# CLARIN-DARIAH.AT - Weaving the network

**Matej Ďurčo, Karlheinz Mörth**
Austrian Centre for Digital Humanities, Austrian Academy of Sciences
Vienna, Austria
E-mail: matej.durco@oeaw.ac.at, karlheinz.moerth@oeaw.ac.at

**Abstract**

The paper gives an overview of recent developments in Austria regarding CLARIN and DARIAH, the two Digital Humanities research infrastructure consortia. The presentation touches on the manifold related international engagements as well as a new wave of national activities and projects. Special attention is directed towards semantic technologies which are becoming the focal point of a diverse range of research areas in the digital humanities, raising the question how HLT can support other disciplines to cope with the "semantic turn".

**CLARIN-DARIAH.AT – spletanje omrežja**

Prispevek poda pregled nedavnega razvoja konzorcijev raziskovalnih infrastruktur za področje digitalne humanistike CLARIN in DARIAH v Avstriji. Predstavljene so številne mednarodne povezave, kot tudi novi val nacionalnih aktivnosti in projektov. Posebna pozornost je namenjena semantičnim tehnologijam, ki postajajo žarišče raznovrstnih področij raziskav v digitalni humanistiki, s čimer se sproža vprašanje, kako lahko jezikovne tehnologije podpirajo ostale discipline in se spopadejo s »semantičnim obratom«.

**Keywords:** research infrastructures, digital humanities, semantic technologies

## 1. Looking back

Austria has been involved in CLARIN[1] and DARIAH[2] already since 2009. It actively contributed to the build-up of technical infrastructures and engaging in the set-up of the organisational structures. At that time, the main contributors were the Centre for Translation Studies at the University of Vienna, the Institute for Corpus Linguistics and Text Technology of the Austrian Academy of Sciences (ICLTT) and the Centre for Information Modelling, University Graz (ZIM). Within CLARIN, the contributions were chiefly related (a) to the Component Metadata Infrastructure (Broeder et al., 2010) – the prototypical development of individual exploitation-side modules, especially the Semantic Mapping Component (Ďurčo, 2013) – as well as (b) to the FCS – Federated Content Search, an initiative aiming at developing a distributed system allowing to search not only in metadata, but also in the content of the resources exposed by individual data providers (Stehouwer et al. 2012).

Even before the beginning of the pan-European research infrastructures the ICLTT (and its predecessor the Austrian Academy Corpus – AAC) had a long tradition in Digital Humanities dating back to the late 1990s. The most prominent example may be the AAC-FACKEL, the digital scholarly edition of the magazine “Die Fackel”[3] authored by Karl Kraus in the years 1899 until 1936. The institute looks also back on a tradition of experimental dictionary-making. More recently, both monolingual and bilingual lexicography has again gained in importance. The focus in these efforts has been on developing tools, working on lexicographic data and eLexicography standards.

Equally, ZIM has a long tradition in conducting DH projects, mostly through the well-proven strategy of accompanying humanities projects, offering them the expertise regarding data modelling, preservation and online publication. The technical heart of these activities is a fedora-based repository system called GAMS[4], which has been developed there since 2003.

These are only a few examples of digital humanities related research in Austria. A survey conducted in 2009 listed more than 30 projects producing digital language resources by 15 different research groups. However most activities have been performed as solitary projects not embedded in any larger framework. Much work needs to be done to achieve a higher degree of integration.

## 2. New phase – Austrian Centre for Digital Humanities

In 2013, the broad range of CLARIN and DARIAH activities carried out in the last years were brought together in a new initiative, the Austrian Centre for Digital Humanities / Digital Humanities Austria (DHA), a project being funded by the Ministry of Science, Research and Economy for the duration 3 years. Digital Humanities Austria has been designed as a platform and a network of excellence for the propagation and dissemination of the digital paradigm and the use of DH methods and technologies. DHA represents the Austrian implementation of the EU’s ESFRI[5] roadmap.

