Towards a DH Knowledge Hub - Step 1: Vocabularies

Matej Ďurčo, Karlheinz Mörth
ICLTT/ACDH-ÖAW
Vienna
E-mail: matej.durco@oeaw.ac.at, karlheinz.moerth@oeaw.ac.at

Keywords: metadata, reference data, controlled vocabularies, interoperability

1. Introduction

We describe a tentative architecture for an integrated online platform for data and knowledge management. It is anchored in the Austrian CLARIN and DARIAH activities and will cater to the needs of the Digital Humanities community in general.

The main goal is to achieve a tight integration of descriptive metadata, reference data (vocabularies, taxonomies) and bibliographic information. The higher integration level shall lead to higher quality of data (curation) and to more consistent user experience when exploring the data.

At the core of the data model, an open set of controlled vocabularies shall act as pivotal point for interlinking disparate but in certain aspects related data sets.

Once completed, the system is supposed to work as a large semantic knowledge base aggregating metadata from a broad spectrum of providers, enriching and mapping this information and offering it through a number of visualization and analysis tools. Likewise, this work is meant as a test-bed for the development of core technical infrastructures for CLARIN.

2. Context

The work is being carried out at the Austrian Centre for Digital Humanities / Digital Humanities Austria (DHA) that started in January 2014 as an initiative to bring together several research infrastructure activities in Austria. The system shall be implemented at and hosted by this centre which is based at the Austrian Academy of Sciences and relies largely on components and resources made available within CLARIN and DARIAH.

The proposed system can be also seen as a technical incarnation of the ongoing effort on metadata curation, coordinated by the SCCTC[1]; it profits from the support for controlled vocabularies added in the new CMD 1.2 version; it relies on the ongoing efforts to convert CMD into RDF (Durco, 2013; Durco, Windhouwer, 2014).

3. State of the Art

There exists a sizeable number of solutions and resources already available covering individual aspects of the proposed system. Regarding metadata aggregators and catalogs, next to CLARIN's well-known VLO (Van Uytvanck, 2010) fed via the CLARIN OAI-harvester, there is the RDF-based portal rechercheisidore[2] (run by the French HumaNum group) or DARIAH-DE's Collection Registry[3] . Within DASISH, another research infrastructure project, a prototypical joint metadata domain has been established[4], harvesting metadata from content providers in CLARIN, DARIAH and CESSDA. Finally there is Europeana with its vast network of national and thematic aggregators which has accumulated huge amounts of resource descriptions over the past few years[5] (Purday, 2013).

In addition, there is a large amount of all kinds of reference data, starting from huge authority files produced by national libraries (GND, LCCN, Getty Thesauri, etc.), classification systems (DDC, LCSH), ISO standardized controlled vocabularies (ISO 639-* for languages, ISO 15924 for writing systems, ISO 3166 for countries), down to domain-specific taxonomies and gazetteers (e.g., Taxonomy of Digital Research Tools DiRT [6] or the gazetteer of historical places Pleiades[7]). See also (Durco, 2013, p. 28-31) for a comprehensive overview of existing vocabularies and reference data.

However all these different data sets are available in rather diverging representations, being accessible through browser-based user interfaces, idiosyncratic web services or even just as static data dumps. To foster the use of reference data in user applications, the access to these disparate resources needs to be harmonized. The issue is furthermore complicated as many of these data sets cannot be easily amended, being in the authority of big institutions like the German National Library or the Library of Congress. All of this calls for more ideally complementary custom vocabularies that can be edited more easily, but would be still available via a common interface.

Finally, we encounter also a growing number of large semantic knowledge bases integrating structured information from a number of sources, the main building block being dbpedia, an RDF-ized version of Wikipedia, that serves as central node of the LOD cloud[8]. Important examples are: YAGO[9] (10 Mio. entities derived from

[1] Standing Committee on CLARIN Technical Centres

[2] http://rechercheisidore.fr
[3] http://demo2.dariah.eu/colreg/colreg/main
[4] http://vmext24-215.gwdg.de/ckan/dataset?groups=dariah
[5] http://www.pro.europeana.eu/web/guest/content
[6] Digital Research Tools http://dirt.projectbamboo.org/
[7] http://pleiades.stoa.org/
[8] http://lod-cloud.net/
[9] http://yago-knowledge.org/

Wikipedia, WordNet and GeoNames), BabelNet [10] (multi-lingual dictionary with 50 languages and 9 mio.synsets) or UBY [11] - a "Large-Scale Unified Lexical-Semantic Resource". LT-World[12] developed and provided by DFKI is not as large as those mentioned before. However, what makes it a particularly interesting resource is its thematic focus and the underlying ontology It provides comprehensive information not only about digital language resources, but also about organizations, persons, publications and events in the world of language technology.

