Secondary Publication

Gradl, Tobias; Henrich, Andreas

A novel approach for a reusable federation of research data within the arts and humanities

Date of secondary publication: 08.09.2023

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-904903

Primary publication

Gradl, Tobias; Henrich, Andreas (2014): „A novel approach for a reusable federation of research data within the arts and humanities“. In: Digital humanities 2014 : conference abstracts EPFL-UNIL Lausanne, Switzerland 8-12 July 2014, Lausanne, S. 382-384, doi: 10.5281/zenodo.58210.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holder(s).

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

A novel approach for a reusable federation of research data within the arts and humanities

Tobias Gradl and Andreas Henrich

Chair of Media Informatics,
University of Bamberg,
An der Weberei 5, D-96047 Bamberg
{tobias.gradl,andreas.henrich}@uni-bamberg.de

1 Introduction

In distributed systems literature the orthogonal but interdependent characteristics of *autonomy*, *distribution* and *heterogeneity* are used to classify distributed systems [1,2]. From a holistic perspective on the arts and humanities, collections have evolved over decades or centuries from highly autonomous disciplines and institutions and are widely spread, which resulted in heterogeneous perspectives and data models [3]. Despite its negative notion as *data integration problem*, the term *heterogeneity* also symbolizes the diversity of research methodologies within the disciplines of the arts and humanities. Resolving heterogeneity hence implies an abstraction from the specifics that are valuable for focused disciplinary and interdisciplinary research projects.

Our approach presents a novel concept for data federation in the arts and humanities, which focuses the needs of research projects as well as interdisciplinary and broad use-cases. We especially address the reusability of explicated knowledge on correlations between schemata and digital collections and show where domain experts are required to bridge semantic gaps.

2 Background

Approaches to data integration often follow the theoretical foundation expressed in [4] by employing the concept of a global view. While being highly distinctive in terms of their underlying concepts, established examples such as ISIDORE [5], OAIster [6] and Europeana [7] share the goal of facilitating access to a wide range of research data through integrated schemata or ontologies. Aside from broad services, an integration need that focuses on a specific topic and related research questions is addressed by the Steinheim Institute, which provides a search in the context of german-jewish history and judaism [8].

Despite usability concerns in having to identify relevant services and accordingly collections, the reappearing need to overcome the same aspects of heterogeneity in reaction to new use-cases is one of the problems we address.

2.1 Use cases

Our federation concept is primarily focused on the realization of the *DARIAH-DE Generic Search*[1], which includes support for queries over large sets of unrelated collections (broad search) and tightly correlated data (deep search)—with different information needs in mind:

- *Broad search*: Due to the quantity and distribution, the relevance of digital collections for particular research questions is not easily assessable. A broad view assists scholars in finding and evaluating possibly relevant data. Figure 1 shows an exemplary, collection-level aggregation of results based on term statistics in the prototype of our search, which will be continuously extended by other relevant visualization techniques (e.g. with respect to spacial and temporal aspects).
- *Deep search*: If the granularity of local data models can be used to formulate more specific queries targeting structure and content (e.g. in search facets), the deep search utilizes mappings specified in the *DARIAH-DE Schema Registry*. Broad search continuously fades to deep search with an increasing count and richness of mappings and hence typically smaller sets of selected collections.

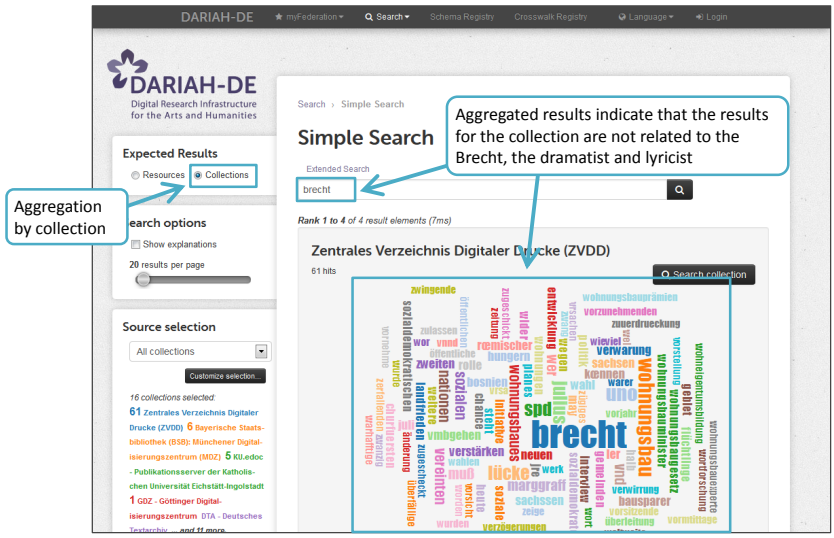

Fig. 1. Result aggregation in the generic search

Despite the focus on the generic search with its virtual integration at query-time, the proposed concept also addresses requirements of a materialized integration of data:

