

GERHARD HEYER / VOLKER BOEHLKE

Text mining in the Humanities – A plea for research infrastructures

Abstract

Research infrastructures for the Humanities can help to share digital resources and content services. In particular, they can help researchers in the Digital Humanities to save time and efforts when developing software to deal with specific research issues. Web services and web applications can be used to build a research infrastructure for sharing data and algorithms. However, the development of such infrastructures and their key software components is a software engineering task that increasingly also poses interesting and challenging research problems for Computer Science.

1. Text, knowledge, and the Humanities

As manifold as the usages of language are the purposes of text. But when looking at text in the Humanities, it looks to me as a Computer Scientist that we are, broadly speaking, always assuming that the texts we are interested in are encodings of knowledge (of a culture at a time). And this is what makes texts the subject of analysis: By looking at texts (and sometimes also at their context of origin) we intend to decipher the knowledge that they are encoding.

Looking at texts from a bird's eye view or taking a close reading perspective has always been the core business of text oriented Humanities. With the advent of Digital Humanities, however, we can scale up this task by using new analysis tools derived from the area of information retrieval and text mining. Thereby all kinds of historically oriented text sciences as well as all sciences that work with historical or present day texts and documents are enabled to ask completely new questions and deal with text in a new manner. In detail, these methods concern, amongst others,

- the qualitative improvement of the digital sources (standardization of spelling and spelling correction, unambiguous identification of authors and sources, marking of quotes and references, temporal classification of texts, etc.);
- the quantity and structure of sources that can be processed at scale (processing of very large amounts of text, structuring by time, place, authors, contents and topics, comments from colleagues and other editions, etc.);

- the kind and quality of the analysis (broad data driven studies, strict bottom-up approach by using text mining tools, integration of community networking approaches, contextualization of data, etc.).

While Computer Science and Humanities so far have acted in their working methodologies more as antipodes rather than focusing on the potential synergies, with the advent of Digital Humanities we enter a new area of interaction between the two disciplines. 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 advances in other areas of Computer Science like Visual Analytics.

In consequence, I think it is important that we distinguish between two important aspects:

- (1) the creation, dissemination, and use of digital repositories, and
- (2) the computer based analysis of digital repositories using advanced computational and algorithmic methods.

While the first has originally been triggered by the Humanities and is commonly known as Digital Humanities, the second implies a dominance of computational aspects and might thus be called Computational Humanities.

To distinguish between both aspects has substantial implications on the actual work carried out. Considering the know-how of researchers and their organizational attachment to either Humanities or Computer Science departments, their research can either be more focused on just the creation and use of digital repositories, or on real program development in the Humanities as an area of applied Computer Science, as is illustrated by Figure 1.

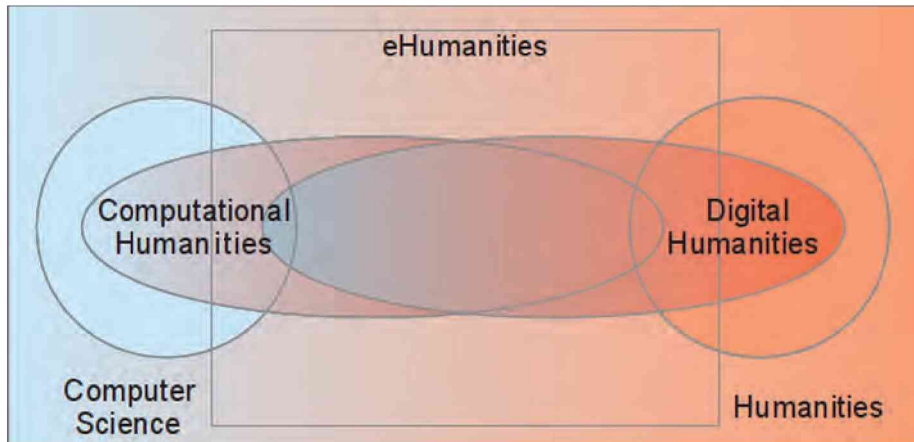


Figure 1: Positioning of Computational and Digital Humanities in the context of Computer Science and Humanities

A practical consequence also in organizational terms of this way of looking at things would be to set up research groups in both scientific communities, Computer Science and Humanities. The degree of mutual understanding of research issues, technical feasibility and scientific relevance of research results will be much higher in the area of overlap between the Computational and Digital Humanities than with any intersection between Computer Science and the Humanities.

2. Research infrastructures for the Humanities

Now, in order to use such computational methods, an individual researcher can proceed by employing two strategies depending on his, or her, own degree of computer literacy. On the one hand, there is the individual software approach. Given a selection of digital text data, the research question is being transferred into a set of issues and methods that can be dealt with by a number of individual programs. This approach allows for a highly dynamic and individual development of research issues. It requires, however, a high degree of software engineering know-how. On the other hand, there is the approach to use standard software. For well-defined and frequently encountered tasks, a Research Infrastructure will offer solutions that provide the users with data and analysis tools that are well understood, have already delivered convincing results, and can be learnt without too much effort.

Both approaches are interdependent. Probably good solutions in one domain of text oriented Humanities can be transferred to other domains by just applying these methods to different kinds of text. A good infrastructure must be capable of making such solutions accessible as best practices.