4. Data

For our purposes, we distinguish two basic types of data processed by the system: descriptive metadata and reference data (vocabularies, taxonomies, authority files, ontologies, ...).

By descriptive metadata we understand not only the resource descriptions as they are encountered within the CLARIN joint metadata domain (the CMD records), but also non-CMDI metadata, both in terms of context (e.g. material from DARIAH-related projects) and type (especially we also plan to include bibliographic records). This approach is implied by the fact that it is impossible to draw a clear line between research data and research results AND the fact that many content providers do not differentiate between these types of material anyhow. So rather than trying to filter out the resource descriptions in a narrower sense, we choose to retrieve all that we can attain and try to handle and sort it out inside the system, the basic premise being to offer as much as possible along with appropriate filtering mechanisms.

Next to the obvious choice, CLARIN's harvester, another possible starting point or at least source of inspiration for collecting the metadata is the tentative joint metadata domain set up in DASISH, especially those providers that have been identified in the context of DARIAH (as there is no official DARIAH-EU wide setup for collecting metadata from the partners so far). The analysis of the records exposed by the "DARIAH providers" shows – besides the usual common dublincore baseline – a number of specialized, partly quite elaborate formats for metadata, offering a nice showcase for mapping metadata formats outside CLARIN's CMD universe.

With respect to reference data, the starting point is the initiative CLAVAS (Brugman, Lindeman, 2012) which provides a number of vocabularies via a dedicated instance [13] of the open source vocabulary repository *OpenSKOS*[14] hosted by the Meertens Institute. Guided by the specific CLARIN needs, CLAVAS currently exposes the following vocabularies: a list of language codes (the ISO 639-3 standard converted to SKOS), a number of closed data categories from the data category registry ISOcat and a list of organization names that has been compiled out of the CMD metadata. It is important to keep in mind though that there exist multiple instances of OpenSKOS operated by different institutions offering a range of taxonomies, e.g. one managed by the Netherland Institute for Sound and Vision [15]. All of them being available via the same system, these reference data can be used by our system at no additional expense.

Within the CLARIN-DARIAH-AT consortium, the Austrian Audiovisual Research Archive (Austrian Academy of Sciences) has committed to curate its internally used taxonomies (musical instruments, languages and language variants, geographical reference data) and make them publicly available in SKOS format. This data will be integrated into the ACDH instance of OpenSKOS.

Another dataset to be included is the Taxonomy of Digital Research Activities in the Humanities or TaDiRAH [16] which was developed within the DARIAH community on the basis of previous work in projects like NeDiMAH and Bamboo's DiRT taxonomy. It is already being used in the DH course registry[17] and in the bibliography collection on DH: *Doing digital humanities - a DARIAH bibliography*[18]. The above mentioned vocabularies are only a starting point, the system will have to be open to new datasets, thus establishing an ever growing pool of reference data to be used internally and externally, both by applications and users. When adding new vocabularies, the focus lies on curation-intensive data such as for instance organization names.

Both the metadata and reference data as described above constitute the input for the system. Internally, a workflow is planned, where in a first step, the curation phase, the mostly XML-based data will be processed, analyzed and enriched. This is done in preparation of the next step, the conversion of the data into an RDF representation. An RDF based data model allows for a high degree of flexibility in (re-)modelling the descriptive information and makes it comparatively easy to link between the metadata and reference data. We need to keep in mind though that RDF in itself is not a universal remedy for all interoperability problems, but rather just another form of information representation. Still it at least offers a common widely adopted syntactic denominator. The critical aspect when modelling data along the LOD paradigm is the selection and reuse of existing semantic resources. In this preliminary stage, we foresee the use of SKOS, VoiD, FOAF, DOAP, Biblioontology, ORE, OpenAnnotation. However, this list is by no means exhaustive. Based on the analysis of the input data, further resources will have to be taken into account in due course[19].

