Report from Dagstuhl Seminar 14301

# Computational Humanities - bridging the gap between Computer Science and Digital Humanities

**Edited by**

**Chris Biemann[1], Gregory R. Crane[2], Christiane D. Fellbaum[3], and Alexander Mehler[4]**

1 **TU Darmstadt, DE,** biem@cs.tu-darmstadt.de
2 **Tufts University, US,** gregory.crane@Tufts.edu
3 **Princeton University, US,** fellbaum@princeton.edu
4 **Goethe-Universität Frankfurt am Main, DE,** mehler@em.uni-frankfurt.de

## Abstract

Research in the field of Digital Humanities, also known as Humanities Computing, has seen a steady increase over the past years. Situated at the intersection of computing science and the humanities, present efforts focus on making resources such as texts, images, musical pieces and other semiotic artifacts digitally available, searchable and analysable. To this end, computational tools enabling textual search, visual analytics, data mining, statistics and natural language processing are harnessed to support the humanities researcher. The processing of large data sets with appropriate software opens up novel and fruitful approaches to questions in the traditional humanities. This report summarizes the Dagstuhl seminar 14301 on "Computational Humanities - bridging the gap between Computer Science and Digital Humanities".

**Seminar** *20.–25. July, 2014* – http://www.dagstuhl.de/14301
**1998 ACM Subject Classification** *I.2.7 Natural Language Processing, J.5 Arts and Humanities*
**Keywords and phrases** *Computer Science, Digital Humanities, Computational Humanities, eHumanities, Big Data, Experimental Methods*
**Digital Object Identifier** 10.4230/DagRep.1.1.1



Computational Humanities - bridging the gap between Computer Science and Digital Humanities,
*Dagstuhl Reports*, Vol. 1, Issue 1, pp. 1–31
Editors: Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

# 1 Executive Summary

*Chris Biemann, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler*



## 1.1 Motivation

Research in the field of *Digital Humanities*, also known as *Humanities Computing*, has seen a steady increase over the past years. Situated at the intersection of computing science and the humanities, present efforts focus on building resources such as corpora of texts, images, musical pieces and other semiotic artifacts digitally available, searchable and analyzable. To this end, computational tools enabling textual search, visual analytics, data mining, statistics and natural language processing are harnessed to support the humanities researcher. The processing of large data sets with appropriate software opens up novel and fruitful approaches to questions in the 'traditional' humanities. Thus, the computational paradigm has the potential to transform them. One reason is that this kind of processing opens the way to *new* research questions in the humanities and especially for *different* methodologies for answering them. Further, it allows for analyzing much larger amounts of data in a quantitative and automated fashion – amounts of data that have never been analyzed before in the respective field of research. The question whether such steps ahead in terms of quantification lead also to steps ahead in terms of the quality of research has been at the core of the motivation of the seminar.

Obviously, despite the considerable increase in digital humanities research, a perceived gap between the traditional humanities and computer science still persists. Reasons for this gap are rooted in the current state of both fields: since computer science excels at automating repetitive tasks regarding rather low levels of content processing, it can be difficult for computer scientists to fully appreciate the concerns and research goals of their colleagues in the humanities. For humanities scholars, in turn, it is often hard to imagine what computer technology can and cannot provide, how to interpret automatically generated results, and how to judge the advantages of (even imperfect) automatic processing over manual analyses.

To close this gap, the organizers proposed to boost the rapidly emerging interdisciplinary field of *Computational Humanities* (CH). To this end, they organized a same-named Dagstuhl Seminar that brought together leading researchers in the fields of Digital Humanities and related disciplines. The seminar aimed at solidifying CH as an independent field of research and also at identifying the most promising directions for creating a common understanding of goals and methodologies.

At the core of the organizers' understanding of CH is the idea that CH is a discipline that should provide an algorithmic foundation as a bridge between computer science and the humanities. As a new discipline, CH is explicitly concerned with research questions from the humanities that can more successfully be solved by means of computing. CH is also concerned with pertinent research questions from computing science focusing on multimedia content, uncertainties of digitisation, language use across long time spans and visual presentation of content and form.

In order to meet this *transdisciplinary* conception of CH, it is necessary to rethink the roles of both computer scientist and humanities scholars. In line with such a rethinking, computer scientists cannot be reduced to software engineers whose task is just to support humanities scholars. On the other hand, humanities scholars cannot be compelled to construe post-hoc explanations for results from automatic data analysis. Rather, a common vision –

shared among both groups of scientists – is needed that defines and exemplifies accepted methodologies and measures for assessing the validity of research hypotheses in CH. This vision motivated and formed a common ground for all discussions throughout the seminar.

## 1.2 Goals and Content of the Seminar

In order to elaborate the vision of CH as a bridge between computer science and the humanities, the seminar focused on questions that can be subsumed under four different reference points of problematizing CH:

1. **The Present State: What works, what does not?**
   - Review of the success of the last 10 years of the digital humanities: Can we identify commonalities of successful projects? What kinds of results have been obtained? What kinds of results were particularly beneficial for partners in different areas of research? Can success in one field be transferred to other fields by following the same methodology?
   - Review of the challenges of the last 10 years of the digital humanities: What are recurring barriers to efficient cross-disciplinary collaboration? What are the most common unexpected causes of delays in projects? What are common misunderstandings?
   - What is the current role of computer scientists and researchers in the humanities in common projects, and how do these groups envision and define their roles in this interplay?
2. **Computational Challenges in Computational Humanities:**
   - What research questions arise for computational scientists when processing data from the humanities?
   - How can the success of a computer system for humanities data-processing be evaluated to quantify its success?
   - What are the challenges posed by the demands from the humanities? In particular, how can computer scientists convey the notion of uncertainties and processing errors to researchers in the humanities?
3. **Humanities Challenges in Computational Humanities:**
   - What research questions can be appropriately addressed with computational means?
   - How can we falsify hypotheses with data processing support?
   - What is and is not acceptable methodology when one relies on automatic data processing steps?
4. **Common Vision: Algorithmic Foundations of Computational Humanities:**
   - Can we agree on generic statements about the expressivity of the range of algorithms that are operative in the digital humanities and related fields of research?
   - Can we distinguish complexity levels of algorithms in the computational humanities that are distinguished by their conditions of application, by their expressiveness or even explanatory power?
   - Which conditions influence the interpretability of the output generated by these algorithms from the point of view of researchers in the humanities?

## 1.3 The Program

In order to work through our set of goals (see section 1.2), the seminar decided for a mixture of talks, working groups and plenary discussions. To this end, four Working Groups (WG) have been established whose results are reported in respective sections of this report:

- The Working Group on *Ethics and Big Data* (members: Bettina Berendt, Chris Biemann, Marco Büchler, Geoffrey Rockwell, Joachim Scharloth, Claire Warwick) discussed a very prominent topic with direct relationships to recent debates about ethical and privacy issues on the one hand and the hype about big data as raised by computer science on the other. One emphasis of the WG was on teaching how to process big data, how this research relates to legal and ethical issues, and how to keep on public dialogs in which such issues can be openly discussed – beyond the narrow focus of the academic community. A central orientation of this discussion was to prevent any delegation of such discussions to closed rounds of experts ('research ethics boards') which do not support open discussions to a degree seen to be indispensable by the WG. The widespread, fruitful and detail-rich discussion of the WG is reported in more detail in section 4.1.
- The Working Group on *Interdisciplinary Collaborations – How can computer scientists and humanists collaborate?*
  (members: Jana Diesner, Christiane Fellbaum, Anette Frank, Gerhard Heyer, Cathleen Kantner, Jonas Kuhn, Andrea Rapp, Szymon Rusinkiewicz, Susan Schreibman, Caroline Sporleder) dealt with opportunities and pitfalls of cooperations among computer scientists and humanities scholars. The WG elaborated a confusion matrix that contrasts commonplaces and challenges from the point of view of both (families of) disciplines. Ideally, scientists meet at the intersection which challenges both groups of scientists – thereby establishing CH potentially as a new discipline. In any event, this analysis also rules out approaches that reduce either side of this cooperation to the provision of services, whether in terms of computing services or in terms of data provisions. More information about the interesting results of this working group are found in section 4.2.
- The Working Group *Beyond Text* (members: Siegfried Handschuh, Kai-Uwe Kühnberger, Andy Lücking, Maximilian Schich, Ute Schmid, Wolfgang Stille, Manfred Thaller) shed light on approaches that go beyond language in that they primarily deal with non-linguistic information objects as exemplified by artworks or even by everyday gestures. A guiding question of this WG concerned the existence of content-related features of such information objects that can be explored by computational methods. As a matter of fact, corpus building by example of such artifacts is in many cases still out of reach so that computation can hardly access these objects. Seemingly, any success in 'computerizing' research methodologies here hinges largely upon human interpretation. Obviously, this is a predestined field of application of human computation with the power of integrating still rather separated disciplines (e.g., musicology, history of art, linguistics etc.). See section 4.3 for more information about this promising development.
- The Working Group on *Literature, Lexicon, Diachrony* (members: Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, David Smith) dealt with the role of information as stored in large-scale lexicons for any process of automatic text processing with a special focus on historical texts. To this end, the WG started from the role of lexica in preprocessing, the indispensability of accounting for time-related variation in modeling lexical knowledge, the necessity to also include syntactic information, and the field of application of automatic text analysis. Special emphasis was on error detection, correction

and propagation. The WG has been concerned, for example, with estimating the impact of lemmatization errors on subsequent procedures such as topic modeling. In support of computational historical linguistics, the WG made several proposals on how to extend lexica (by morphological and syntactical knowledge) and how to link these resources with procedures of automatic text processing. See section 4.4 for more information about the results of this WG.

