Furthering the Exploration of Language Diversity and Pan-European Culture: The DARIAH-CC Science Gateway for Lexicographers

Eveline Wandl-Vogt[1], Roberto Barbera[2], Giuseppe La Rocca[2], Antonio Calanducci[2], Carla Carrubba[2], Giuseppina Inserra[2], Tibor Kalman[3], Gergely Sipos[4], Zoltan Farkas[5], Davor Davidović[6]

[1] Austrian Academy of Sciences, Austrian Centre for Digital Humanities; [2] Italian National Institute of Nuclear Physics (INFN); [3] Gesellschaft für wissenschaftliche Datenverarbeitung mbh Göttingen; [4] EGI Foundation; [5] MTA SZTAKI; [6] Ruđer Bošković Institute

e-mail: eveline.wandl-vogt@oeaw.ac.at; roberto.barbara@ct.infn.it; giuseppe.larocca@egi.eu; antonio.calanducci@ct.infn.it; carla.carrubba@ct.infn.it; giuseppina.inserra@ct.infn.it; tibor.kalman@gwdg.de; gergely.sipos@egi.eu; zoltan.farkas@sztaki.mta.hu; davor.davidovic@irb.hr

Abstract

The paper introduces into a new Science Gateway, developed in the framework of the European Horizon 2020 project EGI Engage - DARIAH Competence Centre, which started in March 2015 co-funded by the European Union, with the participation of about 70 (research) units in over 30 countries.

In this paper the authors focus on transdisciplinary collaboration in the framework of explorative lexicography in cultural context. On the one hand, they give a short overview of the architecture of the Science Gateway, used techniques, and specific applications and services developed during the DARIAH Competence Centre. On the other they mainly focus on possible added value and changes concerning work flow for Lexicographers and researchers on Lexical resources.

This is exemplified on the European network of COST action IS 1305 "European Network of electronic lexicography (ENeL)".

Keywords: science gateway; research infrastructure; cloud infrastructure; social infastructure; open science commons; cultural diversity; cultural lexicography; linguistic diversity; Pan-European heritage; explorative scholarship

1. Introduction: Lexicography in a World of Cultural Diversity

The UNESCO Declaration on Cultural Diversity (2001) states the importance of Cultural Diversity as a "common heritage of humanity" and makes its defence "an ethical imperative indissociable from respect for the dignity of the individual."

Languages can be interpreted as one of the cornerstones part of the larger culture of the community that speaks them. There are about 6000 – 7000 languages spoken in the world. About 96% of the world's languages are spoken by just 4% of the world's inhabitants. Just 3% of the world's total, namely 225 languages, are indigenous to Europe. Same picture shows the Linguistic diversity index measured according to Greenberg [1956]. This surely influences European lexicography.

European lexicography can be seen as a common space of knowledge with shared practices in lexicography (COST MoU 2013). Yet, the big dictionaries' endeavours of the 19th and 20th century were meeting the purpose of nation-building, actually against the background of a (pluri-) linguistic reality (COST MoU 2013). Referring to the index and the UNESCO declaration, plurilingualism is much more the normal human condition than monolingualism, and should be

supported.
Embedded into a world of emerging technologies as well as social movements, Lexicography is at a crossroads. On the one hand, to meet cultural and linguistic diversity reality and support this, on the other, to make rich use of emerging technologies to meet user needs. Hanks (2012: 82) argues, that it is "too early to say what form innovative dictionaries of the future will take".
Although the picture of the future "dictionaries" is quite vague, it is obvious, that technologies and societal needs are challenging lexicography. Lexicographers are facing, on the one hand, rich opportunities such as exploited by Wikimedia and/or Google in different ways from traditional lexicography; on the other hand, funding for traditional lexicographical products is decreasing. There is a trend to still try to develop tools and services in house. Most of small lexicography teams without access to knowledge, novel technologies, financial infrastructure and money cannot succeed in the long run on this (scientific) market.
In this paper we explore an opposite model, making vast use of rich, interdisciplinary, virtual collaboration and shared expertise: the authors are introducing into the DARIAH Competence Center Science gateway, and in doing so embed lexicography into a new cultural, open science paradigm on the example on lexicography of the minority language of Bavarian dialects in Austria.

