DARIAH(-DE): DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES — CONCEPTS AND PERSPECTIVES

ANDREAS HENRICH AND TOBIAS GRADL

Abstract *DARIAH (Digital Research Infrastructure for the Arts and Humanities) is part of the European Strategy on Research Infrastructures. Among 38 projects originally on this roadmap, DARIAH is one of two projects addressing social sciences and humanities. According to its self-conception and its political mandate DARIAH has the mission to enhance and support digitally-enabled research across the humanities and arts. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices. One main distinguishing aspect of DARIAH is that it is not focusing on one application domain but especially addresses the support of interdisciplinary research in the humanities and arts. The present paper first gives an overview on DARIAH as a whole and then focuses on the important aspect of technical, syntactic and semantic interoperability. Important aspects in this respect are metadata registries and crosswalk definitions allowing for meaningful cross-collection and inter-collection services and analysis.*

1. INTRODUCTION

DARIAH[1] (Digital Research Infrastructure for the Arts and Humanities) is part of the European Strategy on Research Infrastructures (ESFRI) and can be best described citing the ESFRI roadmap: ‘DARIAH aims to conceptualise and build an infrastructure in support of ICT-based research practices in the arts and humanities and to support researchers in the creation and use of research

data and tools. DARIAH connects information users (researchers), information managers and information providers, providing a technical framework that enables enhanced data sharing among research communities.'[2]

One background of this initiative is that ESFRI projects usually address research infrastructures for felds such as energy, biological, medical or physical sciences. The analogy to these felds makes clear that in the same way in which astronomers can beneft from research infrastructures such as extremely large telescopes, researchers in the arts and humanities need a digital infrastructure to bring together and collaboratively work with dispersed scholarly resources (e.g. digital content, services and methodologies). DARIAH aims to be such an infrastructure by promoting, supporting and advancing the use of digital content, tools and methods in research. The idea is not only to provide technology, but to connect people, information, tools and methodologies for investigating, exploring and supporting work across the broad spectrum of the humanities. And in the end, the creation of sustainable, distributed research Infrastructures will strengthen the European science landscape.

The timeline of DARIAH is as follows: The preparatory phase started in autumn 2008 and lasted about 2,5 years. Then beginning in March 2011, national projects in nine actively participating EU countries – such as Germany with its subproject DARIAH-DE – started. Hence, DARIAH-EU is currently in what is called the transition phase and this phase will gradually fade into the construction phase in 2012. The start of the so called operation phase is scheduled for 2016. The subprojects in the different countries are funded by the national funding bodies — the federal ministry of education and research in the case of DARIAH-DE.

In summary, the primary objectives of DARIAH are (1) to facilitate the application of digital techniques and methodologies within the arts and humanities in general and (2) to create an infrastructure that supports the federation of knowledge that is distributed among the numerous digital collections of museums, archives and libraries in Europe. To achieve these goals, DARIAH-EU concentrates its effort within four virtual competency centres (VCCs) – each addressing an individual perspective on the digital humanities:

VCC1 (e-Infrastructure) focuses the conceptualisation of a shared technological platform that facilitates the long-term preservation of research data, the federation of formally and pragmatically heterogeneous data sources and ultimately the establishment of services that enable an integrated analysis and retrieval of knowledge. However, DARIAH does not intend to integrate research data of the arts and humanities under a central schema or data store but especially respects the provenance of data and attempts a loose federation with related or collaborating projects.

VCC2 (Research and Education Liaison) intends to promote research and education in the arts and humanities by developing recommendations and curricula that prepare researchers and students for the use of digital

techniques – including those generated by DARIAH. VCC2 will hence create a profound understanding of the dynamic nature within the practices and methodologies of the arts and humanities and derive best-practices for the enrichment and sharing of knowledge in a sustainable manner.

VCC3 (Scholarly Content Management) has the Open-Access paradigm and standardization of content at its core and focuses the lifecycle of scholarly resources to establish best-practices for the generation and distribution of semantically enriched and reusable data and metadata.

VCC4 (Advocacy, Outreach and Impact) has a unifying perspective on the work packages in other VCCs focusing on the coordination of the resources and efforts within DARIAH.

