# Report on eHumanities research topics relevant in the Computer Science (R 4.3.1)

**Version** 1.0
**Work Package** WP 6
**Responsible Partner** University of Applied Sciences Worms

## TextGrid

Networked Research Environment for the Humanities

SPONSORED BY THE
Federal Ministry of Education and Research

**Project:** TextGrid - Networked Research Environment for the Humanities
Funded by the German Federal Ministry of Education and Research (BMBF) by Agreement 01UG0901A
**Project term:** June 2009 to May 2012

**Document Status:** <final version 1>
**Distribution:** <public >
**Authors:**
Marc Wilhelm Küster (UAS Worms)
Thomas Selig (UAS Worms)
Julianne Nyhan (University of Trier / University College London)
Wolfgang Pempe (SUB Göttingen)
Kathleen Smith (University of Illinois / SUB Göttingen)

**Revision History:**

| Date | Author | Comments |
|---|---|---|
| 2010/09/30 | | Version 1 released |

# Table of Contents:


**1. Introduction ..... 4**
**2. eHumanities and eResearch: Paradigms ..... 4**
2.1. eScience and eResearch ..... 7
2.2. eResearch ..... 7
2.3. Data-driven science ..... 8
2.4. Hermeneutic Informatics ..... 9
2.5. eResearch methods in the Humanities ..... 9
**3. Relevant Research Challenges in the Intersection of eHumanities and CS 10**
3.1. eResearch and Data-driven Science ..... 10
3.1.1. eResearch ..... 10
3.1.2. Data-driven Science ..... 10
3.2. Data representation and data formats ..... 11
3.2.1. Data Preparation ..... 11
3.2.2. Data representation: XML, SGML and more ..... 11
3.2.3. Data visualization and publication ..... 12
3.3. Repositories and Long-term archiving ..... 13
3.4. Semantic Web ..... 13
3.5. Architecture ..... 14
3.5.1. Technical interoperability ..... 14
3.5.2. Semantic interoperability / resource registries ..... 14
3.5.3. Digital Ecosystems ..... 15
3.6. Horizontal interaction with other fields of application ..... 16
3.7. Software engineering ..... 16
3.7.1. Tool development ..... 16
3.7.2. Usability and User Involvement ..... 17
3.7.3. Tests and Test methods ..... 17
3.8. Incentive Systems ..... 18
**4. Summary ..... 19**
**5. Bibliography ..... 19**
5.1. Books and Overview Studies ..... 19
5.2. Web Sites ..... 20
5.3. Articles ..... 21
5.4. Reports and Lectures ..... 22

# 1. Introduction

The term eHumanities may be relatively new, but many of the concepts behind it are by IT standards old indeed. Roberto Busa's *Index Thomisticus* started in the late 1940s and was not only one of the first — if not the first — computer-driven project for large-scale text capture, storage and analysis world-wide, it may well also be the oldest IT-based project still up and running today, albeit in a very different incarnation from the punched cards it was born on.

The predecessors of what is now known as eHumanities have at various times and places been known as Computing in the Humanities; Humanities Computing; Literary and Documentary Data Processing; and Literary and Linguistic Computing etc. (sometimes with an implicit restriction of scope to textual documents or with a focus on specific approaches). eHumanities addresses all aspects of the application of computing technology to the Humanities but has a particular, distinguishing focus on computationally intensive inter- and multidisciplinary collaborative work, often carried out by geographically distributed teams and the infrastructure, methods and theories that drive such research. The term Digital Humanities does not necessarily have this dimension and so in this report we propose that eHumanities can be viewed as a school within the larger DH movements.[1]

It is in this framework that eHumanities and hence also the related research in information technology is positioned. This report takes a wide view of the term eHumanities and defines it as a "field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities" [W-1] with a specific focus on collaborative working. Unlike most studies and in particularly McCarty's brilliant [B-1] which look at humanities computing more from a *humanities* computing point of view, this report wants to establish a research agenda for humanities *computing,* i.e. from the point of view of a computer scientist interested in research questions in this intersection of his discipline with the humanities.