Part and parcel of the work of these WGs were the plenary sessions in which they had to present their intermediary results in order to start and foster discussions. To this end, the whole seminar came together – enabling inter-group discussions and possibly motivating the change of group membership. Beyond the working groups, the work of the seminar relied on several plenary talks which partly resulted in separate position papers as published in this report:

- In his talk on *Digital and computational humanities*, Gerhard Heyer shed light on the role of computer science in text analysis thereby stressing the notion of exploring knowledge or text mining. He further showed how these methods give access to completely new research questions in order to distinguish between (more resource-related) *Digital Humanities* and (algorithmic) *Computational Humanities*.
- In his talk, Chris Biemann tackled the field of *Machine Learning* methods from the point of view of their application to humanities data. He clarified the boundedness of these methods in terms of what is called understanding in the humanities. From this point of view, he pleaded for a kind of methodological awareness that allows for applying these methods by clearly reflecting their limitations.
- In their talk on *On Covering the Gap between Computation and Humanities*, Alexander Mehler & Andy Lücking distinguished differences that put apart both disciplines. This includes a methodological, a semiotic and an epistemic gap that together result via an interpretation gap into a data gap. In order overcome these differences, they pleaded for developing what they call hermeneutic technologies.
- In her talk on *Digital Humanities & Digital Scholarly Editions*, Susan Schreibman gave an overview of her work on multimodal, multicodal digital editions that integrate historical, biographical and geographical data. Her talk gave an example of how to pave the way for a people's history in the digital age. To this end, she integrates recent achievements in data mining (most notably network analysis, geospatial modeling, topic modeling and sentiment analysis).
- In his talk on *How can Computer Science and Musicology benefit from each other?*, Meinhard Müller switched the topic of mainly textual artifacts to musical pieces and, thus, to musical artworks. He explained the current possibilities of automatic analysis of musical pieces and demonstrated this by a range of well-known examples of classical music.

This work nicely shows that computational humanities has the goal of covering all kinds of data as currently analyzed and interpreted in the humanities (see also the Working Group *Beyond Text* for such a view).

The seminar additionally included a range of short talks in which participants presented state-of-the-art results of their research: among others, this included talks by Christopher Brown, Anette Frank, Brian Joseph and Szymon Rusinkiewicz. This work nicely provided information about a range of linguistic and multimodal application areas and, therefore, reflected the rich nature and heterogeneity of research objects in the humanities.

A highlight of the seminar was a plenary discussion introduced by two talks given by Gregory Crane and by Manfred Thaller. These talks started and motivated an academic verbal dispute in which, finally, the whole seminar participated in order to outline future challenges of Digital Humanities with impact beyond the border of these disciplines – even onto the society as a whole. Both talks – on *Evolving Computation, New Research Directions and Citizen Science for Ancient Greek and the Humanities* by Gregory Crane (see section 5.1) and on *The Humanities are about research, first and foremost; their interaction with Computer Science should be too* by Manfred Thaller (see section 5.2) – opened a broad discussion about the role of humanities among the sciences and their status within the society.

Last, but not least, we should mention two common sessions with a concurrent seminar on Paleography. These sessions, which took place at the beginning and at the end of the seminars, opened an interesting perspective on one particular field that could be counted as a sub-discipline of Computational Humanities. The paleographers met in Dagstuhl for the second time and discussed some of our CH issues previously; it was fruitful to exchange approaches on how to overcome them.

## 1.4 Conclusion

Most of the working groups used their cooperation as a starting point for preparing full papers in which the theme of the group is handled more thoroughly. To this end, the plenary discussed several publication projects including special issues of well-known journals in the field of digital humanities. A further topic concerned follow-up Dagstuhl seminars. The ongoing discussions around the perceived gap between computer science and the humanities and the various proposals from the participants on how to define, bridge or deny this gap made it clear that the seminar addressed a topic that needed discussion and still needs discussion. The talks, panels and working group discussions greatly helped in creating a better mutual understanding and rectifying mutual expectations.

In a nutshell: the participants agreed upon the need to continue the discussion since CH is a young and open discipline.

## 2 Table of Contents

# 3 Overview of Talks

## 3.1 Digital and computational humanities

*Gerhard Heyer (Universität Leipzig, DE)*


**Joint work of** Gerhard Heyer, Volker Boehlke

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 favoured 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.

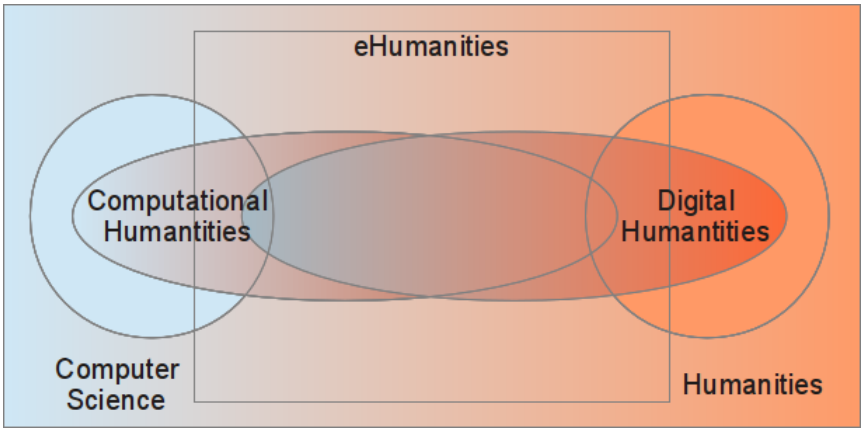


**Figure 1** Positioning of Computational and Digital Humanities in the context of Computer Science and Humanites

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.

## 3.2 Design Principles for Transparent Software in Computational Humanities

*Chris Biemann (TU Darmstadt, DE)*



**Abstract** In this short statement, the importance of transparent software for humanities research is highlighted. Here, three dimensions of transparency are identified: First, software should be freely available so that results are reproducible. Second, software should be easy to use and hide complex underlying algorithmics from the user. Third, to avoid a black box situation where the software's decisions are opaque to the user, the reasons for any of the automatically produced statements should be traceable back to the data they originated from. After elaborating on these principles in more detail, they are exemplified with a basic distant reading application.

**Introduction** The newly emerging field of Computational Humanities (CH) is situated at the interface between humanities research and computer science. Research questions in CH are concerned with aspects of both fields: in Digital Humanities (DH) research, computational aspects either not considered relevant or are merely assigned a subordinated role, while

in computer science, research on computational methods and algorithmic approaches is rather detached from their application domain – e.g. the field of Machine Learning produces methods that learn from data, no matter what kind of data it is. In contrast to this, CH considers humanist's questions and computational challenges both as first-class citizens, and focuses on their interplay. Whereas in both Computational and Digital Humanities, software solutions are needed that support the humanist – typically in accessing electronically available data in her respective field of study – CH research is also concerned with further automatizing the analysis using novel algorithmic approaches. As opposed to generic computer science approaches, however, algorithms in CH software are additionally required to be comprehensible by human(ist)s, in order to be open for scrutiny to allow for a depth of analysis that is satisfactory for the humanities. With respect to these prerequisites, a number of requirements on the software can be deduced. These will be subject of the following section, which discusses three dimensions of transparency that CH software should have in order to be a suitable tool for CH research. On a related topic, but written from the perspective of Computational Linguistics, see Pedersen (2008).

**Transparency of Software for Computational Humanities** The term 'transparency' can be defined in organizational contexts as 'the perceived quality of intentionally shared information from a sender' (Schnackenberg & Tomlinson, 2014) and implies openness, communication and accountability. In this section, these facets of transparency are elaborated on and put forward as desired properties of software used in Computational Humanities research.

**Open Source for Reproducibility** Whether hypotheses are merely empirically verified on data that has been mined by computational approaches, or hypotheses are generated from empirical observations in the first place: research in CH inherently includes empirical aspects, and rational deduction is complemented by a certain amount of experimentation. As in the experimental sciences, such as e.g. Physics, empirical investigations in CH must be reproducible to adhere to scientific standards. Just as it is considered bad science in the field of computational linguistics to rely on commercial search engines for data acquisition and statistics (Kilgarriff, 2006) because their inner workings are secret and they change over time, the CH researcher should not rely on commercial software with closed sources for the same reason. Rather, software in CH and other research contexts should be available open source in versioned public repositories, and the version of the software should be included in the description of the experimental setup. In this way, subsequent research is able to reproduce prior experiments of others and the inner workings of the software are fully transparent, at least for those that can understand computer programs. A further advantage of open source software over proprietary software, especially when distributed under a lenient license, is the possibility for subsequent research to combine several existing software into more advanced and more complex software without having to re-implement already existing methods.

**Intuitive Interfaces and Hiding Complexity** Just as in communication between humans, communication, i.e. human-computer interaction, happens when a CH researcher uses CH software. And just as successful fact-oriented communication between humans just provides enough detail to communicate the intended amount of information, supportive software should be intuitive to operate and hide unnecessary complex aspects from the user. For this, design principles of graphical user interfaces should be adhered to, and e.g. developed according to the visual analytics process (Keim et al., 2010). Abstracting from complexity, however should not be confused with obfuscation – while it is necessary for the acceptance of the software and its methods that algorithmic results are easy to obtain without necessarily understanding the algorithmic details, it is still crucial that the implementation

of such details are transparent (cf. section 2.1) and the algorithmic decisions are backed up by access to the data that leads to these decisions (cf. section 2.3). Only in this way, the CH researcher can build trust in her algorithmic methodology and develop an intuition about its utility and potential. A result of a successful CH research is always twofold: an algorithmic method and/or a mode of its application that allows to easily analyze data from the humanities, and a result in humanities research obtained with the help of such method.

