebegleiter Prinz Johann Ernsts d. J. von
      Sachsen-Weimar; ab 1616 Hofmarschall in Weimar; seit 1620 Geh.
      Rat in Coburg; seit 1621 auch weimar. Geh. Rat von Haus aus.
      </werdegang>
   ...
</fg_mitglied>
```

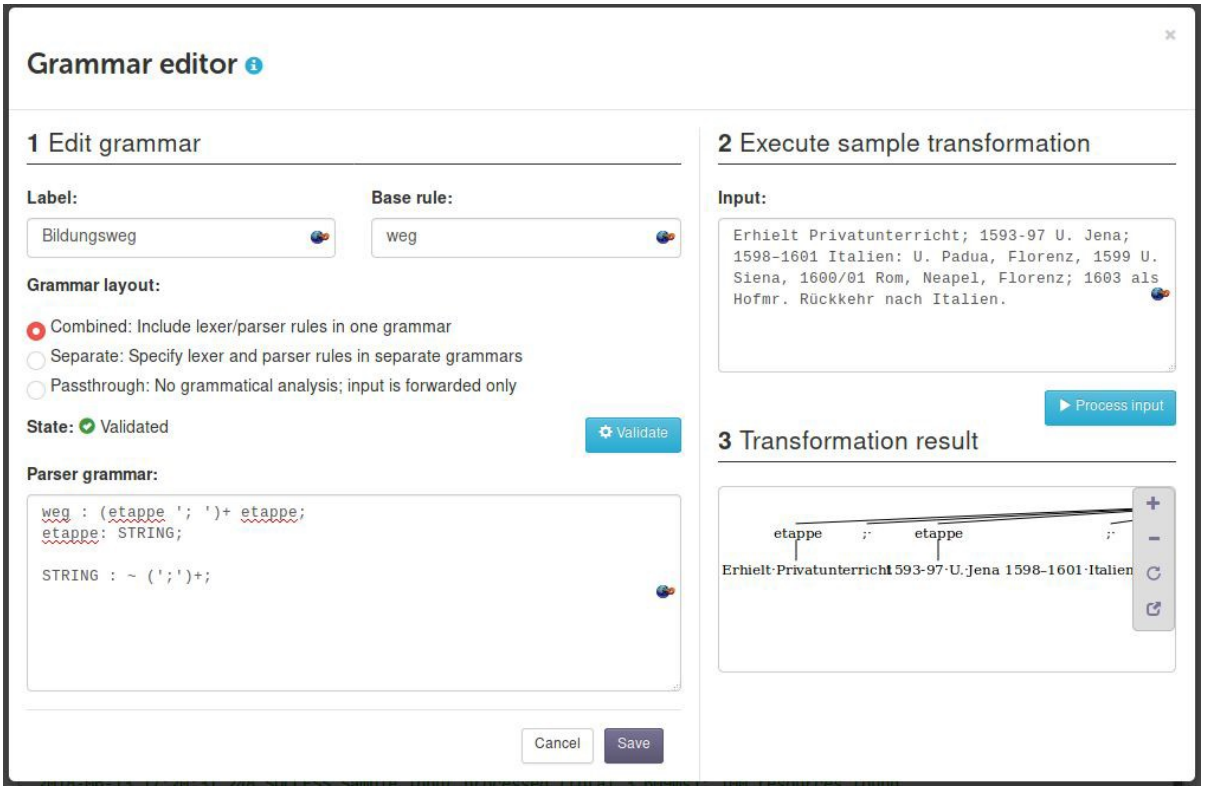
In order to utilize FG metadata outside of its original context and within a comprehensive MWW search, two important steps are carried out:

1. Contextual knowledge is modelled to enrich original metadata (Gradl, Henrich 2016b) and
2. the enriched data model is mapped to the integrative view of the MWW search.

Exemplary aspects of modelling FG metadata can be found at the *Aufnahmedatum* (admission date) and *Bildungsweg* (education) fields as indicated in the screenshot below.