[1] `http://dev3.dariah.eu/search`

– *Data migration and consolidation*: Traditional applications of data integration often do not require a dynamic adaption to selected collections, but determine a set of relevant data sources and an appropriate integration schema or ontology [4]. Examples include data migration induced by the introduction of new information systems (e.g. replacement of outdated archive information software) or the consolidation of selected data sources under a merged schema for the purpose of interdisciplinary analysis and visualization e.g. in the *DARIAH-DE GeoBrowser* [9].

2.2 Problem Definition

The common objective of data integration approaches is to resolve heterogeneity on various levels: *Syntactical* aspects such as the existence of different access and encoding methods can be solved by technical means, whereas *structural and semantic* heterogeneity depend on the application of background knowledge [1]. Despite continuing efforts in the fields of schema and ontology matching, the manual intervention of domain experts—especially for large or complex schemata and ontologies often found in the arts and humanities—has shown to be essential to generate high-quality results [10].

The correlation of the used schemata and ontologies is an inherently complex manual task in our context, which depends on the *fragmented and distributed knowledge* of individual disciplines, collections and scholars. Requiring a common understanding, research projects concentrate knowledge about schemata and semantics used in relevant collections and specify meanings and correlations. In order to integrate the described data and establish technical interoperability, an application of digital methods and tools is required.

3 Concept

Abstracting from aspects of technical and syntactical heterogeneity concerned with accessing, preprocessing and integrating data in a generic fashion, we aim to enable researchers to focus on those aspects of integration, that depend on their knowledge and expertise: the description and correlation of schemata and ontologies. Despite the immediate benefit for individual integration tasks, the centralized formalization and explication of semantics results in the significant advantage of *knowledge reusability*.

3.1 Semantic clusters

The logical architecture of our idea is represented by a directed, weighed graph, where the schemata and ontologies are described by vertices, and mappings between them are symbolized by edges. Whereas correlations between structural elements symbolize a relation of the described concepts (e.g. persons, locations) and could be considered undirected, more specific rules that are required for data

transformation can be composed of non-reversible functions (e.g. the concatenation of fields). For that reason, parallel edges are required for the description of both mapping directions. Differences of schemata in terms of their complexity and expressiveness reduce the achievable level of accumulated completeness, which is represented by the value of *cohesion*.

Figure 2 indicates how the cohesion between schemata can be utilized to suggest semantic clusters: $C1$, $C2$ and $C3$ could be the result of research projects, which needed a high level of mapping completeness between relevant schemata. By interrelating clusters or generically used schemata ($S10$), the expressed semantics can be reused in other contexts.

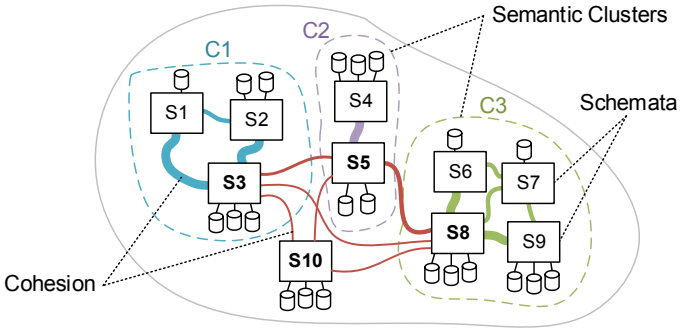

Fig. 2. Semantic clusters of schemata

3.2 Use-case orientation

Our example indicates the difference to the commonly found integration pattern of a global ontology or schema. Despite its theoretical foundation, simplicity and proven applicability for broad integration use-cases [5,6,7], we consider the approach to be impracticable for a holistic context of the arts and humanities because a global structure would either have to be an abstraction from collection or discipline specifics or unmanageably complex.

Narrowing this context to individual domains or research projects, standards could be elected as appropriate integrative structures. As exemplified in figure 2, the schemata $S3$, $S5$ and $S8$ form the integration baseline within our clusters due to their cohesion with other schemata. Considering our *deep search* and *data migration and consolidation* use-cases, these schemata can be utilized to generate a fine-grained view over selected collections accessible within the cluster. In order to support interdisciplinary use-cases, clusters can be combined (symbolized by the strong cohesion between $S5$ and $S8$) to resolve semantic gaps.