# 2. eHumanities and eResearch: Paradigms

In a second century b.c. play written by Terence we find the exclamation *"homo sum, humani nil a me alienum puto*" (I am a human being. I consider nothing human alien to me). This quote is a valuable one to reflect on when trying to sketch the boundaries and concerns of the Humanities to scientists from other disciplines who, in the process of carrying out interdisciplinary research, find themselves on such foreign shores.

The Humanities, then, concerns itself with almost every aspect of what it means to be a human being, whether in the past or present as well as how such interpretations may influence the future. The 2007 Position Paper [R-1] published by the European Science Foundation's Standing Committee for the Humanities (ESF SCH) emphasises the centrality of Culture to the Humanities, and describes the former as 'the result of the complex of conceptual, linguistic, affective, moral and behavioural systems that allow us to define and re-define ourselves in a changing world. This "cultural complex" is what the humanities study.'[R-1: 5]. Accordingly, the Humanities comprises a wide range of disciplines: the ESF SCH mention history and archaeology, literary studies, art history, musicology, psychology, anthropology, philosophy and logic, linguistics, pedagogical and educational research and the history and philosophy of science. However, they also reflect that 'In thus dealing with culture, its contents and its manifestations, its structures and its constraints, the humanities naturally interact with other fields of science and art' [R-1: 5]. It is just such an interface that this report seeks to address, more specifically that of eHumanities. Inherently

[1] Some researchers e.g. in the so-called "Manifesto for the Digital Humanities" want to include the social sciences in the scope of DH [W-23]. However, this view seems to have little currency in literature and amongst practitioners.

interdisciplinary, this emergent field draws computing technology and potentially any item of the Humanities 'data-set' i.e. any knowledge-bearing cultural heritage object into a sustained 'dialogue' with one another. The ultimate aim of such research is to enable otherwise impossible questions to be discovered and asked, and to enable new insights into the role of technology in human culture, communication and collaboration to be posed too. As William Turkell has written of the insights that eHumanities is generating into the Humanities itself "Just because the separation between thinking and making is long-standing and well-entrenched doesn't make it a good idea. At various times in the past, humanists have been deeply involved in making stuff: Archimedes, the Banu Musa brothers, da Vinci, Vaucanson, the Lunar Men, Bauhaus, W. Grey Walter, Gordon Mumma. The list could easily be multiplied into every time and place …" (W-29, par. 5).

In the Humanities, artifacts of, inter alia, human expression, interaction and imagination form the key 'data-sets'. In contrast with the Sciences, where data sets tend to be, in the first instance, generated rather than collected and homogenous in nature e.g. numeric, in the Humanities, data tends to be collected and highly heterogenous in content and format. Examples include images (ranging from the cave paintings of Lascaux over Quattro Centro perspectival art to modern day digital art), all aspects of oral culture (whether international folktale motifs or reconstructions of proto-Indo-European poetry) music and musicology (whether printed tablature, recordings on a phonautograph or reconstructed antique instruments ) and the written word (whether inscribed on the clay tablets of Sumer, such as our proto-dictionaries, recorded on parchment with quill and ink, the incunabular editions of the early printing press or modern day electronic texts). Considering the Humanities from the pan-European perspective further crucial traits emerge too. In contrast with Computer Science Conference proceedings are not the main publication format of the Humanities, nor are Journals, rather scholars publish in a multiplicity of formats (monograph, edited collection, journal, encyclopaedia, dictionary entry, website etc). Furthermore, citation and citation life patterns differ substantially in the Sciences and Humanities. Considering the Humanities from the pan-European level, English does not hold sway as the lingua franca, whether as a language of publication or a language of study.

One the one hand it is not possible to transfer completely unaltered to the Humanities techniques, tools and technologies that have been already developed in computer science. On the other, given the characteristics of the Humanities set out above there is considerable scope for the computer scientist to develop innovative techniques, tools and technologies, in cooperation with Humanities scholars. Not only have such collaborations the potential to push forward the state of the art of eHumanities, but may result in a wechselwirkung this is of value to computer science also.

