European Lexicography Infrastructure Components

Gerhard Budin[1,2], Karlheinz Moerth[2], Matej Ďurčo[1]

[1]Centre for Translation Studies, University of Vienna,
Gymnasiumstrasse 50, A-1090 Vienna
[2]Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Vienna
E-mail: gerhard.budin@univie.ac.at, karlheinz.moerth@oeaw.ac.at,
matej.durco@univie.ac.at

Abstract

Industrial dictionary production has long since started to make use of modern ICT, and while in the world of Academia one can still find many projects working with slip boxes and simple text processors, academic dictionary writing has also begun to move towards digital methods. Although there is plenty of software available, the situation for smaller groups of researchers and individual linguists looks rather bleak. Tools are there, what is – however – needed by many researchers is readily available, standards-based, interoperable, and sustainable infrastructure. In our paper we will describe particular infrastructure components that can be used in building lexicographic infrastructure and describe the work of a group of researchers of several Austrian academic institutions who are currently putting together existing pieces of software to build an integrated modular toolbox for academic dictionary writing that would enable researchers to create, maintain and publish digital dictionaries. In the introduction of the paper, we will also try to give an outline of the institutional settings in which these activities are being carried out which is important in view of the fact that all of the described components are designed as Austrian contributions to the European infrastructures CLARIN-ERIC and DARIAH.

Keywords: research infrastructures; eLexicography; standards, tools

1. Introduction

Industrial dictionary production has long since started to make use of modern Information and Communications Technology (ICT), and while in the world of Academia and smaller lexicographic projects one can still find researchers working with slip boxes and simple text processors, dictionary writing in general has also begun to move, step by step, towards digital methods. Although large amounts of software were developed for use in big publishing houses, the situation for smaller groups of researchers and individual linguists looks rather bleak, as many solutions come at forbiddingly high prices. Infrastructure is there; what is needed by researchers is more common infrastructure: readily available, standards-based, interoperable, and sustainable infrastructure. This report concerns Austrian developments that may help to remedy this problem.

2. Digital research infrastructures

There exist many definitions of research infrastructures. A recent one has been formulated by the European Commission in their *Legal framework for a European Research Infrastructure Consortium (ERIC):*

> *"research infrastructure" means facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and covers major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information; enabling Information and Communications Technology-based infrastructures such as Grid computing, software and communication, or any other entity of a unique nature essential to achieve excellence in research.*

While many institutions in the humanities are concerned with building up basic technical facilities and services, others have already begun to think about next generation research infrastructures: infrastructures that are supposed to foster international cooperation as the key to the "excellence of research" by means of knowledge and technology exchange. Key words in these discussions are the 'Grid', the 'Cloud' and 'big data'.

2.1 ESFRI

In the European Union, the institutional foundation of activities in the field of digital research infrastructures started with ESFRI, the European Strategy Forum on Research Infrastructures, which was founded eleven years ago, in 2002. ESFRI is a group of national delegates and a representative of the Commission, who work together and pool resources to provide Europe with the most up-to-date research infrastructures. It is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.

ESFRI's task is not funding of projects, or realising infrastructures. It is rather an instrument to chart the landscape, to gather relevant information and to direct relevant developments. ESFRI has published a number of reports which describe the situation with regard to research infrastructures in the various scientific fields. In its Roadmap, an ongoing endeavour, it identifies potential new pan-European research infrastructures that are likely to be realised in the next 10 to 20 years. The number of candidate projects has been growing over the years. Roadmap 2006 listed 35 projects; the 2008 Update comprised 44. In 2010, the various scientific disciplines were organised into six major groups (Social Sciences and Humanities, Environmental Sciences, Energy, Biological and Medical Sciences, Materials and Analytical Facilities and Physical Sciences and Engineering) which comprise 48 projects (ESFRI 2010). The next update of the Roadmap is planned for 2015.

One example of a large-scale digital RI that countless researchers in the humanities use (usually without even being aware of its existence) is GÉANT, the pan-European research and education network. GÉANT is a high-speed network interconnecting Europe's National Research and Education Networks (NRENs). It was launched to facilitate cooperation and to enable scientists to share knowledge and resources. An indispensable service many researchers access when travelling across Europe and its universities is *eduroam*, the international roaming service for users in higher education.

An example of best practice and standards of infrastructure components is the Text Encoding Initiative (TEI), which also caters for text-oriented researchers. Most of what the TEI offers belongs in the category of community-based standards. However, the TEI is more than that, as it also provides tools (e.g. standardised schemas, ROMA, OxGarage etc.) and very effective and well used communication channels, such as the TEI mailing list.

