JOCCH Contribution, University of Leipzig

Authors: Gerhard Heyer, Marco Büchler, Volker Boehlke, Charlotte Schubert

Provisional Title: Aspects of an Infrastructure for eHumanities

1. Introduction

Computer Science and Humanities so far have acted in their working methodologies more as antipodes rather than focusing on the potential synergies. However, recent advances in digitizing historical texts, and the search and text mining technologies for processing these data indicate an area of overlap that bears great potential. For the humanities the use of computer based methods may lead to more efficient research (where possible) and the raising of new questions that without such methods could not have been dealt with. For computer science turning towards the humanities as an area of application may pose new problems that also lead to rethinking present approaches hitherto favored by computer science and developing new solutions that help to advance computer science also in other areas of media oriented applications. But most of these solutions at present are restricted to individual projects and do not allow the scientific community in the digital humanities to benefit from other advances in computer science like service engineering. Hence, in this paper we shall attempt to spell out in detail the idea of an infrastructure for eHumanities. Focusing on the notion of re-usability of data and algorithms such as morphological annotation, we sketch how a loosely coupled infrastructure (SOA) helps to reduce the costs while increasing the efficiency of such projects.

We begin by discussing the impact of digitization in the humanities, and propose a distinction between Digital Humanities and eHumanities. The problem of re-using data and algorithms is being discussed by reference to the eAQUA project where data and algorithms have been implemented as SOAP as well as REST based webservices. A detailed report and discussion of these infrastructure aspects is being presented in the final part of the paper.

2. The impact of digitization in the humanities – From Digital Humanities to eHumanites

We review the main areas of computer science applications and discuss its implications for the digitization of humanities. To the extent that applications of computer science always lead to a replacement of analogue by *digital* media and processes, digital media, and processing models have an increasing impact on traditional work flows based on analogue media, also creating new methods and applications. As a first definition of *eHumanities* we propose the notion of digitizing media and work flows in the Humanities. We then discuss in detail a distinction we propose between *Digital Humanities* and *eHumanites*, and relate it to ongoing projects. We conclude by raising the question of where, exactly, there might be synergies between computer science and the humanities that are beneficial for either science, and present two case studies for discussion, eAQUA and the CLARIN infrastructure developments.

3. Project eAQUA

eAQUA is an interdisciplinary eHumanities project set up between the departments for classical studies at the University of Leipzig, Heidelberg, and Hamburg, and the division for natural language processing at the computer science department of the University of Leipzig. During the last years a number of tools such as *Diogenes, Hopper, View&Find,* or *Lector* were developed that give users a dedicated but restricted access to different text corpora. In addition, different data such as the *Lexicon of Greek Personal Names*, or *Perseus' Lidell Scott Jones,* also are available. In most cases, however, such tools often impose copy-right restrictions that correlate to a barrier of their re-usability. Hence, the re-usability of work already done in practice is often much more difficult as it first may seem.

In addition to the text mining technologies that are used in the eAQUA sub-projects, both data and tools are provided by a SOAP as well as REST based webservice. In view of such infrastructure aspects, data, tools (algorithms), and user interfaces are kept separate and combined only depending on a user's intended application. This dramatically increases the re-usability of data, tools, and user interfaces.

4. Infrastructure

We finally address the issue of infrastructure development. We claim that we need a transition from accidental to organized cooperation, and propose to set up *Centers of competence* that mediate the technical and organisational issues. A platform of services like the CLARIN infrastructure allows all participants to share digital resources and helps to establish a culture of best practices. Resources can be implemented as webservices for sharing data and algorithms (based on SOAP). By way of example, we describe in detail the external and internal interfaces of eAQUA, how text mining applications are based on this RDB, and how these applications are realized as web services. We describe a list of services that are provided by SOAP as well as REST. Drawing on 6 years of experience in providing linguistic services we highlight typical first user problems, report lessons learnt, and compare SOAP and REST by usage scenarios. We finally discuss the chaining of webservices by comparing two scenarios for service chaining as an example of how this infrastructure can be used in practical work:

- REST: Given a set of REST based web services (segmentation, tokenisation, NER), a TEI P5 pre-processing chaining will be described that covers about 50% of the total cost in a text mining project.

- SOAP: A set of SOAP services as normalization (upper- and lowercase letters as well as dealing with different diacritics), or lemmatization, is used to highlight the simplicity of building a SOAP based search engine.

5. Summary and conclusion

In summary, we wrap up our main theses for discussion on software engineering in the humanities, and infrastructure building (using distributed resources, standards, metadata, webservices).