Before addressing specific research questions in turn we will say something of the broad outlines of eHumanities, so as to sketch the broad outlines of the research landscape that the computer scientist will encounter as well as its relationship to what may be termed the traditional Humanities.

Considering the centrality of Text to many disciplines of the Humanities (such as, inter alia, History and Biblical Studies), the research questions of the early pioneers of the application of computing technology to the Humanities and the level of technological sophistication then attained it should come as no surprise that Text has been a dominant research object of Humanities computing since the 1940s. This is reflected in the structure of the 2006 "Companion to Digital Humanities" which covers music, multimedia, performing arts and art history, nevertheless its parts II "Principles", III "Applications" and IV "Production, Dissemination, Archiving" are almost exclusively dedicated to textual resources. Though Text remains a central research field, it is clear that in the 2010 research landscape and in numerous recent publications (e.g. Ashgate's Digital Research in the Arts and Humanities series) this balance has been notably redressed. Digital Humanities research on topics such as audio-visual media; visualisation and modelling of complex data such as historical events; mining, and automatic analysis and classification of audio; automated image recognition, interface design and high-resolution digital libraries for the arts, and many more continues to grow. Relevant research topics in such areas will be treated of in this report.

Regardless of the type of medium that Humanities scholars address self-reflection lies at the heart of the Humanities [ESF 5] and much research is characterized by the importance of human endeavour and methodological plurality, with scholars often taking very individual approaches to their research questions and reflecting their source material in the light of their specific premises. This is not to say that theoretical frameworks are not frequently drawn on in humanities research. For example, Structuralism [B-4, B-5], deconstruction [B-3], semiotics, marxist / materialist criticism [B-6], McLuhanesque cultural and media studies [B-7] and many more have been extensively drawn upon in the humanities. Yet, contrary to most natural and social sciences, the coupling of those frameworks to concrete procedures of interpretation is less than evident even for structuralism, the most formalist of them — not to speak of their possible representation in machine-executable algorithms. The interpretation of texts is in the humanities still largely unformalized. In addition to the dominant hermeneutic approach a more quantitative tradition, somewhat modelled on the practice of the empirical methods of the social sciences, is popularized by societies such as the International Society for the Empirical Study of Literature and Media. While gaining in popularity, it is not necessarily widely applied amongst practitioners (as many proponents of this approach themselves deplore [B-8]). Indeed, perhaps most important for the computer scientist to reflect on is that Humanities scholars consider process to be as important as outcome, if not more so, and their aim is rarely to solve problems, once and for all, but rather to rediscover and reinterpret them in order to ask new and better questions. In the field of Humanities Computing this has been addressed by McCarty [R-2], in particular, who has convinvingly argued of the inappropriateness of the 'knowledge juke-box' view of computing for the Humanities. Implicitly referring to the nature and implications of Turing's universal machine he has reflected that "… my field is centred on computing, and so orbits a process that is in principle limited only by the human imagination, and so is for the indefinite future capable of surprising us with new means for investigating whatever we fasten on". A central pillar of the Humanities has always been the building of the "tools of the trade", for example, critical editions of key texts and lexicographic resources (dictionaries and encyclopaedia). Accordingly, Digital Humanites, which focuses on the remediation of cultural heritage in the digital realm, has addressed such resources from its inception. From Father Busa, passing by epoche-making projects such as the digital library of the Thesaurus Linguae Graecae that since the 1970s has become an integral part of the toolkit for the Classics, making its way from tapes to the web [W-10, cf. also B-1:82]. Even today many of the activities in the eHumanities are linked to retrodigitizing and marking up texts, to the elaboration of dictionaries and in general to the preparation, presentation and exchange of knowledge bases that can be the basis for further interpretative research. In a similar vein, as an aid for and first step towards the actual interpretation of texts some projects have from early looked towards the rule-based analysis of electronic text corpora. Initially, algorithms would be used to analyse certain characteristics of texts en masse and with a precision that would have been difficult or impossible to attain manually. An example of this is Wilhelm Ott's metrical analysis of Vergil's Aeneis starting in the late 1960s / early 1970s [W-9], which addressed the goal of providing reliable data on its metrical structure. It is in fact for this task that the first modules of what was later to become TUSTEP were created. In a similar vein, TACT was created as a tool for linguistic / statistical analysis of texts and their vocabulary. This trend is today represented at a more sophisticated level by natural language processing (NLP) tools such as GATE or the Natural Language Toolkit (NLTK) for Python that are situated on the borderline to corpus linguistics, two saplings of the same root.