**Accountability and Provenance** The most precise automatic result will still be subject to doubts and disbelief by human experts, as long as no explanation is provided how the automatic method arrived at such result. As mentioned in the previous section, in order for a method to be trusted, it needs to provide the possibility to drill down into the details of its decision-making process, to be fully accountable and to provide a fully transparent reason why the method arrived at a particular result, which is in software development known as data provenance (cf. W3C.org, 2005; Simmhan et al., 2005). In the context of CH, data provenance means not only to store and use algorithmic derivations of the input data (such as e.g. the number of times a certain term appears in texts of a certain time span), but also the sources from which these derivations were derived from (i.e., pointers to the positions in the documents where the term appeared) and a way to access them via the user interface. Data provenance enables the researcher to judge the software's decisions and to accept or discard algorithmically found evidence.

**CoocViewer – a Distant Reading Tool** In this section, we discuss CoocViewer (Rauscher et al., 2013), a simple tool for distant reading, along the three facets of transparency as outlined above. CoocViewer is an Open Source tool that allows browsing of statistically extracted networks of terms (cf. Quasthoff et al., 2006) extracted from corpora in the format of significant concordances. The figure below shows significant concordances for the term 'bread'. The complexity of the computation of such concordances and details of the concordance are abstracted; the user only notices the most significantly co-occurring terms, for example 'butter' located two positions to the right of 'bread'. To investigate this connection, the user can click on the link and drill down into all 20 references that lead to the link, as shown on the right side of the figure: CoocViewer provides full data provenance by showing – on demand – detailed information about single word frequencies and the references, including document titles and page numbers.

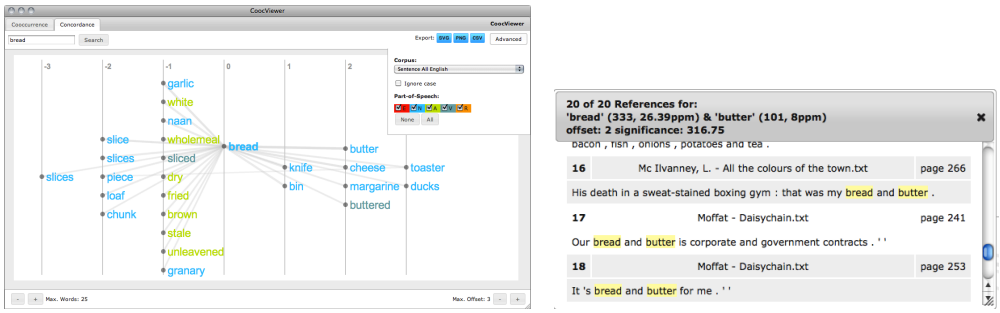


**Figure 2** CoocViewer software, showing significant concordances for "bread" and source text information for "bread and butter"

While not being a very complex example, CoocViewer adheres to the three design principles of transparency for CH software. Additionally, it enables the import and export of data in various formats for improved usability. During its development, several measures of significance, which determine the related terms shown as most significant concordances,

have been examined to investigate computational aspects of distant reading. The tool was productively used in quantitative literary analysis of crime novels, see (Rauscher, 2014).

**Conclusion** This short statement laid out design principles for the transparency of software for the computational humanities. Three important facets of transparency were identified that are desirable for software in the field of Computational Humanities: open source codebases for reproducibility, intuitive interfaces for effective communication between user and software, and data provenance for accountability and to build trust in algorithmic methods. These facets were exemplified on CoocViewer, a distant reading tool that adheres to these principles. Creating software to answer research questions in humanities research and computational research alike is one of the main aspects of the field of Computational Humanities. Adhering to the design principles of transparency, as discussed in this statement, enables a firm basis for reproducible research, the exchange of techniques and components, and the credibility of results through data provenance. Thus, not only the source data should be available freely to other researchers, but also the software that allows us to produce scientific results in the field of computational humanities.

### References


1 Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (Eds.). *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics Association, 2010
2 Kilgarriff, A. *Googleology is Bad Science*. Computational Linguistics 33(1): 147-151. 2006
3 Quasthoff, U., Richter, M. and Biemann, C. *Corpus Portal for Search in Monolingual Corpora*. Proceedings of LREC-06, Genoa, Italy, 2006
4 Rauscher, J., Swiezinski, L., Riedl, M., Biemann, C. *Exploring Cities in Crime: Significant Concordance and Co-occurrence in Quantitative Literary Analysis*. Proceedings of the Computational Linguistics for Literature Workshop at NAACL-HLT 2013, Atlanta, GA, USA. 2013
5 Schnackenberg, A., Tomlinson, E. *Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships*. Journal of Management DOI: 10.1177/0149206314525202. 2014
6 Simmhan, Y.L., Plale, B., Gannon, D. *A Survey of Data Provenance in e-Science*. ACM SIGMOD Record, 34(3):31-36, doi.acm.org/10.1145/1084805.1084812. W3C (2005): http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance


## 3.3 On Covering the Gap between Computation and Humanities

*Alexander Mehler, Andy Lücking (Goethe-Universität Frankfurt am Main, DE)*



Since digital or computational humanities (CH) has started its triumph in the humanities' research landscape, it is advisable to have a closer look at its methodological and epistemological range. To this end, we look at CH from the point of view of preprocessing, machine learning, and the general philosophy of science and experimental methodology. From this perspectives, a number of gaps between CH on the one hand and classical humanities on the other hand can be identified. These gaps open up when considering: (i) the status of preprocessing in CH, its logical work-flow and the evaluation of its results compared to the needs and terminological munition of the humanities. Most importantly, corpus preprocessing often comes *before* hypothesis formation and respective model selection has been carried out, turning the logically as well as methodologically required workflow upside down. (ii)

The predominant role of functional explanations in CH applications vs. the predominant role of intentional explanations with regard to the humanities. While so far computational processes can at most be functionally evaluated, hypotheses made in the humanities are usually embedded within contexts of justification that draw on some intentional statement. (iii) The possibilities of falsifying CH hypotheses and hypotheses in the humanities. Given the different typical patterns of explanations (see (ii) above), the results of computations and of the humanities cannot be put to falsification as known from the powerful methodology from the natural, experimental sciences. This leaves open questions about the validity of these results. (iv) The use of big data in CH vs. the use of deep data in the humanities. Analyses in the humanities usually involve the interpretation and rational reconstruction of their objects. This hermeneutic procedure goes beyond mere preprocessing and parsing of those objects, as is typically within reach of CH applications. When gathering interpreted and preprocessed data into corpora (which is done only seldom in the humanities, though), both approaches result in different kinds of resources which may be only of marginal benefit for the respectively other party. (vi) The lack of experimental methods in both CH and the humanities. In order to implement a notion of falsification in CH, one needs to think of CH-specific experimental settings which give rise to test procedures in the first place.

Based on these assessments, we argue that there are at least five interrelated gaps between computation and humanities, namely (1) an *epistemological gap* regarding the kind of evaluation mainly addressed by computational models in contrast to the kind of explanations addressed in the humanities; (2) a *data-related gap* regarding the build-up of ever growing text corpora in computer science in contrast to the need of controlled as well as deeply annotated data in the humanities; (3) a *semiotic gap* regarding signs as strings in the CH in contrast to rich sign-theoretical notions employed in the humanities; (4) a *methodological gap* with respect to understanding the functioning of methods of computer science by humanities scholars; and (5) an *interpretation gap* regarding the foundation of statistical findings in terms of the theoretical terms of the humanities involved. Having diagnosed these gaps we proceed by delineating two steps that could narrow (some of) these gaps: firstly, the understanding of CH technologies should be fostered by implementing them as part of a curriculum. Secondly, we should think of hybrid algorithmic methods, i.e. methods that at crucial branching points involve humanist expertise from the outset and in this way may pave the way towards "hermeneutic technologies" as a special kind of human-based evolutionary computing.

## 3.4 How can Computer Science and Musicology benefit from each other?

*Meinard Müller (Friedrich-Alexander-Universität Erlangen-Nürnberg, DE)*


**Joint work of** Müller, Meinard;
**Main reference** Meinard Müller, Michael Clausen, Verena Konz, Sebastian Ewert, Christian Fremerey. A Multimodal Way of Experiencing and Exploring Music. Interdisciplinary Science Reviews (ISR), 35(2): 138153, 2010.
**URL** http://www.audiolabs-erlangen.de/fau/professor/mueller/publications

Significant digitization efforts have resulted in large music collections, which comprise music-related documents of various types and formats including text, symbolic data, audio, image, and video. For example, in the case of an opera there typically exist digitized versions of the

libretto, different editions of the musical score, as well as a large number of performances given as audio and video recordings. In the field of music information retrieval (MIR) great efforts are directed towards the development of technologies that allow users to access and explore music in all its different facets. For example, during playback of some CD recording, a digital music player may present the corresponding musical score while highlighting the current playback position within the score. On demand, additional information about melodic and harmonic progression or rhythm and tempo is automatically presented to the listener. A suitable user interface displays the musical structure of the current piece of music and allows the user to directly jump to any key part within the recording without tedious fast-forwarding and rewinding. Furthermore, the listener is equipped with a Google-like search engine that enables him to explore the entire music collection in various ways: the user creates a query by specifying a certain note constellation, some harmonic progression, or rhythmic patterns, by whistling a melody, or simply by selecting a short passage from a CD recording; the system then provides the user with a ranked list of available music excerpts from the collection that are musically related to the query.

