Report

This workshop, entitled *Digital Infrastructure Projects and What they already offer historical languages*, took place on May 9 and 10, 2017, at the Göttingen Centre for Digital Humanities in Göttingen, Germany. Its primary purpose was to focus attention in the Global Philology Planning Project on existing infrastructure projects, such as CLARIN, DARIAH, Europeana, and the German Digital Library, and the resources that these projects can offer to scholars working in historical languages who use digital methods and the role these infrastructures could play in any future Global Philology Project. The agenda for the meeting can be found online[1] and is also appended to this document.

As one can see from the agenda, the first day of the project was given over to talks about possible solutions and resources for scholars working with historical languages. These solutions were presented both by scholars within these fields as well as by the large infrastructures themselves. Since the abstracts and slides for all of these talks will be made available online, this workshop report will concentrate more on the discussion that happened on the second day of the workshop.

At the beginning of the discussion, it was made clear that the goal of any subsequent project that results from the Global Philology Planning Project will be to make existing solutions useful for a wider audience in historical languages. There will also be a strong focus on Open Data, i.e., not just granting users read access to your data but allowing them to download it, change it, and republish it as they see fit.

The conversation started by discussing several existing solutions for the creation of Digital Critical Editions. Besides the representatives of the PANDORA system and TextGrid, who gave presentations on the first day, there was also a representative of the Virtual Manuscript Room (VMR) in attendance. This representative talked about the functionality and extensibility of the the VMR and how it was presently being used both for New Testament Greek and Old Testament Coptic. Besides these solutions, the Digital Latin Library and Transcribus were discussed as possible solutions. It was decided that the first important step in this area would be to fully describe the existing solutions with example use cases, focusing especially on the strengths of each of these platforms.

There followed a brief conversation about Optical Character Recognition. It was pointed out that, for instance, high accuracy recognition of polytonic Greek and classical Arabic has already been achieved. But such high accuracy figures are difficult to generalize since they are typically reached by focusing not only on a specific language but even on a specific type font used by a specific publisher. The idea was put forward that one resource that an infrastructure could provide in this area would be some sort of OCR training platform.

The question of annotations of existing digital objects was also discussed. One solution, which is used by the PANDORA platform, is to use the IIIF standard to precisely annotate images. In

---

1 http://www.dh.uni-leipzig.de/wo/events/digital-infrastructure-projects-and-what-they-already-offer-historical-languages/

terms of textual annotation, there are several possible solutions that exist, including the WebAnno platform from CLARIN. It was also stressed here that any such annotations need to be made available long term, both to the annotator and to the other users of the digital object upon which the annotation is.

At this point, the conversation moved to focus on the different infrastructures that were present at the conference: DARIAH, CLARIN, and Laudatio. Unfortunately, the representative for Europeana and the German Digital Library had to withdraw from the workshop at the last minute for health reasons. Of great interest from Laudatio was the way this platform deals with data and metadata in specific formats. Instead of requiring the scholars producing the data to conform to a specific standard that may not fulfill their needs, Laudatio uses a metadata metaschema to allow the transfer of data from one standard to another. This allows the scholar to produce data in the format they wish as long as they map their own schema to the metaschema. Once this mapping is done, the conversion from one format to any other format that is also mapped to this metaschema is quite easy. In this respect, the DTA Basisformat for historical texts was mentioned.[2]

DARIAH and CLARIN both stressed that they believed they could make the greatest impact in the Global Philology effort by having detailed conversations with the groups involved in the project about the data workflow that each specific group uses. This would allow both of these infrastructures to recognize where they could be of service to these groups either by integrating solutions from the scholarly group into their infrastructure or offering existing solutions from their infrastructure to the scholars involved. Specific examples of the latter were WebLicht from CLARIN, which could be easily customized to use language-specific linguistic tools from any of these historical languages and the authorization services to control access to certain resources and the Open API services to expand the accessibility of tools that DARIAH and CLARIN offer. The representative from DARIAH-EU also stressed the need to find ways to bridge the gap between the generally bespoke solutions that scholars create for their own use and the high-level services that both DARIAH and CLARIN offer.

There was also intense discussion that happened during the coffee and lunch breaks and during the communal dinner on the first evening. The focus was clearly on producing open data that can contain all of the information that a scholar would need to work with the texts involved, from annotations on linguistic attributes to extended sections of complex commentary. In the end, this was a very successful workshop that offered multiple possibilities to involve these large infrastructures in the future efforts on the Global Philology front. The conversations that were started here will continue over the next several months in preparation any subsequent project proposals that result from this planning project.