For broad use-cases we rely on the collaborative and continuous emergence of schemata or ontologies (compare $S10$) within our federation that are used to connect the clusters on the coarse levels sufficient for broad use-cases.

3.3 Scalability considerations

The simplicity of traditional data integration emerges as new local schemata are added to the system and hence an appropriate mapping target needs to be identified. To ensure extensibility and scalability, our proposed federation concept depends on two strategies:

Cluster globals: The concept of semantic clusters builds on the existence or advancement of standards that are considered as appropriate common perspectives by research communities. Although clusters are not predetermined but expected to evolve, established standards such as the *CIDOC Conceptual Reference Model (CIDOC CRM)* or the *Text Encoding Initiative (TEI) Guildelines* could be identified as initial cluster schemata, which can be mapped in a generic fashion [11]. As new schemata need to be added, the standard which promises to achieve the highest completeness is selected to be mapped.

Model inheritance: Our proposal includes an approach to specify the actual usage of schemata more precisely than it is possible at the level of generic crosswalks. Figure 3 shows the exemplary derivation of the Dublin Core element `dc:coverage` to resolve an encapsulated substructure. Mappings are inherited to correlate the refined elements or to specify detailed data transformation rules. As derived schemata are related to their parent, generic mappings remain valid and can be utilized if specific rules are missing.

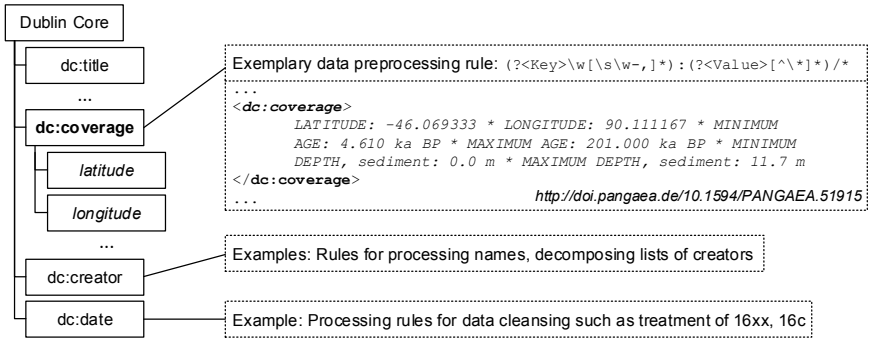

Fig. 3. Exemplary derived version of Dublin Core

4 Conclusion

As we abstract from technical aspects of heterogeneity and reuse the valuable disciplinary knowledge explicated in terms of correlations, processing and transformation rules, the efforts required for integrating research data can be significantly reduced. Another important aspect that is currently being evaluated

consists in appropriate techniques for the visualization of our federation concept and system. After all, domain experts need to be able to recognize clusters, important schemata and ontologies as well as their correlations in order to identify semantic gaps and to collaboratively fill them.

References

1. Sheth, A.P., Kashyap, V.: So Far (Schematically) yet So Near (Semantically). In: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), Amsterdam and The Netherlands and The Netherlands, North-Holland Publishing Co (1993) 283–312
2. Busse, S., Kutsche, R.D., Leser, U., Weber, H.: Federated Information Systems: Concepts, Terminology and Architectures (1999)
3. Henrich, A., Gradl, T.: DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities — Concepts and Perspectives. International Journal of Humanities and Arts Computing **7**(supplement) (2013) 47–58
4. Lenzerini, M.: Data Integration: A Theoretical Perspective. In Abiteboul, S., ed.: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, New York and NY, ACM (2002) 233
5. Pouyllau, S.: ISIDORE : acces to open data of arts & humanities (2011)
6. Hagedorn, K.: OAIster: a "no dead ends" OAI service provider. Library Hi Tech **21**(2) (2003) 170–181
7. Peroni, S., Tomasi, F., Vitali, F.: Reflecting on the Europeana Data Model. In Agosti, M., Esposito, F., Ferilli, S., Ferro, N., eds.: Digital Libraries and Archives. Volume 354 of Communications in Computer and Information Science. Springer Berlin Heidelberg, Berlin and Heidelberg (2013) 228–240
8. Lordick, H.: Vieles finden – die Suchmaschine im Steinheim-Institut (2013)
9. Romanello, M.: DARIAH Geo-browser: Exploring Data through Time and Space (2013)
10. Rahm, E.: Towards Large-Scale Schema and Ontology Matching. In Bellahsene, Z., Bonifati, A., Rahm, E., eds.: Schema Matching and Mapping. Springer Berlin Heidelberg, Berlin and Heidelberg (2011) 3–27
11. Baca, M., Harpring, P., Ward, J., Beecroft, A.: Metadata Standards Crosswalk (2009)