The application domain of the DARIAH e-Infrastructure is constituted of the full spectrum of disciplines in the arts and humanities. In order to provide substantial services allowing scholars to utilise technology within the context of individual research questions the underlying infrastructure is being carefully designed to cope with traditional challenges of data integration while minimizing the necessity to change original data to make it machine-usable.

In the following this paper will provide an overview of current activities within *VCC1 e-Infrastructure* in the German subproject DARIAH-DE[3] – addressing data interoperability and the chances and challenges of comprehensive services. This paper discusses heterogeneity as the main obstacle for data integration and introduces central ideas for establishing interoperability while preventing information loss induced by the harmonization of collections or resources.

Due to recent research efforts, issues and conceptual aspects related to metadata interoperability as well as generic search facilities will be focused.

2. Problem Definition: Interoperability for Collections and Resources

Early attempts to utilise computers in the humanities originated from efforts to alleviate manual tasks in linguistic analysis. A prominent example is symbolised by the *Index Thomisticus*[4] which was intended to be an index of the 11 million medieval Latin words in the works of St. Thomas Aquinas and to provide text search on this corpus.[5] Starting in 1949 Father Roberto Busa began his work on the project and – with the assistance of Thomas J. Watson at IBM – created the frst concordance program based on punch cards. Along with the frst computer-assisted statistical evaluations on ancient text corpora and vocabularies, concordances were in the focus of further research in the 1960s. Since then, literary and linguistic analysis have been among the most prominent research areas of the digital humanities.[6]

Archives of the arts and humanities – and especially digital libraries and museums – are often characterised by detailed and structured object descriptions

which are the result of a mostly manual annotation by domain experts.[7] However, even with the high quality of metadata and the usage of standards, the metadata-based integration of digital libraries is not a trivial task. Heterogeneity arises e.g. from different perspectives on schemas and research data or the usage of different and less powerful standards that need to be aligned with domain-specifc schemas.

Interdisciplinary Research

As an example of interdisciplinary research, literary analysis of the inscriptions on archaeological fndings or epigraphic records can be considered: Epidat is an epigraphic database containing digital editions of 22.749 inscriptions on Jewish tombstones.[8] Assuming an integrated environment, the epigraphic data in Epidat could be analyzed in the context of digital libraries facilitating the translation of ancient inscriptions. Furthermore, comparative research on the usage of phrases in Jewish literature and on tombstones could lead to the identifcation of interesting correlations. There are similar scenarios for archaeological objects such as ancient buildings that could be examined and described on the knowledge base of digital libraries.

The interaction with end-users of the e-Infrastructure is realised by two classes of end-user services: *Generic services* implement functionality that should be usable by a wide range of felds in the arts and humanities while *specifc services* are focusing on individual disciplines, organizations or research groups. DARIAH-DE does not plan to realise all of the specifc end-user services that are needed in individual research contexts, but intends to create an easily extendable platform for services that are developed within the disciplines of the arts and humanities – ensuring applicability to the circumstances of specifc research methodologies and questions.

DARIAH-DE addresses the requirements of generic and specifc services with a layered approach: The *federation layer* encapsulates functionality to overcome the heterogeneity of the disciplines and their data, whereas the *service layer* provides both generic and specifc services to the end users in the arts and humanities.

The possibility of future extensions to DARIAH-DE results in two major implications for the federation of digital collections and the concepts and implementations of technical solutions concerning the interoperability of research data. The e-Infrastructure need to consider both:

(1) *generic questions and tasks* such as search queries that span a wide range of collections and resources

(2) *specific questions* that are limited to a related subset of the available collections and deeply analyse research data and its descriptive metadata.

Metadata Schemas

The landscape of digital archives, libraries and museums is characterised by a variety of existing and utilised metadata-schemas, repository software and best-practices. In the context of digital libraries, the TEI Guidelines[9] represent a widely adopted schema for the annotation of texts.[10] Multiple of its derivations such as TEI lite, the Music Encoding Initiative (MEI)[11] and Epigraphic Documents in TEI XML (EpiDoc)[12] have been adopted in their respective felds. The TEI Guidelines – due to their semantic richness – can be considered as an integrative schema for the description of linguistic digital items. A possible collaboration scenario of digital libraries can build upon the TEI Guidelines being an extensive alignment target for other standards commonly used in the feld [13] – such as METS[14], Dublin Core[15] and MODS[16].