5. System Architecture

The proposed system has to allow for publication and discovery (i.e. search/browse) of descriptive metadata (for resources and publications) and of reference data like vocabularies or taxonomies. The functionality we want to provide for comprises the following basic components: metadata aggregator (harvester), vocabulary repository, indexing engine and a catalogue rich in features as a web

[10] http://babelnet.org/
[11] www.ukp.tu-darmstadt.de/**uby**/
[12] http://lt-world.org/
[13] https://openskos.meertens.knaw.nl/
[14] http://openskos.org

[15] http://openskos.beeldengeluid.nl/
[16] https://github.com/dhtaxonomy/TaDiRAH/
[17] http://dhcoursereg.hki.uni-koeln.de/
[18] https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography
[19] See also LOV for a comprehensive overview of linked open vocabularies: http://lov.okfn.org/dataset/lov

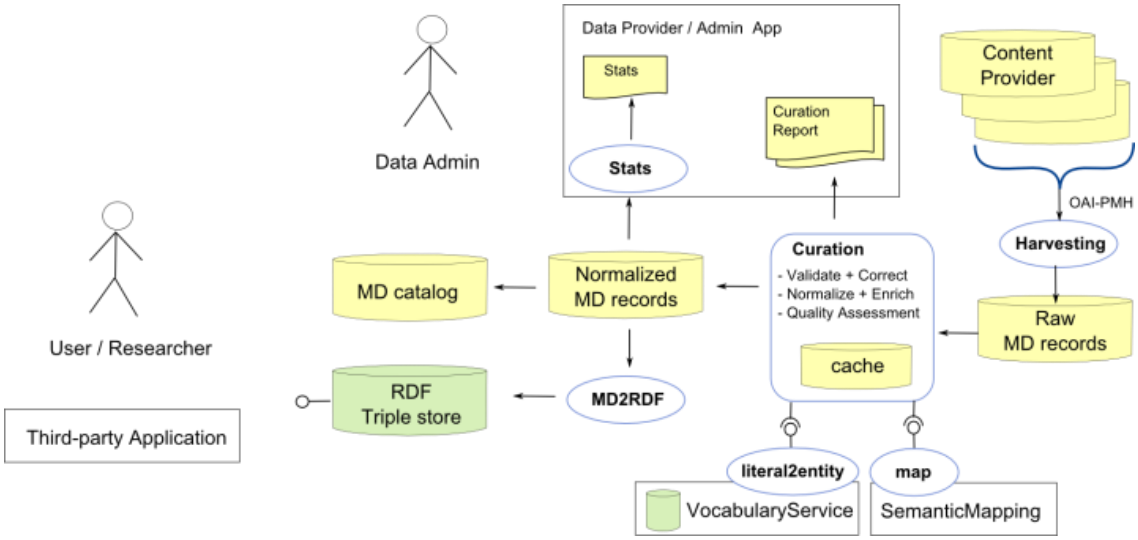

Figure 1: Sketch of a generic architecture with data flow between individual components:

application that enables users to search, browse and visualize the information. Figure 1 depicts the basic setup and flow of data: The raw data harvested from content providers undergo basic curation procedure (such as validation, analysis, normalization, enrichment). This is taken care of by a dedicated "curation" service that processes the incoming metadata, performing semantic mapping of the schemas and matching and normalizing the field values against vocabularies available in the Vocabulary Service. The normalized data is indexed and made available via a browser based catalogue to the users, but also as service via defined interfaces to other applications. Additionally, statistics and information about the curation process is generated and made available to the data administrators.

The catalogue will be implemented on top of a RDF triple store, but (especially during the development) this component will be assisted by an established search engine such as Apache Solr. While this dual setup means more effort, it allows for better control over the system's functionality (as one can compare and check the results returned by the two different indexing systems), ensures a good search performance and offers a fallback solution, in case of problems with the RDF-based system. However the main motivation for the dual approach is to serve as a bridging factor between the XML and RDF "world", bringing the traditionally XML-based metadata towards the Semantic Web.

Optionally, in a later stage also a component for editing the reference data could be integrated into the system. While this would be a very challenging task if started from scratch, such functionality is provided by the vocabulary repository OpenSKOS and could be reused at no extra cost. In any case, the editing functionality would pertain only to the supplementary (custom) vocabularies not to the reference data from external sources.

Equally, no (manual) editing of the metadata records is foreseen. The normalization and enrichment of the metadata records has to be realized as a completely automatic processing workflow that can be reapplied on every incoming batch of data. The main operation in such a workflow will be the matching of values in the records against the reference data. Thus the matching algorithm and the selection of vocabularies to match against are the main configuration parameters (degrees of freedom).