The number of projects bearing the term *infrastructure* in their name, or explicitly aiming to build infrastructures, has risen steadily in recent years. Those of interest with respect to the SSH disciplines include EUDAT (European Data Infrastructure), CENDARI (Collaborative European Digital Archive Infrastructure) and EHRI (European Holocaust Research Infrastructure).

2.2 Digital Humanities

The fields and disciplines with which we are concerned are at the top of the ESFRI list (ESFRI 2010). The two initiatives mentioned there are CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities).

2.2.1 CLARIN

After a preparatory phase of several years, CLARIN was finally granted ERIC status by the European Commission in 2012.

CLARIN aims to provide easy and sustainable access to digital language data and advanced tools to discover, annotate, analyse or combine these data, irrespective of their physical location or format. The data involved are made up of a wide range of different types of language resources: representations of written and spoken language, some are text, others are offered as sound or video files. The target audience of CLARIN-ERIC are scholars in the humanities and social sciences. Currently, CLARIN-ERIC is in the process of establishing a networked federation of European data repositories, service centres and centres of expertise. They are planning to implement simple sign-on access for all members of the academic community in all participating countries. They are working on the interoperability of tools and data across the network, in order to allow researchers to combine distributed and heterogeneous data and to perform complex operations on these. This infrastructure is still under construction and will be so for quite some time. However, a number of participating centres have already started to offer services providing data, tools and expertise. Currently there are nine certified CLARIN centres[1]:

- ASV Leipzig, Bayerisches Archiv für Sprachsignale
- Berlin-Brandenburg Academy of Sciences and Humanities
- Eberhard Karls Universität Tübingen
- Hamburger Zentrum für Sprachkorpora
- IMS, Universität Stuttgart
- Institut für Deutsche Sprache
- MPI for Psycholinguistics
- Universität des Saarlandes

Others are preparing to obtain the official status of CLARIN centre:

- Centre of Estonian Language Resources (CELR)
- clarin.dk
- Austrian Centre for Digital Humanities
- DANS (Data Archiving and Networked Services)
- Huygens Instituut
- INL (Instituut voor Nederlandse Lexicologie)
- LINDAT-Clarin
- MI (Meertens Instituut)

2.2.2 DARIAH

The other focal large scale infrastructure initiative is DARIAH (Digital Research Infrastructure for the Arts and Humanities). As the name suggests it targets a very large community. Its declared goals are to enhance and support digitally-enabled research across the arts and humanities, to develop, maintain and operate an infrastructure in support of ICT-based research practices and to support researchers using ICT-enabled methods to analyse and interpret digital resources (DARIAH-EU Coordination Office 2013). The group of participating institutions and researchers is also aiming to set up an ERIC. DARIAH applied for ERIC legal status in autumn 2012.

In contrast to CLARIN, which organises its activities around physical service centres in the member countries, DARIAH has been operating through a network of four

[1] http://www.clarin.eu/node/2971

virtual competency centres (VCC):

- e-Infrastructure
- Scholarly Content Management
- Research and Education
- Advocacy

So far, each of the VCCs has been headed by two member countries and is formed of mixed groups of stakeholders. The VCCs have their own internal structure and specific workflows which are determined by the necessities of the particular tasks.

2.3 Infrastructure components

Infrastructure can be conceptualised in different ways, though this is beyond the remit of this paper. In a somewhat simplified manner, they can be seen as complex systems formed of a wide range of diverse technical (hardware, software, data) and organisational parts. Not all researchers require the same infrastructure components (ICs), and various disciplines have naturally varying requirements.

Language resources (LRs) are substantial in many fields today. Not only required by content producers and others active in cultural heritage, the work of an increasing body of researchers in SSH disciplines relies on availability of LRs. LRs can be described as a triad of tools, data and interoperability mechanisms. Tools comprise a combination of hardware and software, servers and services being put at the disposal of researchers. Data such as corpora, dictionaries, term-banks etc. constitute the contents, and interoperability mechanisms can be considered the glue that keeps tools and data together; they are the standards and norms that make LRs reusable. Neatly defined and well-documented interfaces are the basis for efficient service-based architectures that function in a distributed and heterogeneous digital biotope. In addition, we must not forget handbooks, documentation of all steps in the lifecycle of digital projects, and best practice guidelines in general to ensure reusability of newly-developed infrastructure components.

One particular type of language resource is dictionaries, which are an indispensable part of the scholarly tool inventory in many fields of the arts and humanities, in particular in all language-related disciplines. Libraries without dictionaries are unthinkable, and professionals, students, teachers, researchers and scholars equally use dictionaries, regardless of their field. Dictionaries have always been one of the most basic and integral elements of arts and humanities infrastructures.