In the Dagstuhl seminar, I have provided an overview of a number of current research problems in the field of music information retrieval and indicated possible solutions. One goal within the Computational Humanities is to gain a better understanding to which extent computer-based methods may help music-lovers and researchers to better access and explore music in all its different facets thus enhancing human involvement with music and deepening music understanding. How may automated methods support the work of a musicologist beyond the development of tools for mere data digitization, restoration, management and access? Are data-driven approaches that can access large amounts of music data useful for musicological research? Vice versa, what can computer scientists learn from historical musicology? How can one improve existing techniques by incorporating knowledge from music experts? How do such expert-based approaches scale to other scenarios and unknown datasets?

## References


1 Meinard Müller, Michael Clausen, Verena Konz, Sebastian Ewert, Christian Fremerey. *A Multimodal Way of Experiencing and Exploring Music*. Interdisciplinary Science Reviews (ISR), 35(2): 138153, 2010.

2 David Damm, Christian Fremerey, Verena Thomas, Michael Clausen, Frank Kurth, Meinard Müller. *A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction*. International Journal on Digital Libraries: Special Issue on Music Digital Libraries, 12(2-3): 5371, 2012.

3 Meinard Müller, Thomas Prätzlich, Benjamin Bohl, Joachim Veit. *Freischütz Digital: A multimodal scenario for informed music processing*. In Proceedings of the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS), 2013.

4 Verena Konz, Meinard Müller, Rainer Kleinertz. *A Cross-Version Chord Labelling Approach for Exploring Harmonic Structures—A Case Study on Beethoven's Appassionata*. Journal of New Music Research: 117, 2013.

# 4 Working Groups

## 4.1 Report of Working Group on Ethics and Big Data

*Bettina Berendt (KU Leuven, BEL), Geoffrey Rockwell (University of Alberta, CAN)*


**Joint work of** Bettina Berendt, Chris Biemann, Marco Büchler, Geoffrey Rockwell, Joachim Scharloth, Claire Warwick

The following is the report of the Working Group on Ethics and Big Data (EBD) at the Dagstuhl seminar on Computer Science and Digital Humanities[1]. This working group was formed to discuss ethical and privacy issues around big data following the Snowden revelations. Some of the questions we asked included:

- What are the ethical and privacy issues raised by big data methods?
- What are our responsibilities as researchers and educators working around big data?

We came to the conclusion that, whatever position one might take on the ethics of big data, we have responsibility to expose our students to the lively discussion around the issue. This led to a more focused question:

- How can we teach the ethics of big data?

During the course of our deliberations we did the following:

- We experimented with a close reading of the CSEC slides.[2] The idea was to use slides leaked by Snowden to both a) explore EBD across disciplinary boundaries and b) to experiment with a way of teaching EBD through current materials. Such close reading of primary source documents about big data and surveillance can bring CH and DH folk together. We need the CH folk to read the software represented and the DH folk to read the documents as rhetorical documents. There is an interesting opportunity also for joint research at this intersection.
- We discussed the literature and archives that need to be explored in this area. (See the Appendix below for some of the archives identified). We agreed to share resources. Rockwell has, for example, create a preliminary reading list to be built on.[3]
- We agreed to share pedagogical materials. Berendt has shared her materials and other plan to as they experiment with teaching EBD.[4]
- We discussed the development of an infographic that makes the case for the importance of ethics in big data.
- We agreed to develop a web site with resources on this subject. Büchler has set up the basic infrastructure for this and we will begin to populate it as we experiment with teaching EBD.
- We agreed to write a short (5000 word) opinion piece for the "Discussions" column of KI - Künstliche Intelligenz (http://www.springer.com/computer/ai/journal/13218). We

[1] See http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=14301
[2] See http://www.scribd.com/fullscreen/188094600?access_key=key-2dvzkv8d3gnowt96adba&allow_share=true&view_mode=scroll
[3] See http://philosophi.ca/pmwiki.php/Main/BigDataEthicsReadings
[4] See http://people.cs.kuleuven.be/~bettina.berendt/teaching/2014-15-1stsemester/kaw/index1.htm – these materials are used in a course described here http://people.cs.kuleuven.be/~bettina.berendt/teaching/2014-15-1stsemester/kaw/

outlined an argument we were all comfortable with as a way of developing a common ethic. (See Appendix A: Discussion Outline).

**Conclusion**: It turns out that the issues are compelling and the reading of original documents like the CSEC slides to understand what the NSA (and others) are doing is one way into a shared discussion about EBD. Some of our conclusions were that:

- There are good ways to get people engaging with the issues – both the issues of ethics and the issues of what be really done get raised.
- We can imagine how we can turn forensic/diplomatic readings into problems for students and a site for interdisciplinary research.
- What the CSEC slides show is a process not unlike what we do ourselves (or want to do). This raises the issue of what the difference is between academic work and SIGINT (Signals Intelligence)? What makes one use of big data methods ethical or not?
- If the big data processes revealed by the Snowden leaks show good (or at least interesting) examples of big data interpretation (or analysis) then can we learn from them? Would it be ethical to copy the tools or processes revealed?
- Ultimately we have to ask how surveillance is different from research or forms of care for the other? Both are a form of knowing another – how is the other different and how is the knowing different?

**Appendix A: Discussion Outline**

| | |
|---|---|
| 1 | Introduction (framing it in terms of the current discussion) |
| 1.1 | How should we do big/data/science in light of the Snowden revelations? |
| 1.2 | Background to Snowden revelations |
| 2 | What do academics have to offer? What is the role of the researcher now? |
| 2.1 | We teach big data, |
| 2.2 | We are researchers developing new methods and tools, |
| 2.3 | We provide data, |
| 2.4 | We are citizens, and |
| 2.5 | We can act as mediators |
| 3 | The standard position on the ethics of big data is that it is not my business – that big data is just a tool/technique |
| 3.1 | How do researchers talk about their developments? |
| 3.2 | What can we learn from philosophy of technology? |
| 3.3 | Mining is about discriminating – one cannot avoid legal and ethical issues |
| 4 | We should beware inventing the other – we need to ask about our activities in the academy too. How is our research a form of surveillance. |
| 5 | Therefore we see the need for public dialogue rather than ethical decision trees that absolve people of the need to think about what they are doing |
| 5.1 | Most people think research ethics boards are the solution – but this delegates |
| 5.2 | We also need thick description – story telling |
| 5.3 | Finally, we need to teach people across humanities/data sciences |

**Appendix B: Archives and Literature**

- ACLU: https://www.aclu.org/nsa-documents-search
- Der Spiegel http://www.spiegel.de/international/the-germany-file-of-edward-snowden-documents-available-for-download-a-975917.html
- Nymrod: http://www.spiegel.de/media/media-34098.pdf
- Cryptome: http://www.cryptome.org
- LeakSource: http://leaksource.info/category/nsa-files/

## 4.2 Report of Working Group on Interdisciplinary Collaborations – How can computer scientists and humanists collaborate?

*Jana Diesner (University of Illinois at Urbana Champaign, US)*


**Joint work of** Jana Diesner, Chistiane Fellbaum, Anette Frank, Gerhard Heyer, Cathleen Kantner, Jonas Kuhn, Andrea Rapp, Szymon Rusinkiewicz, Susan Schreibman, Caroline Sporleder, Caroline

Our group explored obstacles, solutions and different types of benefits for the collaboration between humanities scholars and computing scholars. We formalized common pitfalls as well as opportunity spaces into a novel framework or model as shown in Figure 1. We are currently working on turning this framework and the outcome of our discussion into a paper.

**Table 1** Collaboration Scenarios

| | **Trivial or old Computing** | **Challenging or new Computing** |
|---|---|---|
| **Trivial or old Humanities** | **A)** Routine work; Might still be worthwhile pursuing, e.g. for digitizing data or generating data/ material for a project | **B1)** Humanities as a service, **B2)** Informed reuse of humanities knowledge, material, data (example from comp. linguistics: expertise for data annotation/ markup schemas) |
| **Challenging or new Humanities** | **C1)** Computing as a service, **C2)** Informed reuse of computing knowledge, data, skills (Example from comp. linguistics: building automatic annotation solutions) | **D)** Sweet spot of collaboration; Ideal situation: advancing science/ state of the art in both camps |

Ultimately, the ideal DH project will land in the lower right corner, where it advances science in the humanities and computing. The upper left corner bears no innovation for either domain, but might be a necessary precondition – e.g. data digitization work – for enabling some subsequent projects. The other two cells will entail innovation majorly for either the humanities or computing; with the other discipline serving as a utility or repository of data or methods. We believe that is essential for a successful project to identify where in this grid it belongs. Our planned paper can then help to identify common challenges and opportunities.