However, the application domain of DARIAH-DE is not limited to libraries and the research in a language-focused context, but intends to support other disciplines whose items might not be best described by the TEI Guidelines, but by standards such as the Archaeological DataeXport-Standard (ADeX)[17] for archaeological metadata or Lightweight Information Describing Objects (LIDO)[18] for the description of Cultural Heritage objects. Combining research data from multiple disciplines then implies that not only schemas but perspectives on individual items need to be matched: One particular instance of a book for example can be both an archaeological fnding and an entry in a digital library.

Key aspects of the generic collection federation within the DARIAH-DE initiative can be summarised as follows:

– *Multiple schemas*: Research data is described within different structures, that DARIAH-DE wants to support with generic methods of collection and data federation.
– *Context knowledge*: The description of a collected item is produced with the implicit background knowledge of a scholar, discipline, organization and collection which is typically not explicated in the metadata of the object. Within a federated environment, the context of research data needs to be explicated for further interpretation.[19]
– *Schema ambiguities*: There are ambiguities within the felds of metadata schemas: On a *syntactical* level, the formats for e.g. name and date representations may vary. *Semantic* heterogeneity occurs if felds are flled

with different interpretations of the suggested content. If for instance two collections use Dublin Core to describe cultural heritage objects, the feld *dc:creator* could be used to refer to the creator of the real-world object or the creator of the record.

– *Support for specific services*: DARIAH-DE aims to create a shared infrastructure, that can easily be extended by domain- or problem-specifc services. In order to minimise information loss, the federation needs to respect research data in its original form and allow a query-time transformation into the schema, ontology or vocabulary required two answer a research question.

The central idea for federation in DARIAH-DE is based on crosswalks, which are based on the state-of-the-art in schema and ontology matching.[20] Individual mappings between (groups of) felds are identifed and modelled on the schema level. The analysis of semantic correlations can be based on information on the schemas, instances or external sources such as thesauri, ontologies and reference data.

Access to digital collections

Another level of heterogeneity that needs to be addressed by the interoperability architecture consists in the different access strategies and protocols offered by the digital collections. The loose coupling concept of the DARIAH-DE federation implies that archives are not obliged to implement an application programming interface (API) that DARIAH-DE specifes but that the federation layer adapts to the available APIs of the collections.

The Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)[21] describes an access protocol that is commonly offered by the archives of the digital humanities. Despite the popularity of OAI-PMH, a sustainable concept for the federation infrastructure of DARIAH-DE needs to be adaptive to multiple access standards.[22] Learning from the experiences of an integration project of the National Science Digital Library (NSDL), a common implementation of the rather simple standards of OAI-PMH and Dublin Core by every attachable collection cannot be assumed. While the NSDL was able to specify requirements for considered archives, an comprehensive assessment of the organizations and collections that could provide data through DARIAH is infeasible due to the possible amount of data sources within the digital humanities in Europe.[23]

On a conceptual level, collections that might be accessible to DARIAH-DE can be classifed according to the access strategy:

– *Harvesting*: Metadata can be polled, indexed and locally supplied for information retrieval techniques

– *Querying*: Collections that provide access in the form of search APIs
– *No API*: The data of an archive cannot be retrieved through a dedicated API but rather by crawlers

The facets of interoperability in DARIAH-DE are not limited to the schemas, disciplines and access protocols. Multimodality and multilingualism are just two examples of further heterogeneity levels that need to be addressed by the shared federation layer. However, the discussion of heterogeneity in this section focused on the levels that are most relevant to the concepts of collection federation and generic search, which will be introduced in the next section of this paper.

3. Concepts and Perspectives: Federation and Generic Search

The DARIAH-DE federation architecture aims at the minimization of the information loss by acknowledging and supporting the diversity of the disciplines within the digital humanities, their research methodologies and their data. Mechanisms to realise this goal are being implemented in the components of the federation layer allowing for a fe xible defnition of associations between schemas.

From a bottom-up perspective on the architecture layers, the entry point to the DARIAH-DE architecture is realised by the APIs that encapsulate the physical access to the digital collections and provide a uniform logical access layer.