2.4 The Austrian involvement

Research groups in Austria have been involved in both CLARIN and DARIAH for quite some time. In particular, two institutions played an important role in the establishment of CLARIN and DARIAH in Austria: the University of Vienna and the Austrian Academy of Sciences. The tight institutional connection of CLARIN-AT and DARIAH-AT allows synergism between the two groups.

2.4.1 CLARIN-AT

As mentioned before, the CLARIN technical infrastructure is being built around physical centres; institutions that have sufficient resources and expertise to make long-term commitment more likely. In some countries, several candidates for such centres exist and will undergo an evaluation process before becoming official CLARIN centres. Others have only just begun the process of establishing such centres. Austria is currently establishing a national CLARIN centre, the Austrian Centre for Digital Humanities (ACDH). ACDH will provide the community with several services. One of these will be an OAI-PMH endpoint that will give Austrian researchers the opportunity to feed their metadata into the CLARIN network. The Open Archives Initiative Protocol for Metadata Harvesting (Lagoze et al. 2002) is a standard, offering a comparatively simple mechanism to expose structured metadata in the Internet that has been adopted by the CLARIN community.

Given the wide community with different requirements with regard to metadata (e.g. OLAC, Dublin Core, TEI Headers etc.), CLARIN did not try to impose any one particular metadata scheme for describing the resources, but rather introduced a generic overarching architecture: CMDI (Component Metadata Infrastructure) (Broeder et al. 2012) which is able to accommodate various metadata schemes. Austrian researchers were also active in the development of CMDI.

The availability of research data has become an important issue in recent years. While more and more relevant data are being produced, many institutions conducting research programmes are not in a position to ensure long-term availability of data. Very often, databases move with researchers, corpora are left behind at departments and are no longer traceable once projects have ended. Although funding agencies are getting increasingly aware of the issue and are trying to impose stricter policies, many institutions neither have the required infrastructure nor the funds for long-term preservation of research data generated in these projects.

ACDH is planning to function as a host for such data, while also attempting to access already relinquished and forgotten data. It will offer researchers access to a dedicated repository for linguistically relevant research data.

2.4.2 DARIAH-AT

Austria is heading (together with Germany) the DARIAH Virtual Competency Centre 1. VCC1 is in charge of digital infrastructures; in a manner of speaking, taking care of the infrastructure of the infrastructure. In the context of the overarching project, this implies very particular core services such as authentication and authorisation, persistent identifiers and infrastructure components.

At the moment, DARIAH-AT's top priorities are digital infrastructures for the creation, maintenance and publication of digital language resources, in particular lexicographical data and large text collections. This is motivated by the general interests of the main partners currently involved in the construction work, which are departments concerned with linguistic, lexicographic and terminological research questions.

3. Lexicography infrastructure

The following paragraphs will provide detail about the infrastructure components that have come into existence as part of Austria's CLARIN and DARIAH engagements.

3.1 Dictionary-in-a-box

'Dictionary-in-a-box' is designed as an integrated modular toolbox offering lexicographers, working as individuals as well as in groups, all the necessary software to create, maintain and publish digital dictionaries. This suite is designed as a comprehensive virtual research environment geared towards the needs of researchers collecting lexicographic data. The target group is quite diversified, intended to include linguists from various fields, professionals in need of a simple lexicographic infrastructure, terminologists, etc. The suite will consist of the freely available dictionary editor Viennese Lexicographic Editor (VLE), styles, schemas, and server scripts that can be easily distributed and handled.

3.2 Dictionary editor

There exists a great deal of software for editing lexicographic data. Indeed, the list of well-established dictionary editing applications is quite long (for a short list see Budin and Moerth 2011). Some of these products provide a wide range of functionalities which can be applied to the whole lifecycle of the dictionary creating process: collecting, editing, refining and enhancing lexicographic data. Some packages are fully integrated systems; others are built in a modular way. Some are being used for particular purposes such as endangered languages, while some offer specialised multi-media support. Technically, dictionary writing software is often built around RDBM systems, very often making use of some client-server or multi-tier architecture.

The above mentioned VLE is a fairly new piece of software that came into existence as a by-product of an entirely different development activity. It was developed as part of an interactive online learning system for university students. It was first used in a collaborative glossary editing project carried out as part of university language courses at the University of Vienna. Over time, the tool proved to be sufficiently flexible and adaptable, and was put to work for other purposes in other projects. The interface is built around an XML editor that provides a number of functionalities typically required in editing linguistic data.