2. Towards Open Science Commons

Over the last decade EGI has built a distributed computing and data infrastructure to support over 21,000 researchers from many disciplines with unprecedented data analysis capabilities. EGI builds on the European and national investments and relies on the expertise of the 'EGI Foundation'. EGI partners offer a federated computing and storage infrastructure for scientists, currently including 650,000 logical CPUs, 500 PB of disk and tape storage and 21 clouds. The infrastructure can be accessed for high throughput computing and cloud computing/storage use cases.
EGI-Engage[1] is the current flagship project of EGI, co-funded in the European Commission Horizon 2020 programme between 2015-2017. The mission of EGI-Engage is to accelerate the implementation of the Open Science Commons vision[2], where researchers from all disciplines have easy and open access to the innovative digital services, data, knowledge and expertise they need for their work. The Open Science Commons is a new approach to sharing and governing these assets. EGI-Engage contributes to the establishment of three pillars of Open Science Commons: 1) e-Infrastructure Commons; 2) Open Data Commons; 3) Knowledge Commons.
EGI-Engage expands the current capabilities that EGI offers to scientists and also the spectrum of the user base by engaging with large Research Infrastructures (RIs). The main engagement instrument is a network of Competence Centres, each centre operating as an incubator and provider of e-Infrastructure services for a specific RI community. Such collaboration establishes the 'Knowledge Commons' of the Open Science Commons.
The DARIAH Competence Centre (DARIAH-CC[3]) is one of 8 Competence Centres established during the EU H2020 EGI-Engage project.
The goal of the DARIAH-CC is not only to provide generic solutions and services for all researchers and scholars coming from Arts and Humanities but rather provide concrete solutions and service for particular research domains or groups, one of which is lexicography. Thus, the DARIAH-CC provides a set of specific cloud-based services and solutions tailored for the needs of lexicographers.

[1] https://www.egi.eu/about/egi-engage/
[2] https://www.opensciencecommons.org/
[3] https://wiki.egi.eu/wiki/Competence_centre_DARIAH

3. DARIAH-CC Services and Demonstrator Applications

The DARIAH-CC has developed several applications and services based on the platforms and frameworks that were developed by its members. The basic, underlying platforms, on top of which specific applications and services are developed are:
1) gLibrary - a digital repository system; 2) CDSTAR - a storage system for storing and searching of structured and unstructured data; 3) WS-PGRADE - a Liferay-based, workflow-oriented web portal.
On top of these platforms, during the DARIAH-CC the following applications and services were developed:
1) Parallel semantic search engine (PSSE); 2) Bavarian dialects repository; 3) Multi-source distributed real-time search and information retrieval application (SIR) for processing OCR documents; 4) DARIAH Science gateway.
Furthermore, DARIAH-CC provides access to cloud-based compute and storage resources compatible with the EGI FedCloud.
The following subsections detail the e-Infrastructure services and demonstrator applications that are developed and offered by the DARIAH-CC to lexicographers but which are also applicable and usable by the broader Digital Arts and Humanities community.

3.1. Cloud-based computing and storage infrastructure

The cloud infrastructure for Digital Arts and Humanities established by DARIAH-CC is the most basic service, based on the EGI Federated Cloud technology[4]. This cloud infrastructure currently connects cloud resources from two sites (INFN-Catania and -Bari). Three more sites are expected to join later in 2016 (GWDG, IRB, SZTAKI). The clouds are joint into a virtually single system through an identity federation, which enables researchers and their applications to access all these clouds using a single identity with single-sign on, and enables the CC to decide about user access requests in a centralized way.
The infrastructure is offered as a scalable system for data and computing-intensive applications. These applications can be deployed on the DARIAH-CC cloud in the form of Virtual Machine images, through the EGI Cloud Marketplace portal[5]. The distinct clouds of the federation can all be accessed through agreed, standard interfaces[6] by the users and their applications, ensuring portability across clouds, and sustainability through different cloud generations.

3.2. Parallel Semantic Search Engine (PSSE) service

The PSSE is a service conceived to demonstrate the potential of Open Access Data Infrastructures coupled with semantic web technologies to address issues of data discovery and correlation. Thanks to this service developed by INFN, users can search in parallel keywords in more than 100 languages across more than 30 millions resources contained in the thousands of semantically enriched Open Access Document Repositories (OADRs) and Data Repositories (DRs). Search results are ranked according to the Ranking Web of Repositories[7] and listed in a table view. For each record found, some basic information such as: the title, the author(s) and a short description of the corresponding resource are provided; another click offers additional information: the link to the open access document and to the corresponding dataset, if any.

[4] http://go.egi.eu/cloud
[5] https://appdb.egi.eu/browse/cloud
[6] The main interfaces are: Open Cloud Computing Interface (OCCI), and x509 authentication.
[7] http://repositories.webometrics.info/

The PSSE service is available through the Science Gateway tailored to address the requirements of the DARIAH community. A standard based JSR-286 portlet to specify the keywords to search has been deployed on top of an application server to improve the user's experience.
The service is currently configured to search digital contents in the e-Infrastructure Knowledge Base, OpenAgris, Europeana, CulturaItalia, Isidore, Pubmed and DBpedia. Other lexicographical relevant resources are about to be connected in the future.