Functionality concerned with data interoperability is then provided by the two registries of the infrastructure:

(1) The *Collection Registry* comprises metadata on the collection level. On the basis of the Dublin Core Collections Application Profle (DCCAP)[24], the DARIAH-DE Collection Level Description (CLD) intends both to maintain the interoperability of DCCAP and to provide an extension that is currently being developed in respect of the specifc needs of the arts and Humanities.
(2) The *Schema Registry* contains the schemas that are required for the interpretation of the data contained in any of the collections listed in the Collection Registry. Correlations between registered schemas are specifed as crosswalks that are managed by the Schema Registry as well.

Both registries provide user interfaces for the management of their data and APIs that make them utilizable for the service layer.

Schema Matching and Mapping

The specifcation of machine-readable associations in schemas and collections is a common integration problem that is addressed by the mechanisms of schema and ontology alignment. Current concepts and prototypes in this feld focus on interactive techniques and algorithms that support the defnition of relations between schemas and ontologies. Typical characteristics of current concepts and prototypes are:[25]

- *Semi-automatic techniques* are preferred over strictly algorithmic or manual schema matching and mapping. Various challenges of either manual (e.g. identifcation of relevant schemas or sections within schemas) or automatic (e.g. missing background knowledge, algorithm selection and combination) matching and mapping can be alleviated by the combination of techniques.[26]
- Different *classes of algorithms* analyse terminological and structural features or correlations with external sources.[27] Current algorithms typically focus on the schema level.[28] Semantic correlations are inferred from the combination of individual calculation results and annotated with confdence values and further metadata on the association.
- Various *User-interaction techniques* support manual schema alignment.[29] Among Graph Visualization and Treemaps which – in the context of Schema and Ontology alignment – have not yet found wide adoption in available systems, Linked Tree Widgets seem to be an intuitive representation of schema elements and their associations.[30]

The state-of-the-art in schema and ontology alignment provides functionality that is required for the interoperability layer of DARIAH-DE. However, the process of identifying relevant schemas, collections and their associations is not limited to the core schema matching and mapping task. Figure 1 sketches user-oriented tasks related to the schema and crosswalk management in DARIAH-DE as well as the components that are required.

Considering the amount of the possibly federated collections and schemas, domain-experts that are involved in the continuous process of the generation and adaption of crosswalks need to be assisted in identifying the relevant collections, supported schemas and possible associations with existing entries in the Schema and Collection Registries.

Generic Search

Collections, Schemas and Crosswalks are the fundamental building blocks of DARIAH-DE to support interdisciplinary research in the digital humanities. The effcienc y of services that are based on the DARIAH-DE federation layer highly correlates with the quality of the metadata stored in both the collection and

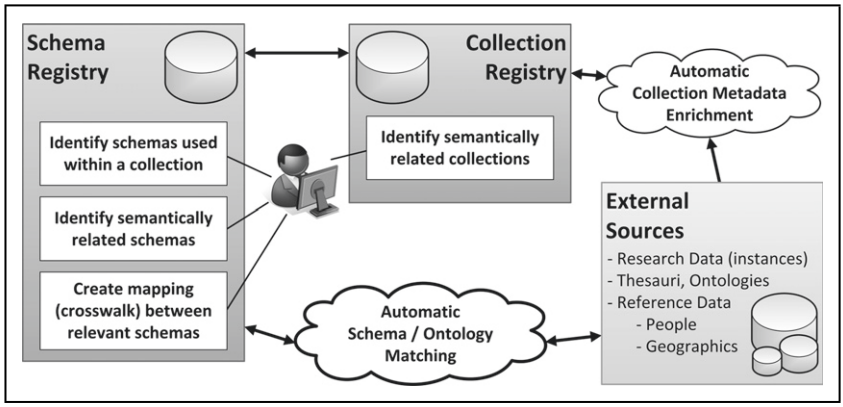

Figure 1. Schema and Crosswalk Modeling.

schema registry. In particular, the task of answering research questions is based on data that is distributed among multiple collections or disciplines is dependent on high quality crosswalks.

With a generic search component DARIAH-DE intends to enable the scholar in the digital humanities to perform a based search on data that is available to the DARIAH-DE federation. Figure 2 indicates our current concept of a metasearch that suits the requirements of a loosely coupled collection federation. On the basis of a distributed information retrieval architecture, the *Metasearch Component* delegates queries to the available search APIs of the collections (Figure 2). Collection *X* represents a Fedora-based archive that provides access to such a search API. In contrast, the collections *F* and *W* do not implement a distributed search API but allow access using the harvesting protocol OAI-PMH. In the latter case a collection indexer component hides the complexity of harvesting collections, indexing the available metadata and ensuring the freshness of the index.