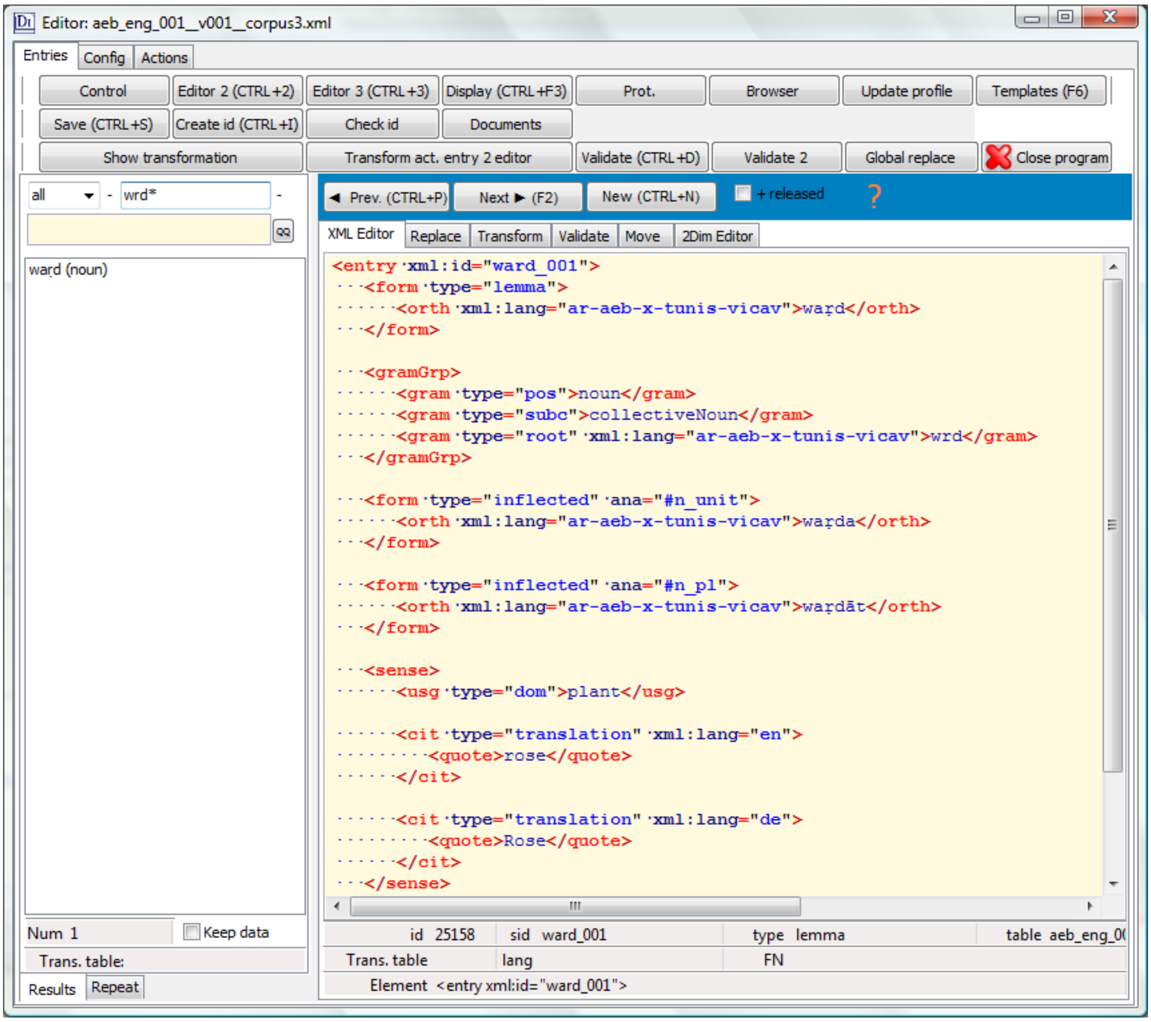
The motives to embark on this project were manifold. Some of the already existing systems were primarily intended for use in big publishing houses, pricing of licences accordingly high and the software consequently out of reach for small projects producing dictionary data. As the software evolved as a by-product of several smaller projects, production costs were manageable. In addition to the economic limitations, our projects were in need of full support for varying XML formats. The application was supposed to process standard-based lexicographic and terminological data such as LMF, TBX, and TEI. We were in need of simple scripting capabilities, a configurable interface allowing access to corpora and offering support for sophisticated validation mechanisms.

One of the particular features of VLE is a special module easing the integration of corpus examples into dictionaries. The main goal when programming this module was optimised access to digital corpora. It was intended to enable lexicographers to gather relevant sample sentences from external resources such as structured corpora (or the Internet) and to integrate these into dictionary entries in a reasonably comfortable manner. The focus in this work was direct access to the data. VLE's corpus interface enables lexicographers to launch corpus queries, and offers functionalities for selectively inserting data into existing dictionary entries without using the clipboard to copy-and-paste, which inevitably results in a lot of inefficient typing or clicking.