In consequence many computer scientists collaborating with humanities scholars work in some form on tools that support these types of tasks. We have already mentioned TUSTEP. TextGrid itself is equally a perfect example, but for many other current and historical projects such as ARCHway / EPPT, TaPOR, this applies just as well. The consequences are not to be taken lightly — using large text corpora allows for a precision difficult to achieve before and the use of IT permits enriching texts with links to their sources in various media in a way unheard of before. We shall look in more detail below at the consequences of this for computer science research and its various subfields.

Interestingly enough, while many scholars galdly integrate some tools and especially digital libraries into their research, the methodology of traditional humanities scholarship seems so far strangely unaffected by these advances — or if it is affected these changes are rarely reflected upon in the humanities themselves. Some still rare exceptions such as Matthew L. Jockers' recent headline-making assignment of 1200 novels to one literature class to encourage the statistical interpretation of texts across a complete period and the ensuing debate on the validity of this approach — "the Humanities Go Google" [W-9, Humanist, 24.86] — if anything rather help to highlight this. Nevertheless, there are indications that this is changing. In addition to well established and internationally recognised centres such as the Centre for Computing in the Humanities, Kings College London and the Centre for Digital Humanities, Universität Trier a recent proliferation of Digital Humanities centres all over the world is evident - for example in University College London, the Universities of Göttingen and Köln as well as the establishment of the Digial Humanities Observatory in Ireland. Furthermore, Digital Humanities is being recognised at both the national European level (e.g. AHRC's recognition of 'Digital transformations in the Arts and Humanities' as a highlight notice[2] ); the US (2007 setting up of the NEH Office for the Digital Humanities[3]) and the pan-European levels (e.g. the identification of Research Infrastructures by the EU as central to realising the European Research Area and the reflection of this in FP7 programmes and the ESFRI road map) lead one to the conclusion that Digital Humanities has a central role to play not only in the academy but also in terms of Europe's research competetiveness. The tide may well be turning.

## 2.1. eScience and eResearch

If eScience — "computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing" [W-8] — and eResearch are also about supporting current research methodologies in the various disciplines, it takes methodological questions more seriously. "e-Science will change the dynamic of the way science is undertaken" (John Taylor, Director General of Research Councils, Office of Science and Technology).[4]

## 2.2. eResearch

The terminological distinction between eScience and eResearch is yet not clearly established. We follow the usual English terminology to see eResearch as the wider term that encompasses all disciplines including the hard sciences, the social sciences and the humanities, which coincides with the approach taken by the Oxford e-Research Centre. In the nature of things it is methodologically more heterogeneous than eScience which in this understanding is Research being applied to the natural / hard sciences. eHumanities inscribes itself squarely as a part of the overall eResearch movement without being eScience.

eResearch is by definition heavily technology-oriented, embracing distributed computing, often specifically Grid computing, and large distributed data sets — not only as key tools but also as key methods. It is collaborative and interdisciplinary in its vision, one of the goals being explicitly to link data sets across institutions and disciplines to get new insights that would otherwise have been impossible.

eResearch also aims to revolutionize the way research results are published by blurring the line between the publications and the research itself. De Roure postulates the need for what he