3.3. gLibrary service and Bavarian dialects application

The second service provided by DARIAH-CC is gLibrary[8], a digital repository system that offers both, access to already existing repositories and provides additional functionalities to build customized and high-level digital repositories using REST technology. gLibrary 2.0 was rewritten from scratch using the Node.JS API Framework LoopBack[9] and new functionalities were developed to help lexicographers. It is now completely agnostic about the storage infrastructures (e.g. XML, TEI, MySQL). Thanks to new functionalities, lexicographers have now the possibility to use the world-wide cloud storage facilities provided by the EGI Federated Cloud Infrastructure as storage back-ends to store digital assets.
The concrete application developed using the gLibrary framework is the Bavarian dialects in Austria dataset repository (Database of Bavarian dialects in Austria [DBÖ / dbo@ema]). The dataset represents a work of 100+ years old collection of Bavarian dialects within the Austrian-Hungarian monarchy from the beginning of German language to nowadays. This use-case demonstrates the better organization of the large datasets which are now stored on distributed, cloud-based storage resources, rather than local storage systems. The benefit of using gLibrary-based repository is that it provides a much larger storage capacities within Cloud infrastructure and, at the same time, increases the resilience of the data on the system failures.

3.4. CDSTAR and optical text recognition

The Common Data Storage ARchitecture (CDSTAR; c.f. Kalman, Tonne, Schmitt 2015) is a system to store and search diverse structured and unstructured research data. The CDSTAR architecture is designed such a way that it provides an easy-to-use system than can store, modify, search, and access a large amounts of diverse research data. CDSTAR has a built-in custom object storage solution that addresses the specific requirements of the research data management according to the good scientific practices. The metadata is kept along with the research data using a flexibe metadata schema. Furthermore, the data stored in CDSTAR is automatically assigned with an ePIC Persistent Identifier (PID)[10]; this way data can be cited in scholarly publications.
The multi-source distributed real-time Search and Information Retrieval (SIR) is a framework based on the CDSTAR technology.
The OCR use case targeted in DARIAH-CC is a very computer-intensive application. The developments enable quick and dynamic provisioning of tools and services. The utilization of the computing power and cloud systems allow the accelerated execution of optical character recognition, processing, and quality assessment. Through this, new levels of quality are possible.

3.5. WS-PGRADE Science Gateway

The DARIAH WS-PGRADE (Kacsuk et al. 2012) gateway will be a high level application

[8] https://glibrary.ct.infn.it
[9] https://loopback.io/
[10] ePIC - European Persistent Identifier Consortium. Online: http://pidconsortium.eu - accessed: 16.05.2016.

development and execution environment on top of the DARIAH-CC cloud infrastructure. At the time of writing this paper, the generic WS-PGRADE technology is under customization by the DARIAH-CC in three directions:
1. Making WS-PGRADE connected to the federated cloud infrastructure through the WS-PGRADE graphical and programming interfaces.
2. Connecting WS-PGRADE to the DARIAH Identity Provider service (IdP) organizing easy access to distributed services and providers.
3. Integrating as many as possible of the above described services and demonstrations into a single environment.
The first version of the DARIAH WS-PGRADE gateway is planned to be released as an online service during 2016, with new features to be added based on user request and feedbacks during 2017.