So far, VLE has been used to edit LMF, TEI, TBX and RDF data. The program provides a number of useful functions to automate editing procedures. Some of these cater to the needs motivated by the underlying XML structure. The editor is capable of highlighting XML elements and performing automatic text completion. The program can continually check the structural integrity (well-formedness) of input on the fly. Technologically, it draws not only on the XML core specification, but also on several cognate technologies. XSLT and XPath play an important role both for visualising and modifying existing datasets. Lexicographers can insert elements on the basis of predefined XML Schemas. Most of the functions can be applied both to single and multiple records.

Validation is a key issue in all XML based document editing. It is the process of checking the data on a level beyond the basic structural XML requirements (well-formedness). When validating the structure of a document, it is checked against a set of definitions of permissible elements and information as to where these elements may appear in the document. Currently, VLE expects document type definitions in the form of an XML Schema which is, like XML, a W3C recommendation. On the to-do-list of the programmers, there is also the implementation of an option to validate against RELAX NG, an ISO standard which has found much support in the TEI and OpenDocument communities.

In VLE, editing of dictionary data can be performed in two ways: the editor works either in XML mode (Figure 1), which may be considered as the expert mode, or in an editor form with predefined entry controls. The second option enables working in an interface made up of controls that are arranged like traditional database input fields. While working on an entry, it is possible to switch between the two modes. The second option, i.e. making use of edit controls for particular XML elements, is useful especially when working in the same field across a number of dictionary entries. Navigating is admittedly more cumbersome in the expert mode than in the edit controls. However, more complex structures, in particular elements nested inside one another, often make it necessary to switch into XML mode.



```
<entry xml:id="ward_001">
   <form type="lemma">
      <orth xml:lang="ar-aeb-x-tunis-vicav">warḍ</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="subc">collectiveNoun</gram>
      <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">wrd</gram>
   </gramGrp>

   <form type="inflected" ana="#n_unit">
      <orth xml:lang="ar-aeb-x-tunis-vicav">warḍa</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">warḍāt</orth>
   </form>

   <sense>
      <usg type="dom">plant</usg>

      <cit type="translation" xml:lang="en">
         <quote>rose</quote>
      </cit>

      <cit type="translation" xml:lang="de">
         <quote>Rose</quote>
      </cit>
   </sense>
```

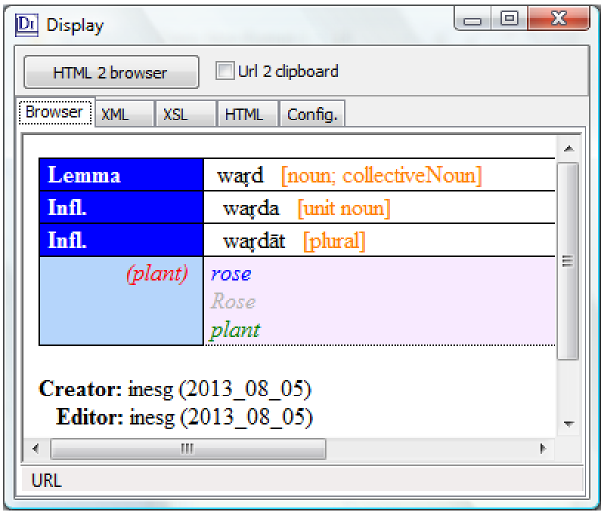
id 25158 | sid ward_001 | type lemma | table aeb_eng_0(
Trans. table | lang | FN
Element <entry xml:id="ward_001">

Figure 1: XML view

The tool visualises data by means of freely configurable XSLT stylesheets (Figure 2). While this functionality is quite commonplace in many XML based applications today, VLE proves to be particularly versatile. It is possible to apply different styles by switching between different views of the same set of data. Automatically generated links in the output data (usually HTML) enable navigation from these visualisations back into the editor control.

The program has a number of features that are intended to ease the lexicographer’s workload. One of these features is a configurable keyboard layout which is designed to support the comfortable input of Unicode characters usually not available in standard key assignments. The software can be configured to automatically choose the appropriate keyboard assignment when moving from one element to another. This functionality is based on the @xml:lang attribute and spares the user from manually switching between keyboard layouts. For example, when working on contents of an element having an @xml:lang="ru" attribute, VLE automatically activates the Russian keyboard layout; on entering an element with the attribute @xml:lang="de", it switches back to German. The program is able to automatically create unique and meaningful identifiers for entries and example sentences on the basis of the contents of the respective items.