[2] See http://www.ahrc.ac.uk/FundingOpportunities/Pages/highlight-notices.aspx

[3] http://www.neh.gov/odh/

[4] John Taylor, Director General of Research Councils UK), cited after http://www.lesc.ic.ac.uk/admin/escience.html

terms Research Objects that would become the "fundamental sharable description of piece of research", replacing current academic papers [W-18]. It would have the "Six Rs of Research Object Behaviours" and be:

1. Replayable
2. Repeatable
3. Reproducible
4. Reusable
5. Repurposeable
6. Reliable

To fulfil these requirements, [W-18] proposes that a research object bundle the data and the corresponding software routines in a collection, possibly described as an RDF graph as specified by the OAI *Object Exchange and Reuse* standard [W-19].

The "myExperiment experiment" http://www.myexperiment.org sets out to create such a living space for research objects. Workflows, data and outside resources can be combined and published as "packs" in the myExperiment environment.

## 2.3. Data-driven science

While many eResearch activities just model existing research workflows electronically, a strong strand of eResearch looks towards data-driven science. Data-driven science clearly postulate a new methodological approach to the sciences and by extension to all research. It attempts to supplement the traditional approach of hypothesis- or model-driven research that has been the criterion for scientific work since Popper's *Logic of scientific discovery* [B-10] by building on data mining approaches to recognize patterns in existing published data (the term "data mining" here taken in a wide sense), very often also across disciplines: "we collect data first, then see what it tells us [...] The basic idea is that if we can collect enough data to form a large, rich picture [...] then we are likely to learn something by looking at it" [W-20, acknowledging that this has been a practice for long in some fields]. In its most extreme incarnations, data-driven science may even be seen to generate and self-test hypotheses in a fully automated way.

This idea is intimately linked to the linked data initiative that Tim Berners-Lee kicked off in 2006 [W-12]. Linked data consciously embraces also the methodological implications of the new approach: Using an example from the medical domain, Berners Lee postulates that "the power of being able to ask those questions as a scientist — questions which actually bridge across disciplines — is really a complete sea-change" (Tim Berners-Lee in his Ted talk on linked data, [W-11]). Hans Rosling's work on statistics on (especially) health and poverty-related issues — which Berners-Lee not coincidentally cites regularly — points much in the same direction: freely available data is considered an essential tool to permit not only a clearer view of the world, but also to allow research and progress to flourish in ways otherwise impossible [W-13].

Other statements mirror this sentiment of linked data as a methodological sea-change. David de Roure speaks in his DEST 2010 keynote [W-21] of linked data as "Datascopes – telescopes for the naked mind". It is clear that in such a vision data mining and pattern recognition algorithms are essential. De Roure speaks in his aforementioned keynote of the "digging into data challenge". A number of domain projects are currently already starting to respond to this challenge, e.g. in the textual sciences, in musicology and in linguistics.

Some proponents of data-science even postulate that the necessary tools will soon be available for "tech-savvy non-programmers" [W-22] using off-the-shelf programs such as Excel to mine the data (connectivity in social-network in the case of [W-22]). In fact, the use of social networks such as Facebook and Twitter is common fare in data-driven science, notably as a means of conducting polls to gain empirical data in response to a given research question, sometimes in combination with serious games, or directly by analyzing the social graph.

While many of the implications of these premises are as of yet not well understood, it is clear that Computer Science plays a central role in the eScience and eResearch world.

## 2.4. Hermeneutic Informatics

Much older than eScience and eResearch, but astonishingly similar in its objectives and language is Father Busa's project of Hermeneutic Informatics, going back to the 1950s, but still very much pertinent [cf. B-2 and A-8]:

> [...] it is thus necessary for the use of informatics to reformulate the traditional morphology, syntax, and lexicon of every language. In fact all grammars have been formed over the centuries by nothing more than sampling. They are not to be revolutionized, abandoned, or destroyed, but subjected to a re-elaboration that is progressive in extent and depth.
>
> Schematically, this implies that, with integral censuses of a great mass of natural texts in every language, in synchrony with the discovered data, methods of observation used in the natural sciences should be applied together with the apparatus of the exact and statistical sciences, so as to extract categories and types and, thus, to organize texts in a general lexicological system, each and all with their probability index, whether great or small. [B-2]

The methods that Busa proposes are closely related to those of today's corpus linguistics, but with the ultimate goal of laying a sound statistical basis for the correct understanding of a given corpus.