[2] http://www.deutschestextarchiv.de/doku/basisformat/

Agenda

Tuesday, May 9

9-9:30 – Gregory Crane

Introduction

9:30-10 – Marco Büchler

A Ten-Year Summary of a SOA-based Micro-services Infrastructure for Linguistic Services

Abstract

From 2004 to 2016 the Leipzig Linguistic Services (LLS) existed as a SOAP-based cyberinfrastructure of atomic micro-services for the Wortschatz project, which covered different-sized textual corpora in more than 230 languages. The LLS were developed in 2004 and went live in 2005 in order to provide a webservice-based API to these corpus databases. In 2006, the LLS infrastructure began to systematically log and store requests made to the text collection, and in August 2016 the LLS were shut down. This article summarises the experience of the past ten years of running such a cyberinfrastructure with a total of nearly one billion requests. It includes an explanation of the technical decisions and limitations but also provides an overview of how the services were used.

10-10:30 – Pietro Liuzzo

EAGLE and IDEA, the International Digital Epigraphy Association: tasks and activities

Abstract

After the end of the EAGLE project (Europeana Network for Ancient Greek and Latin Epigraphy) the association IDEA (International Digital Epigraphy Association) was founded to maintain the services developed and to offer continuity to the networking activities. IDEA aims at continuing the mission of the EAGLE network in supporting small and medium projects with advice and services as well as to keep the aggregated epigraphic data to the best possible standard. The portal and services which provide search across multiple aggregated databases, vocabularies for authority files, and the Story Telling application continue to live and new perspectives open up for a future epigraphy.info project based on the model of papyri.info.

Beta maṣāḥəft and the Ethiopian literary tradition in the digital age

Abstract

The project Beta maṣāḥəft founded by the Academy of sciences aims at taking philological and codicological studies on the Ethiopian literary tradition to the digital age starting with the encoding of primary sources in TEI. The project will build a catalogue of manuscripts, the first Clavis of the Ethiopian literature, a gazetteer of ancient places in Ethiopia and a prosopography

of Ethiopian people. Not an easy task dealing with a still leaving tradition and with scarce access to the sources.

10:30-11 – Coffee break

11-11:30 – Carolin Odebrecht

Corpus metadata for the reusability of historical corpora

Abstract

This talk will address the question of how we can support the reusability of historical corpora with the help of corpus metadata. Historical corpora vary considerable concerning their annotations and formats. Reusing a historical corpus is therefore a challenging task which requires a deep understanding of the corpus architecture and its content. In my talk, I will present our approach to solve this issue with the help of a meta model for corpus metadata. This meta model will provide both the necessary abstraction from the data and the relevant information needed to enable reusability scenarios.

11:30-12 – J. K. Tauber

Greek Linguistic Databases for Better Learning Tools

Abstract

Language learning tools, from vocabulary drills to adaptive reading environments need both models of student knowledge and rich linguistic databases tied to texts. In this talk I will discuss how richer linguistic databases enable better learning tools and how learning tools can motivate better linguistic databases. The focus of examples will be my own work on the Greek New Testament but the ideas and infrastructure discussed will be applicable to the much broader Greek corpus as well as other languages.

12-12:30 – Tariq Yousef

Ugarit Translation Alignment Editor and Dynamic Lexicon

Abstract

Ugarit is an online tool for manual text alignment, users can import texts from perseus cts repository or use their own texts, the editor enables users to align two or three parallel translations in a very simple way. Ugarit serves also as a reading environment for parallel texts, it visualises the aligned texts in very simple and meaningful way showing parallel translation pairs and their frequencies with the ability to export the alignment as XML files or the translation pairs as CSV files. The translation pairs obtained from the manual alignment are used to build a dynamic lexicon.

12:30-1 – Christopher H. Johnson & Jörg Wettlaufer

The PANDORA Linked Open Data Presentation Framework

Abstract

The interconnection of data in the Humanities gets more and more in the focus of research projects. Therefore Christopher Johnson developed a Linked Open Data framework that allows through the combination of a Fedora 4 repository with IIIF APIs and triple stores a SPARQL query driven solution for the Presentation (of) Annotations (in a) Digital Object Repository Architecture (PANDORA). The concrete implementation of PANDORA is a group of distributed web applications that depend on a specification document called a “Manifest” for how they present the data to the client. In PANDORA, the Manifest is a JSON-LD document constructed from Digital Object Repository (FEDORA) resources dynamically using SPARQL. The semantics and conceptualization of the Manifest are in the scope of the IIIF Presentation API, within which is defined how the structure and layout of a complex image-based object can be made available in a standard manner. In this short presentation we will present the architecture and function of the system and like to discuss the possible usage in philological research.