The current VLE version is a stand-alone application that requires Windows operating system. One of the project’s midterm goals is the development of a fully-fledged browser-based interface. While the list of requirements for such an interface is clearly defined, the implementation would require time and resources which are currently being sought.



Lemma	waṛd [noun; collectiveNoun]
Infl.	waṛda [unit noun]
Infl.	waṛdāt [plural]
(plant)	rose Rose plant

Creator: inesg (2013_08_05)
Editor: inesg (2013_08_05)

Figure 3: HTML view

3.3 Dictionary server

Usually, VLE does not work on locally stored data. Data are stored on a remote server that can easily be set up and configured. The system is organised as a client-server architecture. The communication between the dictionary client, i.e. VLE, and the server has been implemented as a REST (Representational State Transfer) web service, which facilitates access to the server and consequently the data with tools other than our own.

The server builds entirely on open and freely available software that can be readily distributed. In the first implementation level, it makes use of the MySQL database, which is connected to clients through a REST-style web service. Querying works on the basis of SRU/CQL (Search/Retrieval via URL + Contextual Query Language). This search protocol was developed by the Library of Congress as successor of the Z39.50 protocol and is being tested and worked on by CLARIN's Federated Content Search (FCS) working group.

The distributed architecture has a number of obvious advantages. Being able to work on the data wherever one has access to the internet is unquestionably a useful feature. Lexicographers can work from anywhere, without having to carry their data around. But, most importantly, this setup also allows for collaborative work on the dictionary data.

VLE allows several editors to work simultaneously on the same dictionary, making use of a simple locking mechanism. When one lexicographer opens an entry, the entry can still be read by other editors, but cannot be edited. An additional feature of the server module currently being developed is an efficient versioning mechanism. We anticipate that this functionality, which might be of particular interest in collaborative settings, will be available by early 2014.

3.4 DictGate

DARIAH-AT is planning to set up a server that will allow (groups of) researchers to host lexicographic data. This infrastructure is intended both for producing and publishing lexicographic data. Thus, the dictionary gate is designed as a two-lane carriageway that will allow both data entry and retrieval. Users will be able to use the central server to produce data and to set up web-based interfaces that make use of the DictGate's web services.

Primary target groups are not commercial entities but researchers working on smaller lexicographic projects that are in need of solutions that can be applied without much logistical and technical overhead. Institutionally, the service will be based at the Austrian Academy of Sciences which has a long-standing and quite diversified tradition in dictionary production.

3.5 Lexicographic data

With respect to data, the DictGate working group pursues several lines. A first stock will be provided by lexicographic data that are being created in the context of the VICAV (Vienna Corpus of Arabic Varieties) project. The contributors of this project are currently setting up a platform to host and exchange a wide range of digital language resources for Arabic studies. Among these data (language profiles, bibliographies, corpora, ...) there are also smaller digital dictionaries of Arabic varieties. Besides of Damascus Arabic, dictionaries for the varieties of Morocco (Rabat) and Egypt (Cairo) are being compiled. A dictionary of Tunis Arabic will be elaborated as part of the project *Lexical dynamics in the Greater Tunis area: a corpus based approach*, which was approved by the Austrian Science Fund in March 2013 and will run for three years. These four dictionaries are being compiled with a special focus on comparative research questions and are structured in a manner that will enable performing queries on the four dictionaries to retrieve integrated datasets. These language resources (tools and data) are intended both for research purposes and academic language instruction.

There are several other research groups that plan to publish their data through the DictGate platform. A first product will be a Persian–English Dictionary of Single Word Verbs and a Russian–German dictionary which is currently being developed. One of the long-term dictionary projects of the Austrian Academy of Sciences, the *Dictionary of Bavarian dialects in Austria*, is also involved and will contribute data to the platform. While the current focus is on linguistics, the project generally targets a wider humanities audience. Among the resources to be made available there are also historical dictionaries that are of interest for disciplines other than linguistics.