## 2.5. eResearch methods in the Humanities

Franco Moretti, a Stanford professor of English, retakes the metaphor of the datascopes and comparative literature, who uses it when speaking of literature studies based on Google Books: "It's like the invention of the telescope. All of a sudden, an enormous amount of matter becomes visible" (cited after [W-9]). It is equally underlying metaphor behind Jockers' 1200 book assignment (cf. above). This approach in fact starts to blur the classical boundaries between the Sciences and the Arts by becoming "more data-intensive, information-intensive, distributed, multi-disciplinary, and collaborative", to cite Christine Borgman's DH 09 keynote [A-5].
[A-2] and [A-6] explicitly discuss these challenges in the light of eScience.
Gregory Crane primarily elaborates on challenges and requirements coming from the "galaxies of data" now available for data-driven science. [A-2] sketches a very ambitious programme for an eResearch in the eHumanities (the article itself speaks of eScience), focussing primarily on data-related issues. He emphasizes the need for "systems that can make specialized content intellectually as well as physically accessible" across language and domains. Given the highly multilingual nature of humanities texts and Crane's background in the Perseus project, tools for automatic translation even for historical languages such as Ancient Greek, Latin, classical Arabic, Chinese, Sanskrit, Old Norse, Syriac, Akkadian, Sumerian, Middle High German figure highly on his agenda. Beyond the pure translation this involves for him OCR tools for historic scripts, non-normalized spellings and especially handwritten material. From there, he sees the need for tools that enrich raw text to structured data: semantic classification, morphological analysis, syntactical analysis,... Many of these currently exist for modern mainstream languages such as English and German, but typically not for historic languages of limited industry relevance.
In addition to the information retrieval aspects, [A-2] looks into spatial queries that link spatial data in the narrow sense coming e.g. from global positioning systems (GPS) systems to corresponding information. This linking process involves for him amongst others information extraction and named entity identification.
For Crane recommendation systems can play an important role in linking up resources (for recommendation systems cf. also [B-11]).
Tobias Blanke in [A-6] looks at the specific challenges that datasets can pose, using amongst others the challenges of historical census data: «As for all data-driven development,

managing the data deluge in the humanities means starting from the requirements of the specific dataset in scope». Furthermore, such datasets can also pose organizational challenges, e.g. in the case of commercially sensitive data or data that might impede on privacy rights. However, he also identifies a number of cross-cutting concerns on the visualization of very high-resolution 2D and 3D scan data, the use of collaborative video annotation in projects e.g. related to choreography as in the e-Dance project. Furthermore, he looks at the example of the *Medieval Warfare on the Grid* project which is in the process of modelling and reenacting the battle of Manzikert on the Grid. Doing so involves 3D modelling and means for synchronous audio-visual interaction with the system. Other activities concern the further development of central infrastructure tools such as Fedora Commons.
John Unsworth [A-1] primarily looks at the collaborative aspects of Digital Humanities in which he underlines the democratization of the academic discourse through discussion lists and similar media. However, for the acquisition of formal social capital he sees many challenges ahead, both for collaborative work, but also quite practically for the citations for digital publications and long-term digital preservation of purely digital scholarship ("No peer review, no publishers, no archives—the situation looks pretty grim for dances with wolves"). While not per se a CS research topic, he identifies here a key challenge for the long-term success of eHumanities – a challenge that exists *mutatis mutandis* also for CS.