Working with digital resources and tools remains a methodological and logistical challenge for many researchers in the humanities. DHA fosters the use of digital data, tools and know-how by easing access, enabling

[1] Common Language Resources and Technology Infrastructure http://clarin.eu/
[2] Digital Research Infrastructure for the Arts and Humanities http://dariah.eu/
[3] http://corpus1.aac.ac.at/fackel
[4] http://gams.uni-graz.at
[5] European Strategy Forum on Research Infrastructures

the production of standards-based data and deepening existing skills. All these activities are conducted in close cooperation with institutes of the Austrian Academy of Sciences, the Austrian universities as well as other institutions in the country that conduct or support relevant research such as libraries, museums, archives etc.

## 2.1 Central concerns

Digital Humanities Austria is based on a new understanding of scholarly research which may not be reduced to simply using digitised materials. The existence of digital data is a prerequisite for digital research, however it is only one of many aspects of DH. The following principles have been agreed upon by many representatives of the digital paradigm as seminal characteristics of the new inventory of methods: systematic use of digital infrastructures, transdisciplinarity, collaborative work, participatory technologies (virtual research environments, web-based research portals), Open Access / Open Source and open life cycle of research data and research results.
The central concerns of DHA can be summarised in three key phrases explained further below:

- Save the Data
- From Data to Knowledge
- The right Toolbox

*Save the Data* covers all the issues related to ensuring long-term preservation and availability of existing and newly created digital research data. In many meetings with representatives of research groups at the Academy and other institutions this has been identified as a central and urgent concern. More often than not, research material produced during projects ends up undocumented on external drives and is lost for future research.
This issue has many facets, quality and format of the data including information about the data (metadata) being one, but also the availability of stable, reliable institutional or national repositories and a pressure or guidance from the funding agencies, to name the most relevant aspects.
One precondition for long-term preservation is the question of standardised formats, when modelling research data. Using standards and de-facto standards makes it far more likely that the research data can be reused by others and is compatible with external third-party systems. Consequently, the overall strategy of DHA revolves around the triad: data, tools and standards, standards representing the glue between data and tools. The preferred/default format for text-based data in DHA – which is in line with widespread usage in the DH community – is the de-facto standard TEI/XML, however it is clear that no one format can cover the diversity encountered in the broad field of DH. To tackle the issue of standardisation, both CLARIN and DARIAH have established bodies responsible for surveying existing practices and working on recommendations and guidelines.
Another aspect of data preservation are dedicated institutional repositories as crucial infrastructure components that are able to handle not just scientific publications (usually documents in PDF format), but also complex structured research data. A good example of such a repository is the GAMS run by ZIM (Graz) that offers an integrated single-sourced, multi-view system relying on community based standards.
A new addition to the repositories landscape and a major achievement of 2014 is the new CLARIN Centre Vienna[6], the first Austrian node in the network of CLARIN Centres[7] which has acquired after a comprehensive assessment procedure the DSA[8] (Data Seal of Approval) and the CLARIN Centre B status as of April 2014. The core of the CCV is the Language Resources Portal, a depositing and publishing service primarily intended for digital language resources with a humanities background (Budin et al., 2013).
The issue of long-term preservation and availability of research data has gained increased importance as funding agencies have become aware of this issue and started to demand strategies for the availability of research results and research data.
In order to raise the general awareness and to intensify the discussion about this issue, a workshop on long-term preservation of data and repositories will be held during the Austrian Days of Digital Humanities in Vienna beginning December this year. In this workshop data producers and providers of scientific repositories shall come together to discuss problems related to data management and possible solutions of these problems.
*From Data to Knowledge* is the second focus of DHA. Scholarly work in DH often means to enrich data, to interpret, to annotate (semantically) and to interlink data. In the build-up of a modern network of knowledge semantic approaches and the paradigm of Linked (Open) Data are expected to play central roles. (We elaborate on this further in chapter 4.)
The goal of providing *the right toolbox* for DH poses quite a challenge considering the great number of involved disciplines with their quite varied traditions and methods, their often very particular research questions that often require highly specialised digital tools. In most fields, we have not yet out-of-the-box solutions. The development and propagation of innovative tools for the digital era, so-called dedicated applications, is a major focus of our efforts to support the continuously growing number of scholars pursuing digital research. Virtual research environments that enable researchers to work collaboratively are one such type of infrastructure components.
The ICLTT has been developing two suites of tools, one being a virtual research environment for lexicographic work, the other one is a platform for online publication of digital editions, called *corpus_shell*. This framewok is developed in collaboration with Telota[9] – the technical group at the Berlin-Brandenburg Academy of Sciences and Humanities.