The usage of the metasearch component by a user-oriented interface is inspired by a methodology that is usable both for breadth-frst and depth-frst search and provides a shared interface for accessing the research data of the digital humanities:

(1) The *Search Interface* interacts with the *Schema* and *Collection Registry*:

   a. Researchers identify the collections that might contain relevant data; the Collection Registry offers basic search functionality to support (relevance-driven) source selection by the user.
   b. Based on the collections that the user selected, the Search Interface interacts with the Schema Registry to identify the depth of crosswalks

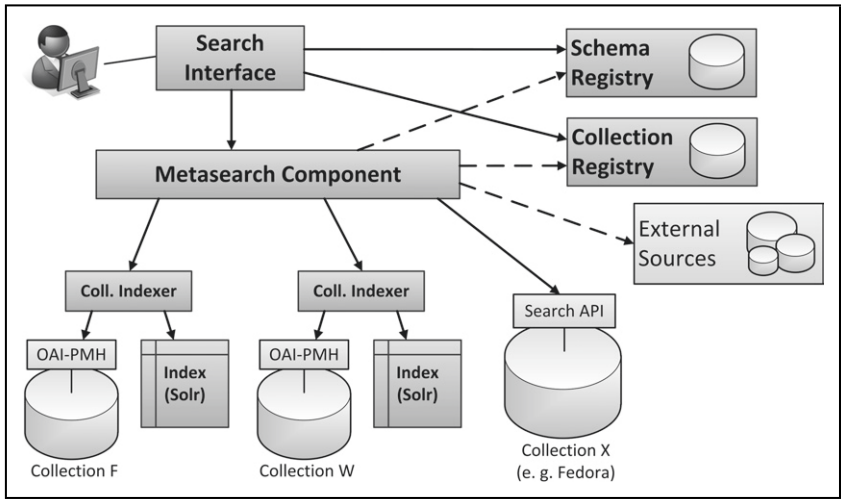

Figure 2. Generic Search in DARIAH-DE.

between those collections, allowing the user to adjust facets of the search.

(2) The *Search Interface* assigns the parameters to the *Metasearch Component* which executes a distributed query, answered either by the available search APIs provided by collections or the locally hosted indices.

By providing both breadth- and depth-oriented search, the generic search can assist the researcher of the digital humanities in fnding relevant research data and collections.

CONCLUSION

This paper focused on interoperability aspects for the federation of digital collections in the digital humanities and the concept of a generic search that uses and supports crosswalks. Interoperability and generic services are two of many important facets that the DARIAH-DE initiative addresses: The tempting perspective of a unifed access to the research data of multiple collections and disciplines – while preserving the semantics of data and minimizing information loss opens possibilities for the generation and answering of new and interdisciplinary research questions.

In contrast to the thematic focus of this paper, the DARIAH e-Infrastructure is not limited to discussions of interoperability and information retrieval. Current research includes topics such as the Authentication and Authorisation

Infrastructure (AAI), the long-term preservation of research data and the establishment of a shared developer portal to supply a complete and comprehensive toolset to the digitally supported research in the arts and humanities.

END NOTES

[1] http://dariah.eu/, last accessed 23 02 2012.

[2] European Strategy Forum on Research Infrastructures (ESFRI), Strategy Report on Research Infrastructures - Roadmap 2010 (Luxembourg, 2011), 27.

[3] http://de.dariah.eu/, last accessed 23 02 2012.

[4] Index Thomisticus, http://www.corpusthomisticum.org/it/index.age, last accessed 23 02 2012.

[5] T. N. Winter, 'Roberto Busa, S.J. and the Invention of the Machine-Generated Concordance.' (1999), http://digitalcommons.unl.edu/classicsfacpub/70/, last accessed on 23 02 2012, 3–20.

[6] S. Hockey, 'The History of Humanities Computing' in S. Schreibman, R. G. Siemens and J. Unsworth, eds., A companion to digital humanities (Malden, Mass, 2004), 1–19.

[7] A. Isaac et al., 'Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies', Library Review (2008), 187–199.