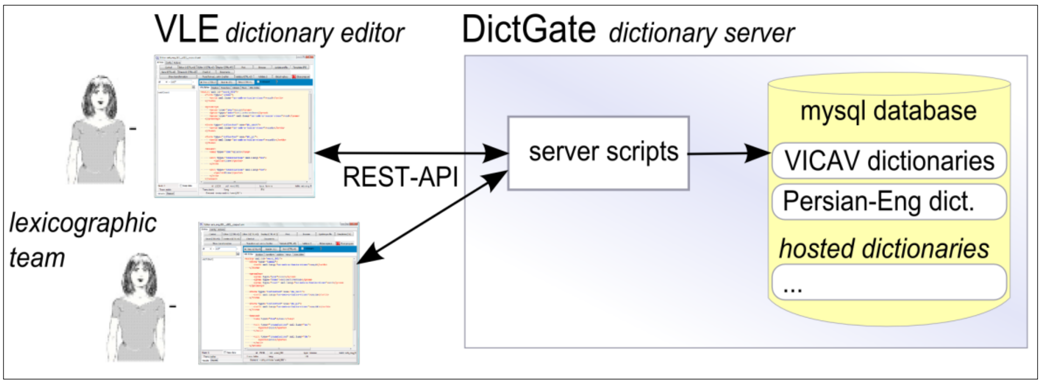


Figure 4: Client-server setup of the lexicographic infrastructure

We will also strive to access data without institutional backing to host or maintain them, to work towards efficient service-based lexicographical infrastructures that also offer data that can be used for NLP applications.

3.6 Access Policy

Basically, the focus of all these activities is on open-access resources. So far, no binding decision has been made as to the licence under which DictGate offerings will be available. However, there is a strong case for a Creative Commons licence, CC-BY being the favoured option. Discussions with interested researchers and other stakeholders have shown that permission to create derivative works is usually regarded an important prerequisite in order to ensure reuse of data.

Free access will not be a preclusive condition for the incorporation of data in the DictGate platform. However, open access will be strongly encouraged as funding organisations increasingly demand open access to publicly funded research data. In this respect, it will be important to get to a point where truly open access to data implies more than the availability of pdf documents, but direct access to data in reusable (i.e. standardised) formats.

4. Standards

Both CLARIN-ERIC and DARIAH consider standards a major concern of their activities and have institutionalised their respective work. CLARIN-ERIC has set up a Standards Committee to advise the Board of Directors on the adoption of standards to be supported by infrastructure. In the DARIAH network, various working bodies share the declared intention of working on standards and the formulation of best practises. As a particular form of language resource, standards, technical specifications and best practises are thus to be regarded as important cornerstones of digital infrastructures and should be considered infrastructure components in their own right.

When creating digital lexicographic resources, several standards and de-facto standards have to be considered. There are, for example, LMF (Lexical Markup Framework, ISO 24613:2008) and the dictionary module of the Guidelines of the Text Encoding Initiative. The bundle of documents created by ISO-TC37 (Terminology and other language and content resources) also contains a number of relevant specifications, such as ISO 639 (Codes for the representation of names of languages) or ISO 24610-1:2006 (Language resource management – Feature structures – Part 1: Feature structure representation), that should be considered.

As to the format used by the software components of the proposed infrastructure services, the goal was to come up with solutions that would be as open and flexible as possible. The core data of the initial phase of the project will be encoded in TEI P5[2]. This is in particular due to the fact that the contributing partners provide data in this format. The involved projects are mostly rooted in humanities disciplines that have a long tradition in making use of the TEI guidelines.

The guidelines of the TEI comprise an ample set of well-tried, and in many parts thoroughly discussed, specifications for a wide range of encoding scenarios. It has grown as the de-facto encoding standard for dictionaries digitized from print sources. Interestingly, the most recent versions of the TEI Guidelines contain a passage that indicates that the authors are actually aiming at a much wider range of dictionaries:

[2] http//www.tei-c.org/Guidelines/P5/

> *... The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for word lists, lexica, glossaries, etc. included within other documents. (TEI Consortium P5 2012, 247)*

The idea of extending the scope of the TEI dictionary module for use with language-processing software is not as far-fetched as it may seem at first glance. The interest in the issue has been clearly documented by the large audience of the workshop “Tightening the Representation of Lexical Data: A TEI Perspective”, which was held at the 2011 Annual Conference and Members’ Meeting in Würzburg (Germany).

The dictionaries to be published in the first round share a common schema which was developed on the basis of the TEI dictionary module. This schema is made up of a comparatively small subset of elements and imposes a number of clearly defined constraints to make the resulting dictionaries interoperable with one another and some other language resources.

In using the Guidelines of the TEI for linguistic and lexicographic purposes, encoders usually combine them with other standards in a complementary manner. Thus, it has become common practice in TEI encoding to make use of the global attribute @xml:lang which has been incorporated into the Guidelines from the World Wide Web Consortium’s XML Specification. TEI prescribes this attribute to identify both linguistic varieties and writing systems. In this hybrid approach, the value of the attribute should be constructed in accordance with the Internet Engineering Task Force’s *Best Current Practice 47* (BCP 47) which in turn refers to and aggregates a number of ISO standards (639-1, 639-2, ISO 15924, ISO 3166).