[6] http://clarin.oeaw.ac.at/ccv
[7] http://www.clarin.eu/centres
[8] https://assessment.datasealofapproval.org/assessment_121/seal/html/
[9] http://www.bbaw.de/telota

As mentioned before, ZIM has been developing an integrated fedora-based repository system (GAMS) that comes together with a versatile open-source client for the management of the data in the repository which also includes batch editing of data.A representative of ZIM also contributes to DARIAH as task leader for "Reference software packages" in DARIAH's Virtual Competency Centre I (eInfrastructure), inventarizing existing software usable by DH research teams.

### 2.2 Organizational setup

The virtual network is organized in a national consortium comprising, a number of Austrian academic institutions. Next to the core members of the preparatory phase who still ensure the continuity of activities, the Technical University Vienna, the University of Innsbruck and additional new institutes of the Austrian Academy of Sciences, the University of Vienna and the University Graz joined the consortium. The network is funded by the Federal Ministry of Science, Research and Economy and coordinated by the Austrian Academy of Sciences.

### 2.3 Work packages

The activities of the new initiative are organised in three main thematic areas: Research Infrastructures for Digital Humanities (*RI4DH*) which chiefly perpetuates the involvement in CLARIN and DARIAH, a digitization initiative *go!digital*, and a bundle of activities to strengthen the DH in the education *dh-curriculum*.

*RI4DH* comprises the technical aspects of building research infrastructures and the various engagements in the European Research Infrastructure Consortia [10] CLARIN and DARIAH, as well as the coordination of the national efforts with the respective institutions on the European level.

The overall goal is to strengthen inter- and trans-disciplinary research and development in the humanities, on the basis of European research infrastructures DARIAH and CLARIN implementing the ESFRI roadmap. This also includes the construction of a research and service platform for the collaborative work of Austria DARIAH and CLARIN partners and their operational embedding in the two ERICs.

The main part of the RI related work lies in the procurement of so-called in-kind contributions, contracted between the national consortium and the ERICs on an annual basis. For CLARIN, the contributions consist mainly in digital language resources, but also other data (such as controlled vocabularies), software packages (e.g. lexicographic tools), services (like the establishment of the Language Resources Portal) or international events (workshops, conferences). The DHd conference[11] – Digital Humanities in German-speaking area – organized by the University of Graz in February 2015 is an example of a major event as an in-kind contribution.

One source for potential new Austrian in-kinds is the Language Resources Survey conducted back in 2009 together with its update planned for this year.

*go!digital* was set up as a call for innovative digitization projects in Austria. 5 projects were selected by an international jury out of 36 submissions. The projects have a duration of 1,5 - 2 years, and dispose of a budget of roughly 100.000 EUR each. The call put a strong emphasis on the use of standards and the integration with research infrastructures. The high number of proposals shows also the high potential in the Austrian research landscape.

Together with a related call "Digital cultural heritage" for proejects based at the AAS (with longer duration of projects and higher volume) 10 new DH projects will start by the end of the year, which constitutes an unprecedented surge of coordinated activities in this area in Austria (though the calls were explicitly inspired by similar setups in Netherlands and Germany). Under the motto *innovation*10* all the new projects will be presented in a kick-off event on 1 December 2014 as part of the Austrian Days of Digital Humanities, organised to foster collaboration and exchange among the projects and also in the broader community.