# 3. Relevant Research Challenges in the Intersection of eHumanities and CS

## 3.1. eResearch and Data-driven Science

Based on the discussion in the previous sections on eResearch and data-driven science we can identify the following CS research challenges:

### 3.1.1. eResearch

*Research Challenge:* If the prospect of a Research Object is key to the eResearch vision, it will be highly non-trivial to put into practice outside a fixed framework such as myExperiment, given that both the workflows, the data and the software routines are typically not easy to bundle, but are intimately linked to a given software stack (often involving proprietary tools) and especially for the hard sciences sometimes even to specific hardware (sensors...). It will be a major IT challenge to even define the technical preconditions and the needed specifications for interoperable Research Objects — one that in the end has a lot in common with the work on mid- to long-term digital preservation of research data.
*Research Challenge:* Support the use of modelling of specific historic events as currently done for the battle of Manzikert. Much of such work will probably be project specific before a generalized framework could be envisonaged

### 3.1.2. Data-driven Science

*Research Challenge:* Find new pattern recognition algorithms. Those algorithms will be in large part media-dependent and specific to the requirements of the humanities' discipline in question.
*Research Challenge:* Elaborate new, probably domain-specific methodologies to visualize and validate possible patterns that have been identified.
*Research Challenge:* Develop a methodology and a framework for pattern recognition across account multilingual data (including historic language forms)
*Research Challenge:* Develop project-specific and dataset tools for pattern recognition across specific datasets

*Research Challenge:* Develop tools and methodologies for the crowd sourcing especially for polls to support empirical research methods for the humanities

## 3.2. Data representation and data formats

### 3.2.1. Data Preparation

Very much work is necessary to meet the challenges set out in [A-2], in particular:
*Research Challenge:* Develop OCR tools for historic writing systems and language forms

*Research Challenge:* Develop tools for transforming raw text into structured data for historic writing systems and language forms. In particular, work on their

1. semantic classification
2. morphological analysis
3. syntactical analysis
4. entity recognition
5. In the long term the full corpus linguistics tool set currently available for modern language forms will have to be available also for historic forms.

*Research Challenge:* Develop tools for the semi-automatic detection of links between raw texts (citation detection / text-text linking, including typed links and links between concepts), taking into account the humanities' concept of citation
*Research Challenges:* Strategies and tools for the creation of annotations to existing texts in the domain using:

- non-hierarchical structures (e.g. critical editions) beyond the current models proposed by the TEI, including the text's genesis and responsibilities for parts of the text
- stand-off annotation for the annotation of potentially volatile media (text, images, music, video), especially relating to synchronization between the annotated text and its annotations
- tools for adding semantics to Manuscript Images and other media, e.g. Text-Image-Link-Editor [W-4] and audio-visual media as in the case of the eDance project.

### 3.2.2. Data representation: XML, SGML and more

Digital Humanities were instrumental in popularizing markup standards. From the earliest days onwards Busa used some form of positional markup. In the late 1980s, humanities computing became one of the early adopters notably of SGML through the Guidelines of the Text Encoding Initiative (TEI), the first official version (P1) of which was published in 1990, followed in 2002 by the first XML version (P4). TEI had and continues to have an immense influence in the humanities as the quasi-standard for encoding humanities texts. Indeed, Michael Sperberg McQueen, the then editor of the TEI Guidelines was the co-Editor of the XML specification.
Since its creation in 1996 XML has of course been popularized across industry and a whole set of specifications have been created around it, but even now enough research questions remain specifically for the representation of humanities and linguistic texts, notably:
*Research Challenges:* Explore strategies and build tools for the representation of annotations to existing texts in the domain using:

- non-hierarchical structures (e.g. critical editions) beyond the current models proposed by the TEI, including the text's genesis and responsibilities for parts of the text
- stand-off annotation for the annotation of potentially volatile media (text, images, music, video), especially relating to synchronization between the annotated text and its annotations

*Research Challenge:* Develop tools for the visualization of specific media, e.g. very high resolution 2D and 3D scan data.