[8] epidat - Steinheim-Institute, http://steinheim-institut.de/cgi-bin/epidat, last accessed 23 02 2012.

[9] Text Encoding Initiative (TEI) Guidelines, http://www.tei-c.org/Guidelines/, last accessed 23 02 2012.

[10] Hockey, 'The History of Humanities Computing' in S. Schreibman, R. G. Siemens and J. Unsworth, eds., *A companion to digital humanities* (Malden, Mass, 2004), 1–19.

[11] Music Encoding Initiative (MEI), http://music-encoding.org/, last accessed 23 02 2012.

[12] EpiDoc: Epigraphic Documents in TEI XML, http://epidoc.sourceforge.net/, last accessed 23 02 2012.

[13] R. Gartner, Metadata for digital libraries: state of the art and future directions (2008), http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf, last accessed 23 02 2012; C. L. Palmer, O. L. Zavalina and M. Mustafoff, 'Trends in metadata practices: a longitudinal study of collection federation.' Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07) (New York, NY, 2007), 386–395.

[14] Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets/, last accessed 23 02 2012.

[15] Dublin Core Metadata Element Set, Version 1.1, http://dublincore.org/documents/dces/, last accessed 23 02 2012.

[16] Metadata Object Description Schema (MODS), http://www.loc.gov/standards/mods/, last accessed 23 02 2012.

[17] Archaeological DataeXport-Standard, http://www.landesarchaeologen.de/verband/kommissionen/archaeologie-und-informationssysteme/projektearbeitsgruppen/adex/, last accessed 23 02 2012.

[18] Lightweight Information Describing Objects (LIDO), http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/lido-technical/specifcation.html, last accessed 23 02 2012.

[19] Palmer, Zavalina and Mustafoff, 'Trends in metadata practices: a longitudinal study of collection federation.' *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07)* (New York, NY, 2007), 386–395; M. Foulonneau et al., 'Using collection descriptions to enhance an aggregation of harvested item-level metadata', Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL '05) (New York, 2005), 32–41.

[20] M. Granitzer et al., 'Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques', Future Internet 2010, 3 (2010), 238–258; P. Shvaiko and J. Euzenat, 'A Survey of Schema-Based Matching Approaches', in D. Hutchison, ed., *Lecture Notes in Computer Science* (Berlin, Heidelberg, 2005), 146–171.

[21] The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html, last accessed 23 02 2012.

[22] M. Brogan, Contexts and Contributions: Building the Distributed Library (Washington, 2006).

[23] C. Lagoze et al., 'Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience' Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL '06) (New York, NY, 2005), 230–239.

[24] DCMI, Dublin Core Collections Application Profle (2007) http://dublincore.org/groups/collections/collection-application-profle/ last accessed 23 02 2012.

[25] S. M. Falconer and N. F. Noy, 'Interactive Techniques to Support Ontology Matching', in Z. Bellahsene, A. Bonifati and E. Rahm, eds., Schema Matching and Mapping (Berlin, Heidelberg, 2011), 29–51; M. Granitzer et al., 'Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques', Future Internet 2010, 3 (2010), 238–258; P. Shvaiko and J. Euzenat, 'A Survey of Schema-Based Matching Approaches', in D. Hutchison, ed., Lecture Notes in Computer Science (Berlin, Heidelberg, 2005), 146–171.

[26] P. Shvaiko and J. Euzenat. 'Ten Challenges for Ontology Matching.', in R. Meersman and Z. Tari, eds., Lecture Notes in Computer Science (Berlin, Heidelberg, 2008), 1164–1182.

[27] Shvaiko and Euzenat, 'A Survey of Schema-Based Matching Approaches', in D. Hutchison, ed., Lecture Notes in Computer Science (Berlin, Heidelberg, 2005), 146–171.

[28] E. Rahm, 'Towards Large-Scale Schema and Ontology Matching', in Z. Bellahsene, A. Bonifati and E. Rahm, eds., Schema Matching and Mapping (Berlin, Heidelberg, 2011), 3–27.

[29] Falconer and Noy, 'Interactive Techniques to Support Ontology Matching' in Z. Bellahsene, A. Bonifati and E. Rahm, eds., *Schema Matching and Mapping* (Berlin, Heidelberg, 2011), 29–51.

[30] Granitzer et al., 'Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques', 238–258.