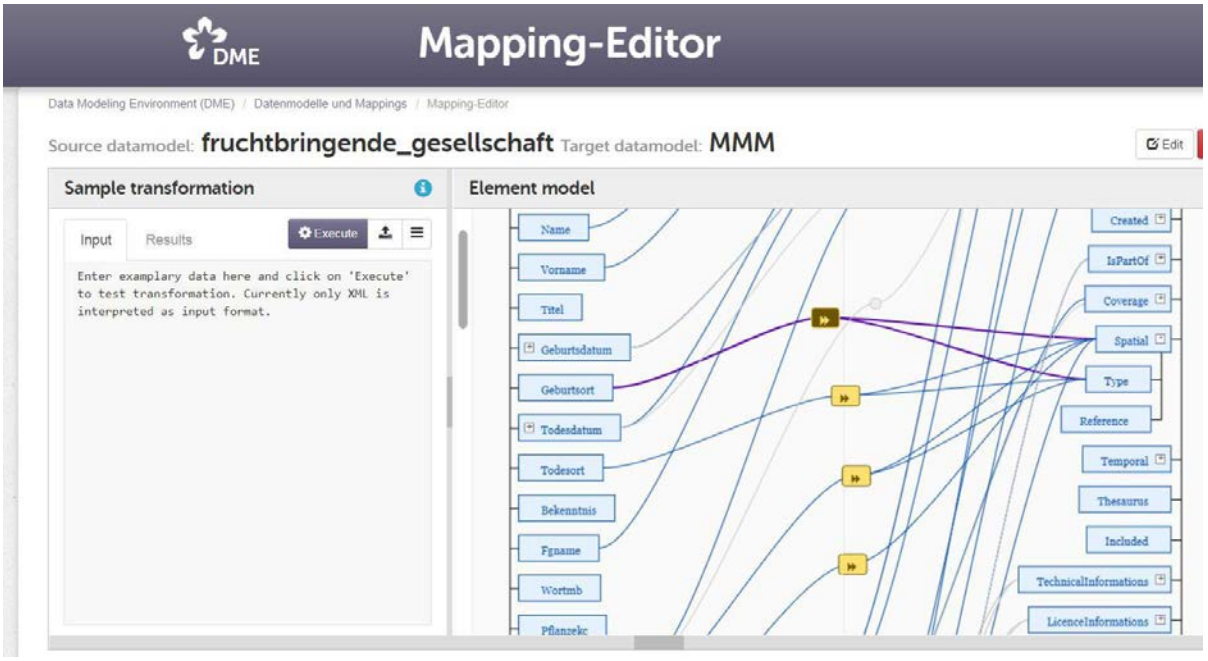


Rules in the form of domain specific languages and transformation expressions (Gradl, Henrich 2016c) are formulated to process the provided fields in order to enrich datasets.



Although the rules can be applied to the provided sample immediately, the generated data model functions as a ruleset that can be reapplied at convenience and by consuming services or other users. Hence, when indexing data, the MWW search applies these specifications to fresh data.

In order to finally prepare the unified view within the existing integration model of MWW, a mapping of the enriched FG model is developed. The screenshot below shows how the birthplace of a member of the FG is mapped to an equivalent element of the integrated model. This spatial element is also linked to the place of death because the integrated model describes a type and not a contextualized meaningful element.



Conclusion

The DARIAH-DE Data Modeling Environment allows domain experts to model data independent of technical difficulties such as conversions between encodings and protocols. Allowing for individual integrated models of individual projects, the DME has proven to be applicable for iterative modelling requirements of MWW and led to an initial version of an MWW generic search (see below), which can be dynamically extended as digitization efforts proceed and contextual knowledge is further explored.

Bibliography

Gabriele Ball, Anne Dickel, Andreas Herz, Timo Steyer (2016): Der gedruckten Edition eine digitale Schwester. Das AEDit-Projekt und die digitale Edition der Fruchtbringenden Gesellschaft. In: Denkströme. Journal der Sächsischen Akademie der Wissenschaften, 16, pp. 69-82.

C. Batini, M. Lenzerini, S.B. Navath (1986): A comparative analysis of methodologies for database schema integration. ACM Computing Surveys 18(4), pp. 323-364.

Tobias Gradl, Andreas Henrich (2016a): Die DARIAH-DE-Föderationsarchitektur – Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen. Bibliothek Forschung und Praxis, 40(2), pp. 222-228. Retrieved 13 Jun. 2018, from doi:10.1515/bfp-2016-0027.

Tobias Gradl, Andreas Henrich (2016b): Data Integration for the Arts and Humanities: A Language Theoretical Concept. Research and Advanced Technology for Digital Libraries 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, pp. 281-293. Retrieved 13 Jun. 2018, from https://doi.org/10.1007/978-3-319-43997-6_22.

Tobias Gradl, Andreas Henrich (2016c): Extending Data Models by Declaratively Specifying Contextual Knowledge, in: DocEng '16: Proceedings of the 2016 ACM Symposium on Document Engineering, pp. 123-126. Retrieved 8 Jun. 2018, from https://doi.org/10.1145/2960811.2967147.

Tobias Gradl, Andreas Henrich (2017): Explicating knowledge on data models through domain specific languages, in: INFORMATIK 2017. Gesellschaft für Informatik, Bonn, pp. 1125-1136. Retrieved 13 Jun. 2018, from https://doi.org/10.18420/in2017_114.

Schreibman, Susan, Siemens, Ray & Unsworth, John, eds. (2004): A Companion To Digital Humanities. Blackwell Publishers.