Research infrastructures are concerned with the systematic and structured acquisition, generation, processing, administration, presentation, reuse, and publication of contents. Content services make available the resources and programs needed for that. Public digital text and data resources are linked together and made accessible by common standards. New software architectures integrate digital resources and processing tools to develop new and better access to digital contents. A good example at hand is the ESFRI initiative CLARIN that aims at “Providing linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities”¹. CLARIN-D, the German sub-project, is primarily designed as a distributed, centers based project. This means that centers are at the heart of an infrastructure that aims at providing persistent data services (see also Boehlke/Heyer/Wittenburg 2013). Different types of resource centers form the backbone of the infrastructure and provide access to data and metadata and/or run infrastructure services. Access to data, metadata and infrastructure services is usually (but not solely) based on web services and web applications. The protocols and formats of infrastructure services like persistent identifiers or metadata systems and standards have already been agreed upon in an early stage of the project. Additional infrastructure or discipline specific services are built upon those basic infrastructure services. The usage of general services like registering and resolving persistent identifiers, however, is not limited to CLARIN itself. The usage of such common services by other infrastructure initiatives is intended and already in place.

Research infrastructures for the Humanities can also help to reconcile the current debate of where the Digital Humanities should be institutionalized – individual Humanities departments or, more central, at Computer Science. On the one hand, there will clearly be no gain for the scientific community in Digital Humanities as a whole when know-how will be duplicated at different Humanities departments. On the other hand, it will be difficult to foster Digital Humanities applications in the Humanities, and to develop new ones, unless the digital research is driven by the Humanities themselves. Deciding for one

¹ <http://de.clarin.eu/en/home-en.html> (last accessed: January 29, 2015).

or the other alternative is no solution, as either way will be hampered by massive drawbacks. However, the building, furnishing, and maintaining of a research infrastructure for the Humanities clearly is a task for Computer Science, while the use of the research infrastructure clearly must be left to the individual researchers and communities in the Humanities themselves. This way we obtain a division of labour that has proven to be most useful in other areas of applied Computer Science and that can lead to substantial and rich contributions of the Humanities to the field of Digital Humanities without getting involved in programming and Computer Science beyond necessity.

3. Infrastructure text mining services

Making language resources and language processing tools available as services enables researchers and developers to use exactly the amount and kind of data that is needed for a specific application. As an example from the area of linguistics, let us finally consider the Webservice access to digital text and lexical data as well as NLP algorithms and tools that was established at the Natural Language Processing Department of Leipzig University already in 2004 (Biemann et. al. 2007, Böhler/Heyer 2009). These services, Leipzig Linguistic Services (LLS) for short, comprise, amongst others,

- a very large, frequency sorted dictionary of German word forms including POS information, sample sentences and co-occurrences,
- monolingual corpora of standard size for currently 48 different languages,
- a tool for sentence boundary detection,
- graph based clustering,
- co-occurrence statistics,
- synonyms and similar words computed on co-occurrence profiles of words,
- automatic terminology extraction, and
- named entity recognition.

Let us assume that a scientist wants to use one of these tools, viz. the named entity recognition, on specific parts of a collection of texts that is encoded in TEI-P5. The task at hand is to extract the needed information from the TEI-P5 document collection, encode it in a way the Named Entity Recognizer web service is able to work on, fetch and interpret the results and probably also to perform a manual correction. In the end, the result is intended to be published

in a way that allows other scientists to validate or reuse the process through reiteration.

Research infrastructures give support on multiple levels depending on the know how of the researcher. Instead of installing, configuring and using a set of offline tools, a scientist who is part of the Digital Humanities community is able to work with programmes that are provided in the form of existing web applications facilitating the usage of low level functionality like converting TEI-P5 documents into simple text or other formats, invoking a Named Entity Recognizer webservice, sending the results to an online annotation platform and archiving the results in a repository. Researches with this level of know-how may be limited in the usage of research infrastructures since they mostly have to rely on the existence of all those tools that make up the single steps of the workflow described above. But they do not need to maintain their own local software stack and they are also able to work in an environment that allows their research to be reproducible.

A scientist who works in Computational Humanities may tap the full technical potential of research infrastructures by creating himself new web services and bundling existing and new application specific algorithms in web applications. This process results in new or alternative workflows becoming available for the whole scientific community. When doing so, the planning, implementation, and deployment process is getting more efficient due to the fact that it is possible to build upon the basic functionality that the research infrastructure provides. Just to name a few:

- Instead of implementing, deploying and hosting one's own version of a simple storage facility that allows to store intermediate results, a common workspace concept of the research infrastructure can be used that is compatible with other tools and services deployed in the infrastructure.
- Exhaustive documentation on how to generate metadata and how to plug services into the infrastructure is available, reducing the time needed to interact with other components.
- Basic concepts and services that allow to make workflows reproducible are in place (e.g. PID systems).
- The question on where to host services that cannot be run by the researcher himself (due to lack of hardware, legal reasons, lack of time to provide long-term support, ...) is answered.

References

- Biemann, Chris/ Heyer, Gerhard/Quasthoff, Uwe/Richter, Matthias (2007): The Leipzig Corpora Collection: monolingual corpora of standard size. In: Proceedings of Corpus Linguistics 2007, Birmingham, UK. www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/190Paper.pdf (last accessed: January 29, 2015).
- Boehlke, Volker/Heyer, Gerhard/Wittenburg, Peter (2013): IT-based research infrastructures for the Humanities and Social Sciences – Developments, examples, standards, and technology. In: *it – Information Technology* 55(1): 26-33.
- Büchler, Marco/Heyer, Gerhard (2009): Leipzig Linguistic Services – A 4 years summary of providing linguistic web services. In: Heyer, Gerhard (ed.): *Text Mining Services – Building and applying text mining based service infrastructures in research and industry.* (= Leipziger Beiträge zur Informatik XIV). Leipzig: Universität Leipzig, 55-65.