4. Science Gateway for Lexicographers

Lexicographers are invited to exploit the above services.
The following case studies discuss the potential of the services intended for lexicography.
Parallel semantic search engine (PSSE):
1. Semantics are essential when interlinking languages and cultural content, as well as they are core of lexicographical work ("definition / meaning").
PSSE is an easy accessible tool, allowing lexicographers to quickly overview a word and its potential uses in the chosen repositories and data collections. Opening up new (lexico-graphical) resources to the PSSE-service would further strengthen interdisciplinary collaboration and visability of the lexicographers work in other research communities.
On the example of queries for a certain common name for living organisms, e.g. daisy / Bellis perennis, scientific usage as well as non-scientific usage of the word "daisy" might be explored and be studied in European cultural context (or even global context). Naming concepts in different disciplines might be discussed. The virtual collaboration group on biodiversity and linguistic diversity coordinated by the Austrian Academy of Sciences offers an anchor to explore related research questions.
Furthermore, based on this rich dataset, data modelling for linked data framework can be discussed.
gLibrary offers a repository framework for lexicographical data.
In the use case of Bavarian dialects dataset it is explored for cloud storage of lexicographical resource data. A rich collection of paper slips and questionnaires (1911-1998) is stored on the cloud, easily accessible for internal staff members, as well as for collaboration partners.
The easy access to different data-formats, the result of a long history in lexicography compiling on the one hand and explorative lexicography on the other, offers rich potential to explore open innovation in the framework of lexicography. Visual interfaces allow to modify queries on complex (heterogeneous) data sets in a user friendly way.
Furthermore, the infrastructure framework allows the lexicographer to concentrate on his/her lexicographical work as well as on exploring new methods, whereas the cloud-repository offers stable technologies for sustainability concerning data access, storage etc.
Nowadays, lexicography usually is based on (digital) corpora. Lexicographers, exploring and challenging emerging technologies to answer their research questions, might make use of CDSTAR engines to process very large datasets / collections, e.g. analytics platform which implements Hadoop / Spark.
Additionally, lexicographers can benefit by running computing-intensive applications on EGI cloud, either directly via EGI application marketplace, long tail of sciences platform or via (future) DARIAH-Competence Center gateway. This gateway will offer easy-going user interface which will not require technical knowledge on cloud technologies from end users, e.g.

lexicographers.

5. Conclusion

To design "complex, sustainable digital humanities projects and publications requires familiarity with both the research subject and available technologies" [Schreibman et al. 2004].
In this paper we offer insight into selected EGI-technologies and discuss possible ideas on their use in lexicography. To support familiarity with the technologies, EGI-ENGAGE offers tutorials.
Especially for the case studies, similar to those discussed above, exploitation of the existing services and tools seems promising and easily adaptable.
The infrastructure services introduced offer examples to make use of emerging technologies and technological innovation even for small lexicography data providers or students. In doing so, we interpret this as a good development towards strengthening cultural diversity via supporting linguistic diversity or minority language lexicography.
Furthermore, there are social infrastructures emerging, allowing us to strengthen collaboration and intercultural learning such as COST ENeL, exploreAT!, Digital language diversity project (DLDP) as well as the just recently established ELRA working group on "Lesser resourced languages".
Both social infrastructures and technological innovation are key essentials to design explorative cultural lexicography in an iterative, open, interdisciplinary process.

References

Boivet, M., Chambaut, O. (2002): *Unesco Universal Declaration on Cultural Diversity*. http://unesdoc.unesco.org/images/0012/001271/127160m.pdf [accessed: 2016.05.28].

European Commission / COST (2013): *Memorandum of understanding for a European concerted research action designated as COST action IS 1305 "European Network of electronic Lexicography (ENeL)"* http://w3.cost.eu/fileadmin/domain_files/ISCH/Action_IS1305/_mou/IS1305-e.pdf [accessed: 2016.05.28].

European Council / European centre for modern languages – European Day of languages (2016): *The celebration of linguistic diversity.* http://edl.ecml.at/Home/Thecelebrationoflinguisticdiversity [accessed: 2016.05.28].

Greenberg, J.H. (1956): *The Measurement of Linguistic Diversity*. Language 32 (1): pp 109-115.

Hanks, P. (2012): *Lexicography from earliest times to the present*. In: Allan, K. (Ed.): The Oxford Handbook of the History of Linguistics. http://www.patrickhanks.com/uploads/5/1/4/9/5149363/2012dlexicography_from_earliest_times.pdf [accessed: 2016.05.28].

Kálmán, T., Tonne, D., Schmitt, O. (2015) *Sustainable Preservation for the Arts and Humanities*. New Review of Information Networking 20, 1-2, pp. 123–136. doi:10.1080/13614576.2015.1114831.

Kacsuk, P. et al. (2012): *WS-PGRADE/gUSE Generic DCI Gateway Framework for a Large Variety of User Communities*. Journal of Grid Computing 10/4, pp. 601-630.

Schreibman, S., Siemens, R., Unsworth, J. (Ed.; 2004): *A Companion to Digital Humanities*.

United Nations Educational Cultural and Scientific Organisation (2009) Unesco-Report: *Investing in Cultural Diversity and Intercultural Dialogue*. http://unesdoc.unesco.org/images/0012/001271/127160m.pdf [accessed: 2016.05.28].

Wandl-Vogt, E. (Ed.; 2010): *Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)*. https://wboe.oeaw.ac.at/projekt/beschreibung/ [accessed: 2016.05.28].

Acknowledgements

The work presented in this paper has been supported by the EGI-Engage H2020 project (Grant number 654142), COST IS 1305 ENeL and explore.AT!, funded by Nationalstiftung für Forschung, Technologie und Entwicklung (Nationalstiftung FTE) at the Austrian Academy of Sciences.