The third area of action is education. The DHA initiative *dh-curriculum* has been motivated by the evident lack of young DH experts in the country. The initiative's particular concern are consciousness raising activities and the training of young researchers. In addition to workshops, seminars and summer schools, participating researchers work on a DH curriculum, which aims to ensure the anchoring of related know-how in the academic education. Specialized courses are supposed to enable so-called data scientists to work in their respective disciplines, to support DH projects and to curate digital data collections (corpora, editions, digital archives, etc.).

While the University of Graz already offers a complete module on DH practices, there is nothing comparable to be found in the rest of the country. However, this deplorable state of affairs is going to be changed as a number of stakeholders have come together to establish a new cross-faculty department for Digital Humanities at the University of Vienna. This activity is in line with the initiatives on the European level where a working group in DARIAH is developing a DH course registry and a reference curriculum for DH teaching and training.

## 3. ACDH-ÖAW

Rooted in the long tradition of RI activities for DH at the Academy (in particular at the ICLTT), there are plans to setup a whole new institute dedicated to this task – the Austrian Centre for Digital Humanities (based at the Austrian Academy of Sciences). This institute will grow out of the ICLTT's technical group, but is planned to be substantially expanded to better cover the whole range of digital humanities, especially archaeology and historical studies. The ACDH-ÖAW will assume the role of a national coordinator and represent Austria in international RI bodies.

[10] or ERIC – a new European legal entity for research infrastructures

[11] http://dhd2015.uni-graz.at/

The strategy is directed towards a tight interaction between ACDH-ÖAW and the other institutes of the Academy, bundling and sharing development resources and technical solutions (don't create a new repository for every institute or even project) in a matrix organization, i.e. staff from individual other institutes is also actively involved in the activities. Institutes delegate colleagues to cooperate with the ACDH on common solutions working on particular problems in projects running at the institutes.

## 4. Semantic turn

With the extraordinary diversity of disciplines and communities of practice that constitute the "digital humanities" a major challenge is to find a common language, a common understanding of the problems the various disciplines share. A promising technical approach to tackle the issue is the advancement of semantic technologies and the Linked Open Data paradigm (LOD, Berners-Lee, 2006). Although RDF and related technologies in themselves are not a universal remedy for all interoperability problems (rather just another form of information representation), it at least offers a common widely adopted syntactic denominator. Combined with the unifying force of the RIs on the organizational level this approach seems to have a high integrative, harmonizing potential.

In this respect, it has to be acknowledged that the bulk of existing research data exists in databases or XML-based formats, which means that before being able to take advantage of the new technology, a major effort is required to transform or enrich the data. This does not necessarily imply that all of the data needs to be converted into RDF right away. The change, the "semantic turn" as it is called by many, can and should happen gradually, in small steps.

As a first step, it may just be enough to semantically annotate existing data using well defined semantic reference resources as vocabularies, the trivial example being annotating persons as named entities in texts using the GND[12] (or dbpedia) resolvers. For data structured in databases it may be rather worthwhile to try to remodel the data in RDF, but here too, it needs to be decided if the added value is worth the effort. Adding a field with a URI identifying or classifying a given entity based on selected reference resources may be enough in the initial phase.

So, before moving into the world of complex ontologies, the existing data has to be normalized and enriched with links to semantic entities defined in well-established reference resources such as taxonomies or authority files. Accordingly, a number of activities have been started within CLARIN and DARIAH, both on the European and the national levels, aiming to coordinate the creation and maintenance of controlled vocabularies and other reference data. Within CLARIN, the initiative CLAVAS (Brugman, Lindeman, 2012) 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 taken from the data category registry *ISOcat*[15] and a list of organization names that has been extracted from the metadata collected from collaborating content providers. However, 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[16]. All of these are available via the same system and constitute a large pool of valuable reference data that can be tapped into at no cost, taking advantage of a uniform interface.

It is also planned to use the vocabulary repository *OpenSKOS* as a core module of the Knowledge Hub, a new integrative system for data and knowledge management that is currently being developed at the ACDH-ÖAW and will be available via CCV as an infrastructure service. In an integrated, largely automated environment metadata will be aggregated from a number of sources, it will be normalized and enriched using controlled vocabularies integrated in the system, before it is made available for browsing and searching (Ďurčo & Mörth, 2014).