For each cell in this table, we have discussed preconditions, insights from both perspectives, pitfalls aligned with possible solutions and best practices. We believe that such an overview can help scholars to be more systematic and comprehensive in addressing the following problems:

- Identification and definition of the objectives and advantages for everyone involved in a CH/DH project? Possibly different ones for the computing versus the humanities people. A project might not advance both disciplines.
- Norms and standards, e.g. for publishing, co-publishing, performance, data gold standards

- Evaluation
- Barriers to collaboration: fears (traditional humanists: becoming obsolete, computing people: results of too low accuracy to get published, both: requirements associated with complexity of interdisciplinary projects including learning)
- Intellectual property (data, tools)

We have translated this model into implications for the actual process of starting and working through a collaborative project. We have concluded that prior to a project, extensive communication and discussions are needed to clarify on a couple of critical points. These points are outlined below and will be further detailed in our paper:

**Identification of which cell a project falls into**

- Is everybody ok with that?
- Collect arguments as to why either side could not carry out the work alone

**Funding**

- Hard for data collection, digitization (cell A)
- Easier for computing people, discuss role of humanities scholars in that case to ensure balanced responsibilities (cells C, D)

**Expectations**

- Computing "burns methods", Humanists "burn data". This requires a discussion on the value of methods and data, and expected standards for both.

**Success criteria for each camp**

- Methodological quantitative questions (focus in computing) versus substantive questions (focus in humanities)
- Data: amount, quality
- Performance: What matters? Speed? Accuracy? Theory? Understanding?
- Publishing

**Mutual learning**

- What amount of learning about knowledge and/ or skills from other camp is a) needed and b) expected?
- What learning resources are available? Add time for training into the grant application? Record training material as reusable resource?
- "Celebrate the gap"? Under what conditions does not each side necessarily need to be intimately familiar with the other

**Team composition**

- Mediators needed?
- Student research opportunities?

**Standards**

- Data collection
- Data analysis
- Evaluation
- Level of formalization, generalizability
- Publishing

We will bring the model shown in Table 1 together with these criteria for each cell in the matrix and align them with implications for the steps needed in every research process. Our team entails members from the humanities, computing and both, which we believe is essential for fleshing out these pitfalls and remedies.

## 4.3 Report of Working Group *Beyond Text*

*Andy Lücking (Johann Wolfgang Goethe University Frankfurt/Main, GER)*


**Joint work of** Siegfried Handschuh, Kai-Uwe Kühnberger, Andy Lücking, Maximilian Schich, Ute Schmid, Wolgang Stille, Manfred Thaller

The Working Group *Beyond Text* deals with any kind of media except text (i.e. written language). Accordingly, the group started by enumerating kinds of media as objects for digital humanities (DH). Due to the personal constitution of the group, the prime examples discussed are artworks (primarily paintings) and communicative everyday gestures. The example of paintings leads directly to a huge challenge for the feature-oriented focus of digital, corpus-based methods prevalent in DH: paintings exhibit properties bound up with their *expressiveness* that cannot straightforwardly reduced to (sets of) material features of the paintings – if they can be reduced at all. In particular, aesthetic judgments, for example, draw on normative backgrounds that are not part of the painting proper. As a consequence, such properties are out of reach for computational methods that only have (a digital representation of) the painting in question at their disposal. Such higher-order aspects of images, therefore, still rely on human interpretation, probably made explicit in annotation. Thus, respective work in DH seems to involve a hermeneutic dimension that so far is out of reach of computational automatization. This line of thinking, therefore, pinpoints a gap between humanities and DH and shed some light on a division of labor.

This result leads to curricular issues: what kinds of knowledge and which skills does a DH researcher need to have? Obviously, a genuine DH researcher optimally can decide which part of preprocessing or analysis can be done automatically and which part requires human interpretation. In order to make such a decision, the DH researcher needs to have a basic understanding of DH technology on the one hand, and of the hermeneutic methods in the humanities' discipline in question on the other hand. At this point, a self-evident connection to groups discussing curricular issues emerges.

A particular feature of paintings (though clearly not an exclusive one) is *vagueness*. Accordingly, the group discussed vagueness as a sample topic for DH dealing with media beyond text. Vagueness in paintings comes in a variety of manifestations: the colors of a painting give rise to a graduation known very well from categorization and prototype theory. The painting technique itself (e.g., sfumato) may result in a "visual vagueness" due to blurring the depicted scene and thereby preventing a clear recognition. Some features of the text may simply be unknown or uncertain like the name of the painter or the year of painting. Furthermore, paintings often draw on ambiguities of different kinds, ranging from flip-flop images over superimposed encodings to iconographic stylizations on top of figurative painting. A precondition for DH therefore is to distinguish different kinds of vagueness. According to the above-given list, at least the following phenomena have to be distinguished: *epistemic vagueness*, *visual vagueness*, *fuzziness*, *ambiguity*, and *interpretational vagueness*. Whether or not all these phenomena are subsumed under vagueness or a divergent terminological rendering is preferred, DH tools and techniques have to deal with them. This pertains to information storage (databases) as well as to computational modeling (e.g. fuzzy logic).

A special problem in this context is due to logical inconsistencies. Such inconsistencies can be the result of merged perspectives in paintings (think of the famous paintings of Escher) or of conflicting descriptions in texts (for instance, if the protagonist is sometimes described to be a left-hander, other times to be a right-hander). Problems of fictional speech acts and statements in fictional theory aside, a useful DH application has to provide even conflicting

information. Of course, contradictory details can simply be gathered in, say, a database. But this would come at a high prize: the application of inference engines would be blocked. The group discussed some application scenarios and possible technical solutions, though a realizable joint project had to be postponed to further collaboration.

It has to be emphasized that this summary is highly streamlined in the sense that it neither reflects nor exhausts the thematic and rhematic dynamics of discussions. Although only few talking threads converged into a viable proposal, the involvement of discussions shows that there is a great need for exchange of researchers from different backgrounds working in roughly the not yet delineated field of DH.

## 4.4 Report of Working Group on Literature, Lexicon, Diachrony

*Loretta Auvil (Illinois Informatics Institute, Urbana, IL, USA), David Bamman (Carnegie Mellon University, Pittsburgh, PA, USA), Christopher Brown (The Ohio State University, Columbus, OH, USA), Gregory Crane (University of Leipzig, DE, and Tufts University, Medford, MA, USA), Kurt Gärtner (University of Trier, DE), Fotis Jannidis (University of Wrzburg, DE), Brian Joseph (The Ohio State University, Columbus, OH, USA), Alexander Mehler (Goethe University Frankfurt, DE), David Mimno (Cornell University, Ithaca, NY, USA), David Smith (Northeastern University, Boston, MA, USA)*


**Joint work of** Loretta Auvil, David Bamman, Christopher Brown, Gregory Crane, Kurt Gärtner, Fotis Jannidis, Brian Joseph, Alexander Mehler, David Mimno, David Smith

### 4.4.1 Introduction

The Working Group on *Literature, Lexicon, Diachrony* identified three key issues or themes that pertain to the computational study of structured linguistic resources (prototypically, the lexicon) and unstructured text. These themes are the following:

- characterizing the nature of the information that has been captured in existing lexica written for human use and the possibilities for rendering these linguistic resources useful for automatic processing;
- exploring the possibilities of creating and augmenting linguistic resources by analyzing texts, and in particular in capturing diachronic variation; and
- analyzing, classifying, and mitigating errors introduced at each stage of processing, from optical character recognition and human annotation, to the construction of word frequency distributions and topic models, to part-of-speech (POS) tagging, lemmatization, parsing, and narrative analysis.

Schematically (as depicted in Table 2), these themes fit within a typology of complementary human and machine annotations. In what follows, we elaborate on each of these themes and develop within each various related sub-issues, some of which overlap with one another or serve as a bridge linking one theme with another.

### 4.4.2 The Nature of the Lexicon

The value of digitized lexica is well established: even elementary steps of text processing like OCR correction gain a great deal from access to lexica – not to speak of more challenging

**Table 2** Stages of lexicon formation contrasted with automatic automatic processing and human annotation.

| Stage | Human | Automated |
|---|---|---|
| Text creation | Double-keying | OCR |
| Combining variant forms | Morphology, lemmatization | String-edit clustering, morphological classification, named-entity recognition |
| Lexical disambiguation | Examples of textual citations, usage | PoS-tagging, contextual clustering |
| Sense disambiguation | Query expansion from existing definitions, organizing examples into categories | Latent semantic and topic analysis, contextual clustering |
| Relationships: phrases, synonyms, antonyms, frames, names | Examples of connections between documents | Collocate detection, parsing, lexical patterns (e.g. *not just X but Y*) |

tasks like textual entailment or discourse parsing. Our discussion began by asking what a dictionary is and what purpose it serves. More specifically, we asked whether it is a repository of information, an authoritative statement that users can turn to for answers, a snapshot of a language at a particular point in time, or just what (for a comprehensive international survey of lexica see Hausmann et al 1989).

For each stage of lexicon creation, there are both manual and automatic methods. We argue that modern workflows should incorporate both types of analysis. Table 2 shows correspondences between methods at each stage.