1-2 – Lunch

2-2:30 – Stephan Bartholmei (CANCELED)

German Digital Library and Europeana

2:30-3 – Thorsten Trippel

CLARIN

Abstract

CLARIN-D is a research infrastructure for the humanities and social sciences. The infrastructure provides easy and sustainable access for scholars to digital language data (in written, spoken, video or multimodal form) and to advanced tools to discover, explore, exploit, annotate, analyze or combine them. For projects and scholars creating and using data, CLARIN supports their data management with services for preparation and for depositing of data. In this presentation, we will show selected services and resources provided by CLARIN that can be utilized for historical languages and that can be integrated by scholars in their own projects and software.

3-3:30 – Susanne Haaf

Historical German Data in CLARIN’s user involvement phase: status and perspectives

Abstract

The Language Center at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) constitutes the CLARIN-D data center for historical German text. It is home to the German Text Archive (Deutsches Textarchiv, DTA), a platform for the publication and exploitation of historical corpora for the German language. The talk will focus on the components offered by CLARIN-D via the BBAW data center that enable and support work with historical data and corpora, including standard formats, documentation, research and archiving platforms, as well as analysis tools. In addition, the talk will outline perspectives for further developments with regard to historical data during CLARIN’s user involvement phase which has started in September 2016 and follows up to CLARIN’s implementation phase.

3:30-4 Coffee

4-4:30 – Mike Mertens

What makes DARIAH-EU, DARIAH-EU? (Or, how I stopped worrying about definitions and learned to love Research Infrastructures)

Abstract

When one first hears the word “infrastructure” in a research or academic context, one perhaps immediately visualises something akin to CERN: a large, multinational complex that manages a purely physical and often unique asset. DARIAH-EU has long however insisted that the knowledge inherent in Arts and Humanities endeavours requires three distinct elements to flourish – people, skills as well as hardware. Each of these requires a clear framework and shared resources.

In the presentation I will give a brief history of DARIAH-EU, will focus on the importance of international cooperation; the necessary interrelationship between research, archives, museums and other memory institutions, and the broader public; the link with the creative industries, what DARIAH offers in terms of skills and training, and how we see the future of sustainable digitally-enabled Arts and Humanities.

4:30-5 – Stefan Schmunk

DARIAH-DE – Generic usable components for disciplinary requirements

Abstract

DARIAH-DE offers a variety of IT components and tools, which can be used by research projects and institutions and integrated into their own developments. This includes a number of basic components, but also a number of generically usable special tools and services. This includes, for example, services from the fields of annotations, big data and research data. Within the scope of the lecture, some of these components are to be presented and at the same time the requirements of the project are to be determined.

5-5:30 – Jan Brase

The research data Alliance (RDA)

Abstract

The RDA was founded in 2013 as a research community through a joint effort of the European Commission, the American National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation. The RDA defines itself not as a digital infrastructure project, but as a global platform to bring together specialists for research data issues. RDA’s main vehicle for outputs are 18-month long working groups that generate recommendations aimed at the RDA community. In addition to working groups, interest groups with no fixed lifetime can produce either informal or “supported” outputs which carry some degree of RDA endorsement.

In this overview we will have a look at those working groups that are of interest for the field of historic languages and discuss how to cooperate with them

Wednesday, May 10

9am-1pm

Wednesday will be given completely over to discussions, which can include whole-group discussions as well as breakout sessions for participants wanting to focus on specific issues or technical solutions.

Participants:

Prof. Heike Behlmer, Universität Göttingen
Jan Brase, SUB Göttingen
Marco Büchler, Universität Göttingen
Shih-Pei Chen, MPIWG Berlin
Prof. Camilla Di Biase-Dyson, Universität Göttingen
Prof. Gregory Crane, Universität Leipzig
Jakob Epler, DARIAH-EU
Maryam Foradi, Universität Leipzig
Susanne Haaf, BBAW Berlin
Brent Ho, MPIWG Berlin
Chris Johnson, Akademie der Wissenschaften zu Göttingen
Pietro Liuzzo, Universität Hamburg
Mike Mertens, DARIAH-EU
So Miyagawa, Universität Göttingen
Matthew Munson, Universität Leipzig
Carolin Odebrecht, Humboldt Universität zu Berlin
Ulrich Schmid, Universität Göttingen
Stefan Schmunk, DARIAH-DE
Masoumeh Seydi, Universität Leipzig
James Tauber, Eldarion
Thorsten Trippel, CLARIN-D
Jörg Wettlaufer, Akademie der Wissenschaften zu Göttingen
Tariq Yousef, Universität Leipzig