An equally important tool applied in the encoding of these dictionaries is ISOcat, the ISO TC 37 (Terminology and Other Language and Content Resources; Kemps-Snijders et al. 2009) Data Category Registry[3] that has been set up as a publicly available pool for definitions of widely accepted linguistic concepts. ISOcat can, for instance, be applied in TEI when annotating word forms with word class information. The ISOcat database assigns each data category a unique persistent identifier which makes them universally identifiable.

Additional infrastructure components to be contributed by the Austrian partners of CLARIN-ERIC and DARIAH belong to the third type of the above introduced data-tools-interoperability triad. It is not only important to adhere to standards. To enable others to work along similar lines, thorough documentation and examples are needed that in turn can serve as the basis of new projects and further developments in ongoing standardisation processes.

[3] http://www.isocat.org

5. Status

At the time of preparing this report, most of the components described here are functioning and in use by researchers for their everyday work. Distributable prototypes of Dictionary-in-a-box are currently being tested and a first version will be available by early next year. The dictionary editor is already available and can be freely downloaded through the Language Resources Portal of the Institute of Corpus Linguistics and Text Technology[4].

6. Conclusions

In this report, we introduced a suite of easy-to-adopt tools for collaborative lexicographic work and their embedment in evolving SSH research infrastructures. All of this work is driven by a vision of a growing ecosystem of freely accessible and distributable lexical resources being used by growing communities of researchers. We hope that our open concept and the readily available infrastructure will create new and sustainable dynamics in the field of lexicographic data production.

7. Acknowledgements

The work described in this paper has been made possible by funds provided by the Austrian Federal Ministry for Science and Research (Bundesministerium für Wissenschaft und Forschung), the University of Vienna and the Austrian Academy of Sciences.

8. References

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T. & Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*, p. 1.

Budin, G., Mörth, K. (2011). Hooking up to the Corpus: the Viennese Lexicographic Editor's Corpus Interface. In *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10–12 November 2011*, edited by Iztok Kosem and Karmen Kosem. Ljubljana: Trojina, pp. 52–59. Institute for Applied Slovene Studies.

Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries. In *Journal of the Text Encoding Initiative (TEI and Linguistics)* 3.

DARIAH-EU Coordination Office (2013). *Introducing DARIAH-EU*. Accessed at:

[4] http://oeaw.ac.at/icltt/vle

www.dariah.eu/ indexcc3b.pdf.

European Commission (2010). *Legal framework for a European Research Infrastructure Consortium – ERIC. Practical Guidelines.* Accessed at: ec.europa.eu/ research/infrastructures/pdf/eric_en.pdf.

European Science Foundation (2011). Research Infrastructures in the Digital Humanities. In *Science Policy Briefing* 42.

European Strategy Forum on Research Infrastructures (2010). *Strategy Report on Research Infrastructures. Roadmap 2010.* Luxembourg Publications Office of the European Union. (doi:10.2777/23127)

Ide, N., Kilgarriff, A. & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In *Proceedings of the Ninth EURALEX International Congress: EURALEX 2000*: Stuttgart, Germany, August 8th–12th, 2000, 113–126. Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

ISO-24612:2012 (2012). *Language resource management – Morpho-syntactic annotation framework.*

ISO-24613:2008 (2008). *Language resource management – Lexical markup framework (LMF).*

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. & Wright, S.E. (2009). ISOcat: Remodelling Metadata for Language Resources. In *International Journal on Metadata, Semantics and Ontologies* 4: pp. 261–276.

Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. June 2002. Accessed at: http://www.openarchives. org/OAI/2.0/openarchivesprotocol.htm.

Romary, L., Salmon-Alt, S. & Francopoulo, G. (2004). Standards Going Concrete: From LMF to Morphalou. In *Workshop on Enhancing and Using Electronic Dictionaries*. Geneva: Coling.

Romary, L. (2010). Standardization of the Formal Representation of Lexical Information for NLP. In *Dictionaries: An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography. Accessed at: http://arxiv.org/ abs/0911.5116.

Romary, L. (2010). Using the TEI Framework as a Possible Serialization for LMF. Paper presented at RELISH workshop, August 4–5, 2010, Nijmegen, Netherlands. Accessed at: http://hal.archives-ouvertes. fr/docs/00/51/17/69/PDF/NijmegenLexicaAugust2010.pdf.

TEI Consortium (2013). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 2.3.0. Last updated on 17th January 2013. Accessed at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Wandl-Vogt, E. (2010). Multiple access routes. The Dictionary of Bavarian Dialects in

Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In *eLexicography in the 21st century. New challenges, new applications, Proceedings of eLex2009.* S. Granger & M. Paquot (eds), Louvain-la-Neuve, Presses Universitaires de Louvain: pp.451-455.