- **On the value of dictionaries:** There are various types of lexicon/dictionary serving different functions. For literary and linguistic research, lexica/dictionaries on historical principles are essential aids for the diachronic study of texts from the first records of a language up to its present-day varieties. Information technologies can contribute enormously to enhance the uses of existing dictionaries in various ways, thus satisfying the requirements of linguists and philologists studying texts (textual data), words and their histories. (Retro-)digitized lexica/dictionaries play a key role in transforming lexicographical resources from book form with alphabetic macro structures into more efficient means of locating reliable, accurate and comprehensive information; the user is no longer restricted to entries in alphabetical order, but can perform complex searches and exploit all the riches of information stored in a lexicon. The Perseus project[5] (see Crane 1996, also Lidell & Scott 1996) is one example of this.
  In the field of the vernacular languages, the scholar of *Middle High German* (MHG) in pre-electronic times had to use at least four dictionaries for this language period (ca. 1050 up to ca. 1350). These dictionaries have been digitized and all the essential information positions have been encoded carefully in order to allow complex searches related to lemma and word formation, word class, languages of loanwords, diachronic and diatopic features and document types of sources. The digitized dictionaries have been interlinked, so that an entry can be searched in all four lexica displayed synoptically on the screen (see:

[5] www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057

mwv.uni-trier.de). In the off-line version the search can be restricted to specific sources, e.g. the Arthurian novels, the writing of the mystics etc., or to a single text e.g. the *Parzival* by Wolfram von Eschenbach (see Fournier 2001). Furthermore, the existing MHG dictionaries are interlinked with the new MHG dictionary (*Mittelhochdeutsches Wörterbuch*) which is being published since 2006 in book form and concurrently on the Internet (www.mhdwb-online.de); for more information about the electronic text archive, lemmatization procedures etc. see Gärtner (2008). The interlinking techniques via normalized lemmata allow for the creation of dictionary nets for a certain period of a language. The period-related subnets can be interlinked with other historical dictionaries of a certain language, e.g. the *Deutsches Wörterbuch* (DWB, 2*nd* edition DWB$^2$) by the brothers Grimm (dwb.uni-trier.de). The interlinking can be achieved in various ways, e.g. via period-specific lemma forms in the head of an entry or even by semantic features (see: woerterbuchnetz.de). An even more global net of dictionaries could comprise dictionaries through more subnets e.g. for the Germanic languages: Gothic, Old English, Old Saxon, Old High German and Old Norse.

Interrelations of language stages, linguistic borrowings etc. can be studied in new and more reliable ways, if linguists and philologists are willing to look over the fences of their national languages and collaborate. Scholars of the classical languages with a long and interrelated history (Greek, Latin) have set an example and could play a leading role in this.

- **Extent to which morphological and syntactic information needs to be built into lexical representations:** when tagging texts, a first source of information about the parts of speech of tokens are lexica. A very obvious pitfall here concerns the distinction of the lexical *Part of Speech* (PoS) of a wordform – normally stored in the lexicon – and the syntactic PoS of a token of that form in a sentence. Obviously, these two assignments need to be distinguished. On the one hand, the PoS of a wordform stored in the lexicon can be used as a reference when tagging sentences in order to reduce the number of unknown tokens (obviously, this reduction supports any statistical tagging). On the other hand, the tagger may need to overwrite the lexicon information.

  Take the example of past participles, which in the lexicon are normally subsumed under a corresponding verb lemma: in German, for instance, participles can be used to derive adjectives (like in 'Der zerbrochene Krug'/'The Broken Jug') which have to be tagged appropriately. Thus, one has to balance the information taken from the lexicon against what has to be overwritten by the tagger. A way out of this problem is to tag both kinds of PoS: the lexical and the syntactic one. In any event, derivational knowledge (e.g., about the derivation of adjectives from participles) has to be included in the lexicon so that the search space of the tagger can be reduced. Given, for example, a verb like *lesen* ('to read') in German, a large set of nouns can be derived from it: *der Leser*, *der Lesende*, *das Gelesene*, *die Lesbarkeit*, *die Leserei*, *die Lesung* etc. Thus, one should not underestimate the additional amount of information to be stored in the lexicon if one has to consider, for example, a set of 20,000 verbs of a language. Derivational knowledge is morphological knowledge that is included here in the lexicon in order to guide the tagging of syntactic information in sentences.

  Note that the range of 'syntactically motivated' PoS can be much larger than what is distinguished lexically in the lexicon. Take the example of conjunctions where one can distinguish between subordinating conjunctions (subjunctions) and coordinating ones. Obviously, dependency parsing can be boosted by making this distinction during PoS-tagging. Thus, morphological or lexical ontologies of parts of speech can depart from

their syntactically motivated ones. Relying on some 'universal' PoS tagsets (Petrov et al 2012) does not solve this task. Once more, the reason is simply that if we want to reduce error rates of text parsing we need to include more and more information in the lexicon, that is, we need to make more distinctions, distinctions that are abstracted by universal tagsets or universal rule sets Marneffe et al (2014). In other words: while universal tag- or rule sets aim at the interoperability or comparability of methods, the humanities need in many cases rather fine-grained models that map the specifics of a given language or corpus and, thus, contrast with interoperability.

Another obvious example in favor of including syntactic information in the lexicon relates to the valency of verbs. Knowledge about this valency can guide dependency parsing and corresponding disambiguation processes (when distinguishing, for example, between complements of verbs and nouns). Once more, the amount of information to be considered here is enormous – it is even higher if we consider the requirement to account for variation of this information over time (see below). However, in order to meet the very low error rates acceptable to humanities scholars, there seems to be no alternative to more ambitious projects of building lexica.

To be more precise: any decision about what to include in the lexicon hinges upon the need to reduce error rates of tagging (historical) texts for humanities scholars. In this line of thinking, we always get a reason to extend the lexicon as much as possible: given the plethora of annotation desired by scholars (and not just from the point of view of NLP), most of the relevant information units still cannot be tagged automatically. Thus, it is desirable to put as much information as possible into the lexicon in order to make tagging less error prone. From this point of view, present-day full-form lexica (although usually including information about PoS and inflectional paradigms) are insufficient.

This approach may further interdisciplinary collaboration between computer scientists and scholars in the development of lexica and taggers based thereon. An example of such an interdisciplinary project is reported in (Mehler et al 2015) where historians work together with computational linguists in the exploration of Latin texts. The project deals with the genre- and register-related classification of medieval Latin texts based on their pre-processing in terms of lemmatization and PoS-tagging. To this end, the authors developed a large-scale Latin lexicon (primarily based on inflection patterns that produce around 11 million wordforms out of ca. 250,000 lemmata). The lexicon is used as a reference for PoS-tagging whenever it lists a single PoS for a form. Beyond that, tagging is done by means of a CRF that is superimposed by a set of short-scale 'syntactic' rules. In this sense, a hybrid approach is followed where lexical information is combined with a syntactic knowledge base and a statistical tagger. One should not underestimate the amount of work entailed by an approach in which the syntactic rules are handcrafted as are many of the patterns and even entries of the underlying lexicon. As it stands, such a labor-intensive approach (somehow reminiscent of human computation) is indispensable for text processing and, thus, for the generation of classification results acceptable to historians.

Several tasks undertaken by the Herodotos Project for Ethnohistory (Ohio State University/Ghent University) illustrate the necessary interplay of human correction with machine generation of data. These include: the determination of error rate and causes of error in the application of the Stanford Classifier to the identification of group names in the English texts of the Perseus corpus of ancient authors; the refinement of the classifier to deal with authors of different genres and periods; the development of Latin and Greek language classifiers suitable for identifying group names; automated XML markup of

the texts for which we have complete lists of (edited/corrected) group names, using the *TTLab Latin Tagger*[6] (TLT) for Latin and XML-/TEI-based tagging (Mehler et al 2015), and marking up Perseus code for group names by an automated process.

- **Relationship between dictionaries and chronology:** A key role for finding information about the history of a word and its usage is played by the dating of its sources in historical dictionaries. Changes of spelling and morphology of a lexical item, the first record of its use and meaning etc. are usually documented in the great national dictionaries (OED, DWB$^2$, TLF etc.). The definitions of a lemma connected to a certain language stage could also be looked up in a period specific lexicon. Of special interest in searching historical texts for definitions are borrowings, especially from Latin into German, English and other European vernaculars. The Latin borrowings e.g. of German from its first recordings in bilingual word lists in the 8$^{\text{th}}$ century through all the following periods up to the 19$^{\text{th}}$ century are immense. In religious texts of the Middle Ages as well as in scientific writings of today there is hardly any sentence without Latin traces. Latin loans in German books printed from about 1500 were marked for a long time by a change of fonts: Fractura had been used for German, antiqua for Latin.
  In lemmatizing historical texts the change of fonts is essential in order to filter out loans and find the appropriate time related definitions in the *Deutsches Fremdwörterbuch* (Schulz et al 1995), the sister of the DWB which from its inception was not meant to contain loanwords (*Fremdwörter*). The borrowings from Latin consist not only of loanwords, which are usually taken over together with a specific meaning (see the definitions in the national dictionaries to Medieval Latin), but also of loan translations (e.g. Latin *re-sur-rect-io* and its morpheme based rendering in German *Auf-er-steh-ung* which goes back to MHG *f-er-stand-unge*). Translating key Christian terms in the Middle Ages led to a variety of synonyms of which in the course of time often only one has survived (of six synonyms for Latin *gratia* with its specific Christian meaning in OHG only one made it into MHG *genäde*, NHG Gnade). For determining which concept is represented by which lemma and definition we need a semantic index to the historical dictionaries. This is a real challenge for digital humanists trying to explore the lexical history of an expression and its definitions through time and place. An inspiring example is the *Historical Thesaurus of the OED* by Christian Kay which has been integrated into the *OED online*.
  It is a commonplace that the meaning of a lexeme changes over time. However, it does not do so according to a single timescale. Thus, by analogy to Domingos (2008), we may speak of modeling the variation of lexical items in terms of *structured time* in order to account, for instance, for different processes of temporal variation (e.g., function words change according to a longer timescale than content words). This variation can be conditioned by the dynamics of genres and registers in which the lexical items are preferably used (Halliday 1977, Halliday 1991). In such cases, models of genres and registers are additionally required. Thus, beyond morphological and syntactical knowledge we may also include pragmatic knowledge in the lexicon. Time is just a gateway for this kind of knowledge.
  Thus, a central challenge of automatic, lexicon-based text analysis of historical texts concerns the requirement to cover time as a constitutive parameter of lexicon formation. That is, the variation of the morpho-syntactic realizations of lexical items over time have to be considered as an integral part of the lexeme/syntactic word/wordform relation. So far, little is done in this respect: either the lexica do not contain information about lexical