During the curation and normalization step custom vocabularies can be used to complement and – by applying precedence rules – even override information coming from external reference data like dbpedia or GND.

For the actual implementation, many of the components are already available either in the context of CLARIN or as open-source software: a number of OAI-PMH compliant harvesters, the open-source vocabulary repository OpenSKOS, the high-performant RDF triple store Virtuoso, together with SPARQL endpoint, plugin for faceted browsing etc. All this components will be integrated into the system running as own service instances on a dedicated server. The Semantic Mapping component (Ďurčo, 2013) providing crosswalks between metadata schemas is also available as a prototype and will be further developed and integrated as a module within the Knowledge Hub. The SMC Browser is envisaged as one of the dedicated visualization components for exploring the data. Having most of the components available as stable software the focus in the development is on an overarching system that glues together the individual parts and ensures the data flow between them. The one component that still needs a substantial development effort is the Curation module. There is a simple Java application for validating CMD data being developed by Oliver Schonefeld at IDS that could be used as a starting point. However there is no implementation yet available for the interaction with a vocabulary service, neither is there a defined format or workflow for the curation output (the Curation Report).

All in all we can identify following interfaces of the system. On the consuming side, there is the OAI-PMH protocol for harvesting the records. On the output side (to be consumed by other systems), the system will offer an FCS as well as a SPARQL endpoint. There will be a separate output channel for the outcome of the curation and for the statistical data, however the interface and the format need yet to be defined. And finally, there is the two interfaces for entity lookup and semantic mapping that are both used by the internal component (Curation), but will also be available to other applications.

6. Users and Use cases

As primary users of the system metadata creators and data administrators of individual content providers are foreseen that shall be able to inspect the metadata they provide and get feedback on potential problems and possible improvements.

But the "normal user", i.e. the researcher should equally profit from the integrative aspect (more data to explore), the normalization (better recall), the enrichment (alternative exploration paths) and the additional overview and statistical information.

Another possible usage scenario represents the use by other applications. The curation component could be called as a service by a metadata authoring tool, to validate and assess the quality of the metadata record being edited, give feedback and possibly suggest amendments. Also search engines like the VLO could profit from a semantic mapping component delivering correspondencies between metadata fields in various schemas as well as from a harmonized curation and normalization. Once the proposed system is established and running the search engines could ingest already the preprocessed, normalized and enriched data. Finally, making the data available as LOD opens doors to the world of the Semantic Web.

7. Current status and outlook

The proposed system is still in planning. However several components and datasets are already available and have been installed and employed for testing at the ACDH: the OpenSKOS vocabulary service, Apache Solr indexer, Virtuoso Triple Store are up and running. Also the SMC browser has been successfully deployed and is already in productive use. The next steps are the identification of content providers, the setup of the harvesting component, and implementation of the curation service (harmonized with the work of the Metadata Curation Task Force). In an iterative workflow the results of curation and analysis will serve to identify potential new vocabularies.

8. Conclusions

Though the proposed system draws resemblance with (or even reuses) already established systems and resources, we believe that the integrative efforts go well beyond the current state-of-the-art solutions – we discern four integrative aspects: 1) collecting metadata from a wide range of content providers, irrespective of the metadata format, 2) collecting metadata also on research results (bibliographic information), 3) focus on interlinking of vocabularies and metadata, 4) the system as integration of a number of technical components.

9. Acknowledgements

The project is run by the Institute of Corpus Linguistics and Text Technology as part of the Austrian Centre for Digital Humanities and is jointly funded by the Austrian Academy of Sciences and the Federal Ministry of Science, Research and Economy.

10. References

Brugman, H. & Lindeman, M. (2012). Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*, pp. 66.

Ďurčo, M. (2013). *SMC4LRT - Semantic Mapping Component for Language Resources and Technology* Technical University, Vienna.

Ďurčo, M. & Windhouwer, M. (2014) From CLARIN Component Metadata to Linked Open Data. In *LDL 2014, LREC Workshop.*

Purday, J. (2013). *Making connections, Annual Report & Accounts 2013* The Europeana Foundation

Uytvanck, D. V.; Zinn, C.; Broeder, D.; Wittenburg, P. & Gardellini. (2010) Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In M. Calzolari, N.; Choukri, K. & others (Eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* European Language Resources Association