As part of its CLARIN-DARIAH-AT commitment, the Austrian Audiovisual Research Archive (Austrian Academy of Sciences) has started to work on taxonomies (musical instruments, languages and language variants, geographical reference data). So far these have been used internally only and will be made 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[17] (Borek et al., 2014) which was developed in the DARIAH community. It is based on experiences in previous work in other projects like NeDiMAH and on Bamboo's DiRT taxonomy. It is already being used in the DH course registry [18] and in the bibliography collection on DH (*Doing digital humanities - a DARIAH bibliography* [19] *)*. The above mentioned vocabularies are only a starting point, the system will 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. In adding new vocabularies, the focus lies on curation-intensive data such as for instance various named entities, e.g. organization names.

Another new service that has been launched recently as a DARIAH-DE contribution, offers a proxy service for the GND data (Gemeinsame Normdatei – the Integrated Authority File) which are maintained by the German National Library (GNL). The GND is a major normative reference resource in the German-speaking area and

[12] Gemeinsame Normdatei – the Integrated Authority File of the German National Library
[13] https://openskos.meertens.knaw.nl/
[14] http://openskos.org
[15] http://www.isocat.org
[16] http://openskos.beeldengeluid.nl/
[17] https://github.com/dhtaxonomy/TaDiRAH/
[18] http://dhcoursereg.hki.uni-koeln.de/
[19] https://www.zotero.org/groups/doing_digital_humanities_-_a_dariah_bibliography

beyond, however the native service endpoint provided by the GNL is available under very restrictive conditions. The unrestricted service which is made available by DARIAH-DE for the scientific community constitutes a major extension to the inventory of resources used for the task of semantic annotation.

The ACDH is involved in the activities as an early adopter and will be testing the new endpoint, planning to employ it in a number of ongoing DH projects requiring named entity recognition technology.

## 5. Conclusion and Outlook

As of 2014, a new phase in the institutional establishment of digital research infrastructures has begun. While on the European level CLARIN and DARIAH have both become official RI consortia formally installed by the European Commission, in Austria a new initiative was introduced to merge the hitherto rather fragmented activities, ensuring continuity by building on existing infrastructure components, but also breaking new ground through the orientation towards innovative cutting-edge technologies.

Next to the continuation of the "usual" RI work, ten new DH projects start this year which promises a substantial tide of new contributions in the years to come. The main challenge will be to safeguard the long-term preservation and availability of research data.

## 6. Acknowledgements

The initiative Austrian Centre for Digital Humanities / Digital Humanities Austria is funded and supported by the Federal Ministry of Science, Research and Economy and the partner institutions of CLARIN-DARIAH.AT[20].

## 7. References

AAC-Austrian Academy Corpus (2007). AAC-FACKEL Online Version: "Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936", http://www.aac.ac.at/fackel

Berners-Lee, T. (2006). Linked Data. online: http://www.w3.org/DesignIssues/LinkedData.html

Borek, L., Dombrowski, Q., Munson, M., Perkins, J. and Schöch, Ch. (2014). Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects. In *Proceedings of Digital Humanities 2014*. Lausanne, Switzerland.

Broeder, D., Kemps-Snijders, M. et al. (2010). A data category registry- and component-based metadata framework. In M. Calzolari, N.; Choukri, K. & others (Eds.). *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).* ELRA, Valetta.

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.

Budin, G., Moerth, K. and Ďurčo, M. (2013). Working towards European Infrastructures - The ICLTT Language Resources Portal. In *49. Jahrestagung des Instituts für Deutsche Sprache*, Poster-Session, Korpora geschriebener Sprache. IDS, Mannheim.

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. ELRA, Reykjavik.

Stehouwer, H., Ďurčo, M., Auer, E. and Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. pp. 3255-3259. ELRA, Istanbul.

Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P. and 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 2010).* ELRA, Valetta

[20] http://acdh.ac.at/consortium