[6] See prepro.hucompute.org and collex.hucompute.org

variation (applying, for example, lexica of classical Latin to medieval Latin texts) or the taggers do not operate in a time-sensitive manner. In order to understand possible pitfalls of the latter case consider the task of tagging multilingual texts: taggers are typically language-specific; if an input text of language $A$ contains text spans (e.g. citations) of another language, the tagger tries to tag these spans as instances of language $A$ – obviously, this is an erroneous procedure. Rather, what should happen here is that the tagger starts with language detection for any text span in order to select the corresponding language-specific tagger for it. The same should happen along the time axis where time period-sensitive taggers are selected to tag corpora of historical texts that instantiate several stages in the development of one or more languages. As it stands, current taggers are not powerful enough to account for such requirements of stratified tagging – stratified with respect to time, language, register, genre etc.

- **Linking lexica via *hyperlemmata*:** above, we argued that rather than abstract tagsets, fine-grained lexicon models are needed to meet the requirements of, say, philologists, who look for the specifics of certain texts rather than for a generalized model, say, of the PoS realized by them. Such an approach runs the risk of adapting its lexicon model to the specifics of the underlying corpus in such a way that interoperability of methods and comparability of findings is negatively affected. In order to provide a way out of this fallacy, we may think of using *hyperlemmata* to establish links between the lemmata of different lexica. This model is in line with approaches like Petrov et al (2012) and Marneffe et al (2014), but with a focus on lexemes instead of PoS or dependency rules. Given a unified lexicon model based on hyperlemmata one can envision 'translations' between different lemmatizations of the same text. Alternatively, one can envision abstract search queries based on hyperlemmata that are automatically mapped onto the specifics of the underlying lexica. Such an additional layer of modeling lexica entails a further level of labor-intensive research. However there seems to be no alternative to such an approach if our goal is to switch between different lexical ontologies.
- **Compiling lexica automatically (definition generation):** since processing historical languages is reminiscent of processing low-resourced languages in that it faces related challenges, it is necessary to think of standardized procedures for the rapid, less error prone compilation of lexica even out of (small) corpora of historical texts. Here, we envision a combination of methods of (i) computational linguistics for learning, for example, inflection patterns, valency patterns or word-order patterns, (ii) text-technological methods of building and maintaining lexical databases and (iii) methods of human computation for the fine-grained adaptation and extension of the resulting lexica. On the basis of such a procedure, one can envision an application that allows for estimating the complexity of building a lexicon for a given historical language starting from a given corpus of a certain size. Such an application could help interdisciplinary projects distribute the various tasks of compiling the lexicon among project members.

### 4.4.3 Computational Analysis of Literary Texts

In addition to the structured information in human- and machine-readable lexica, computational linguists and digital humanists work with increasingly large bodies of unstructured text. To speak very broadly, this text varies greatly in the specificity of its metadata, the consistency of its editing, and the standards and accuracy of its transcription. On the one hand, creators of lexica and other linguistic resources have always used corpora to investigate and illustrate linguistic facts, and textual critics have always been concerned with the basis of our knowledge of texts. On the other hand, the wide availability of electronic texts and

means for their automatic analysis encourage us to think more systematically about the interplay of lexicon and corpus.

We believe, therefore, that important research questions will continue to center around our ability to augment structured resources such as lexica with inferences from unstructured text and how to exploit lexica to improve automatic processing. Among the specific problems we discussed were:

- practical problems in compiling corpora to work with, in particular for long-term diachronic analysis including multiple language stages, typefaces (e.g., Fractura), and genres;
- constructing corpus-specific lexica and refining existing lexica with corpus data;
- adapting standard NLP tools to domains (e.g., literary texts) that may be divergent from the newspaper texts on which they were trained;
- interpreting automatic clustering methods such as topic modeling extraction from texts: the intersection of computational analysis of text and the lexicon, since here word-meanings make a difference;
- automated thematic analysis;
- automated plot summaries; and
- computer-aided stylistic analysis.

For example, one problem in applying topic models and related approaches to historical texts is that any semantic analysis should not only consider wordforms, but rather lexemes or – better – lexeme groups (*Lexemverbände* in German) which subsume lexemes based on the same stem even if they belong to different part-of-speech classes (an example is *fliegen*, *Flug*, *Flieger* etc.). In order to do this, a very good lemmatization is needed. As discussed above, this is a task that is not completely solved in the case of historical texts. Here, we still need to do a lot even in terms of lexicon building. However, presentations of clouds of wordforms subsumed under the same topics will hardly convince philologists or historians who – as outlined above – expect very low error rates. A wordform is a formal unit and not a semantic unit. Lexemes in the lexicon are dually articulated in the sense of de Saussure: formally by the wordforms by which they are realized and semantically by the meaning that they carry. If topic modeling aims at drawing level with this view, it should be combined with a very thorough pre-processing of historical texts – beyond what is currently done in many approaches to topic modeling. Here, historians, philologists and computational linguists should go hand in hand in order to further develop their methods – possibly by example of topic models that are well established on the ground of taggers as described in section 4.4.2. This can be a way out of detecting, for example, function words as part of word clouds attributed to a certain topic, where these function words do not occur in the cloud because of carrying a certain meaning, but just due to the statistics of the given text.

What kind of structure of the lexicon would enable a better analysis of literary texts? Strategies to improve text analysis, which informations of a digital lexicon can be employed (for example hypernymy/hyponymy)?

In order to better meet the requirements of text analysis with the help of lexica, the following information objects should be included into lexicon formation: the lexicon should cover derivation relations in order to allow for modeling lexeme groups (see above). It must consider time as an attribute of any relation and any attribution in the lexicon (e.g., Which lexeme is realized during which period by which wordform carrying which grammatical information? etc.). Beyond time, each lexicon entry should be equipped with an expressive attribute model that allows for mapping various syntactic, semantic, pragmatic, genre- or register-specific information units (e.g., sentiment/polarity, connotations, semantic classes (e.g., anthroponyms, oikonyms, chrononyms etc.)).

### 4.4.4 Error Detection and Correction

An important issue with any application of computational methods to text is the degree to which errors occur in the automatic processing. While scholars in some areas of computer applications may be satisfied with a small error rate (say 1%, for optical character recognition of documents printed within the last hundred years or so), humanists tend to be very concerned about the integrity of the text that they are working with, and tend to express great dissatisfaction with even tiny error rates smaller than 1%. Thus, it is a concern to be able to detect errors, to predict the rate at which they are likely to occur, to characterize their effects on subsequent processing, and to be able to do something about the errors if possible. Among the applications we considered that could generate errors (while at the same time, of course, generating electronic output that is very useful and usable) was optical character recognition to create electronically manipulable texts.

- classification of errors (OCR errors, lemmatization errors, POS-tagging errors, parsing errors etc.);
- consequences in statistical analysis;
- how error rates affect results;
- the extent to which errors are random or patterned; and
- understanding the impact of errors in functions and propagation of errors relationship of dictionaries, functions and errors.

A central challenge of any error analysis concerns the availability of online tools for comparative error analyses by which errors can be classified and displayed in terms of summaries (e.g., by decreasing frequency). This is needed to allow for a more rapid and comprehensive detection and elimination of errors. Current systems either only provide summary data (in the form of F-measure statistics) or only selected error analyses by discussing some use cases. However, what is needed is a systematic overview of the whole range of errors being made by automatic text analysis, an overview that human users can use to guide future processes of text analysis in order to guarantee lower error rates. To this end, computational linguists building annotation tools, digital humanists (providing web-based interfaces for the usage of these tools), and humanities scholars should cooperate much closer in order to meet the low-error-rate requirement of the humanities.

### 4.4.5 Conclusions and Future Research

Digital linguistic resources such as lexica are necessary for making progress in many areas of natural language processing; moreover, the availability of digitized corpora and automatic annotation methods can make creating these resources a collaborative effort between linguists, philologists, and computer scientists. We see several opportunities for strengthening these collaborations, for creating new linguistics resources, and for analyzing and mitigating the errors in human and computational annotation processes.

### 4.4.6 Acknowledgement

The group thanks Bettina Berendt for her fruitful hints, comments, and discussion of the topics discussed by this working group.

#### References

1 Gregory Crane. Building a digital library: the perseus project as a case study in the humanities. In *Proceedings of the first ACM international conference on Digital libraries*, DL '96, pages 3–10, New York, NY, USA, 1996. ACM.

2 Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, pages 4585–4592, 2014.

3 Pedro Domingos. Structured machine learning: Ten problems for the next ten years. *Machine Learning*, 73:3–23, 2008.

4 Johannes Fournier. New directions in middle high german lexicography: Dictionaries interlinked electronically. *Literary and Linguistic Computing*, 16:99–111, 2001.

5 Kurt Gärtner. The new middle high german dictionary and its predecessors as an interlinked compound of lexicographical resources. In *Digital Humanities 2008, Oulu, Finland, Book of Abstracts*, pages 122–124, 2008.

6 Michael A. K. Halliday. Text as semiotic choice in social context. In Teun A. van Dijk and J. S. Petöfi, editors, *Grammars and Descriptions*, pages 176–225. De Gruyter, Berlin/New York, 1977.

7 Michael A. K. Halliday. Towards probabilistic interpretations. In Eija Ventola, editor, *Functional and Systemic Linguistics*, pages 39–61. De Gruyter, Berlin/New York, 1991.

8 Hans Schulz et al. *Deutsches Fremdwörterbuch, begonnen v. Hans Schulz, fortgeführt v. Otto Basler, weitergeführt im Institut für deutsche Sprache, Bd. 1-2 [A - Pyramide], Straßburg 1913 und 1942, Bd. 3-7 [Q bearb. v. Otto Basler, P - T bearb. v. Alan Kirkness, U - Z bearb. v. Gabriele Hoppe; Bd. 7 Quellenverzeichnis, Wortregister, Nachwort hg. v. Alan Kirkness], Berlin 1977-1988. - Deutsches Fremdwörterbuch, 2. Auflage, völlig neunearb. im Institut für deutsche Sprache, Bd. 1ff. (bisher Bd. 1-7 [A-Präfix - hysterisch] bearb. v. Gerhard Strauß u.a.*, volume 1. de Gruyter, Berlin/New York, 1995ff.

9 Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, and Ladislav Zgusta, editors. *Wörterbücher. Ein Internationales Handbuch zur Lexikographie. 3 Teilbände.* Handbücher zur Sprach- und Kommunikationswissenschaft 5, 1; 5, 2; 5, 3. de Gruyter, Berlin / New York, 1989; 1990; 1991.

10 Henry George Liddell and Robert Scott. *A Greek-English lexicon: With a revised supplement.* Clarendon Press, Oxford, 1996.

11 Alexander Mehler, Tim vor der Brück, Rüdiger Gleim, and Tim Geelhaar. Towards a network model of the coreness of texts: An experiment in classifying Latin texts using the ttlab latin tagger. In Chris Biemann and Alexander Mehler, editors, *Text Mining: From Ontology Learning to Automated text Processing Applications*, Theory and Applications of Natural Language Processing. Springer, Berlin/New York, 2015. appears.

12 Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC 2012, pages 2089–2096, Istanbul, Turkey, 2012.

# 5 Panel Discussions

## 5.1 Evolving Computation, New Research Directions and Citizen Science for Ancient Greek and the Humanities

*Gregory R. Crane (Tufts University, US)*



**URL** http://sites.tufts.edu/perseusupdates/2014/09/29/opening-up-classics-and-the-humanities-computation-the-homer-multitext-project-and-citizen-science/

Increasingly powerful computational methods are important for humanists not simply because they make it possible to ask new research questions but especially because computation makes it both possible – and arguably essential – to transform the relationship between humanities research and society, opening up a range of possibilities for student contributions and citizen science: http://homermultitext.blogspot.de/, http://www.homermultitext.org/.

My departure point for this paper were questions during a workshop on Computational Humanities, at Schloss Dagstuhl in July 2014 that Manfred Thaller posed to me in response after a talk I delivered, and then as part of a podium discussion. I had argued that we needed to advance research projects that provided our students with an opportunity to make substantive contributions to research as a central part of their education. In so doing, I echoed Wilhelm von Humboldt, who argued that students in a university should always be engaged in advancing human understanding – mastering set curriculum was for primary and secondary school. For Humboldt, the challenge of new questions and interests from students was one of the great intellectual advantages of working in a university rather than in a research institute. Manfred reminded me that if we focused too much upon serving the students, then we would do a disservice to research.

The challenge is to establish a productive tension between the established interests of the faculty and the fresh questions of the students. I conclude this paper by describing the Homer Multitext Project and how, in opening up new opportunities for student contributions and research, this project shifted my idea of how the field should move forward. Research should not solely serve students (then it runs the risk of being dumbed down) but it should not only serve specialist researchers (then it runs the risk of becoming detached scholasticism). In the ideal case, we identify research that challenges (and captivates) the most advanced researchers but that also engages a broader audience and provides multiple opportunities for contribution. In the study of Greek and Latin, I see this productive tension leading to re-organization (and, in the US at least, a revival) of very traditional philological questions with very new methods and in a much more decentralized and collaborative culture.

This paper thus argues that the most important consequence of computation for the Humanities lies not in the new research questions that are now appearing but in the fact that research with increasingly large bodies of digitized source materials opens up opportunities, indeed the necessity, for a new, more open culture of intellectual production, one that is less hierarchical, more focused on collaborative inquiry, more dynamic, and, in my view, more effective as an environment for broad and deep intellectual development. Computation allows – perhaps more accurately, challenges – humanists to redefine their relationship to their students and to society as a whole.

A link to the full document can be found at http://tinyurl.com/kner5dk.

## 5.2 The Humanities are about research, first and foremost; their interaction with Computer Science should be too.

*Manfred Thaller (University of Cologne, Germany)*



This statement describes with the privilege of hindsight a view on a controversy about the relative weight to be assigned to various priorities which the application of computer science to Humanities research questions should support. The controversy was started by my response to a statement of Greg Crane about the importance of pursuing educational goals by computer supported systems in the Humanities. I argued, that the application of Computer Science to the Humanities at the university level must gain its intellectual merit primarily from the contribution it makes to the research agenda of the Humanities. The controversy about the relative importance of these two goals can be quite sharp, depending on the implications one of these two priorities has within a political argument for the way in which Computer Science and the Humanities are to cooperate in a given university context. Obviously, when such a political context does not exist, the two goals are not contradictory as such. Nevertheless, the following theses shall summarize and clarify my appeal for a strong focus on the research agenda in the Humanities to be supported by Computer Science.

**Assumption 1.1:** The interest of the Humanities in Computer Science has, from the very beginning, been directed by two goals, which can easily be mixed up with confusing results. Humanities research requires many routine tasks, which are plain drudgery: Busa's dream of never have to go through Aquinas line by line any more to look for the forms of the root of a specific word, is the obvious example. On the other hand, there has always been the promise that some methods supported by Computer Science might open up the way to explanations for Humanities phenomena, which open up new epistemic vistas for the disciplines producing these explanations.

**Assumption 1.2:** While the interest in both aspects of this interdisciplinary field is perennial, the relative emphasis assigned to them depends intellectually on the state of the epistemological discussion in the Humanities at large or their individual disciplines. Humanists who are impressed by C.P.Snow or K.R. Popper assign different priorities to interdisciplinarity, than those following P.K. Feyerabend.

**Assumption 2.1:** The Humanities consider themselves currently in a crisis, or rather, primarily the Anglo-American branch of the Humanities does. In those countries they react by an intensive discussion on how low-level applications of technology can make the disciplines look more modern and relevant. As this is a defensive movement, it frequently reacts frightened against applications which require a more thorough understanding of Computer Science, which is seen as additional competition, which is more likely to sharpen the crisis of the Humanities, rather than alleviate them.

**Assumption 2.2:** This I consider a dangerous fallacy. The Humanities do not become more respectable by showing that they employ the same simple tools which everybody else uses in the year 2014. Nor does the ability to teach critical thinking at the level of the gymnasium justify a research agenda.

**Assumption 2.3:** The Humanities need a broad vision, why they are so fascinating, that society at large should support them. This is not done by inventing short term economic benefits, but by presenting goals which have such a wide appeal, that society is willing to support them, even if no short term benefits are generated. Hubble does not lower unemployment rates; it promises to unveil fascinating secrets.

**Assumption 3.1:** Computer Science for a long time has been a discipline which emphasized the necessity of data being highly structured and free of contradictions as a precondition for their processing, even if in some rather exotic branches a theoretical interest in applications vulnerating these preconditions has always been existing. As it has changed from a discipline supporting individual solutions for specific problems into the conceptual and theoretical backbone of an integrated infosphere, Computer Science can less and less define what properties the data should have, it intends to process and has therefore to handle whatsoever comes along.

**Assumption 3.2:** The Humanities have since their earliest inception always been focusing on the ability to draw a maximum of conclusions from a rather limited amount of information, they could access physically. The only start to notice, that this barrier has broken down. The primary qualification of a Humanities' researcher of the year 2050 will not be, how to lovingly extract insights from a few isolated bits of information, but how to meaningfully integrate the information contained in the largest possible set of data.

**Assumption 3.3:** There is a convergence, therefore, between the approach towards data to be supported by Computer Science and the Humanities.

**Thesis 1:** To reach the vision postulated as necessary by assumption 2.3, the Humanities have to focus more strongly again at the epistemic implications of methods which can be supported algorithmically (cf. assumption 1.1 / 1.2), as only so the challenge posed by assumption 3.2 can be answered successfully.

**Thesis 2:** Care has to be taken, that the high visibility of low level approaches to "Digital Humanities" described in assumptions 2.1 and 2.2 do not obscure the developments needed for the support of thesis 1.

**Thesis 3:** To support thesis 1, a joined research agenda between the Humanities and Computer Science is necessary, which does not restrict itself to the application of known algorithmic solutions on the knowledge domain if the Humanities, but uses the challenges described in assumption 3.2 to help solving the challenges described in assumption 3.1. Realizing, that is, what has been postulated by assumption 3.3.