## 3.2.3. Data visualization and publication

The visualization of humanities data is first of all very much a topic for the humanities themselves. While the classical publication formats of the Gutenberg age had centuries to mature and are typically not controversial, digital editions are even today “incunabula: productions that replicate in new media the limits of the old“ [A-2, cf. also A-9]. We must "Develop[…] the forms of publication" [B-1, p.209ff] adequate for the digital age, both for scholarly output and for source material such as manuscripts, early print versions etc. that at present are not yet systematically published.
Nevertheless, the topic does have a CS impact in that publications need to meet a number of requirements:

- Humanities publications are a specific kind of research object and hence sharing the research challenges mentioned there, including long-term availability of the text itself and any dependencies it may have (sources, software,...)
- Publications are typically multi-channel, hence the reuse of the data for visualization and publication in various publication channels including paper, the Web, eBooks and other, as yet unknown forms of publication must be maximized – all this taking into account that the data in question may depend on other data and / or on specific software modules
- The various manifestations of the text must be interlinked, e.g. by offering the page numbering of the print manifestation also in the electronic versions
- Since texts are being elaborated increasingly in collaboration or with contributions of third parties through discussion lists, chats etc., we need visualization strategies for this, including the need to identify when needed the respective responsibilities for parts of the data
- The genesis of the data must be visualized, possibly across versions.
- The data must face both the revolution in access and convenience of use, and in the way a digital edition itself is conceived and used, i.e., rather than a fixed, authoritative work, the goal being to create a work that is open to updates / collaboration / change while still being citable [W-4, A-10].
- Many text types such as critical editions and commentaries are heavily interlinked, non-linear texts. For print publications there are widely accepted methodologies to present them e.g. using critical apparatuses, for their digital equivalents no definitive conventions yet exist
- The data may be from the start or become at a later stage part of research collections and then must support cross-corpus searches, more generic display options etc.
- The currently fixed boundaries between text and source image and even between genres is becoming blurred by bundling both in research object.[5]

This list of requirements is by no means complete. This research is intimately linked to ongoing research on digital editions and long-term archiving.
*Research Challenge:* Representation of data on the net, developing together with humanities scholars best practices for data visualization and envisioning tools to support it
*Research Challenge:* Representation of data in print, developing tools to support the publication of XML data in line with the requirements above.

[5] The journal Vectors (http://www.vectorsjournal.org/journal/ index.php?page=Introduction) is perhaps relevant here because it offers a forum for works that «that need, for whatever reason, to exist in multimedia» since they involve «moving- and still-images; voice, music, and sound; computational and interactive structures; social software; and much more»

## 3.3. Repositories and Long-term archiving

In a manner of speaking, digital repositories[6] are a focal point where the above mentioned issues of data representation, visualization and publication, data-driven science, and collaboration encounter each other. Unlike most of the hard sciences like particle physics, humanities' repositories provide a wide variety of discipline- and institution-specific standards and implementations. Especially when working with individually collated corpora integrating/addressing resources from more than one repository, human as well as machine-based agents have to deal with heterogeneous interfaces, data formats, object representations, metadata standards, and recently modified or deleted contents. Therefore the most important (and promising) Research Challenge is the development of an Open Repository Environment enabling the federation of distributed repositories and to link them with external added-value services, an environment "in which – rather than monolithic repository systems sitting on dedicated hardware – repository functionalities are part of a larger, open environment of decentralised, collaborative agents (e.g. repositories, registries, re-representation and preservation services)" (cited after [A-20]). This challenge is subject of current research activities (e.g. [B-12]) – taking into account the interoperability issues discussed in the following sections. The development of a federated repository infrastructure for the Arts and Humanities in Germany will be subject of another TextGrid report (R1.3.1, forthcoming).

Though crucial not only for the eHumanities, the long-term archiving of research data doesn't impose any serious domain-specific challenges on CS research. This may change over the years, when the first productive systems will have been running long enough to allow considering how and whether the technical results met the expectations and needs of the research communities. Since the requirements for long-term archiving/preservation of research data are in a large part not of a technical but organisational and administrative nature[7], the corresponding challenges have to be solved primarily by the research communities themselves. In [W-26] the WissGrid project gathered the requirements concerning the long-term archiving of research data of some of the research communities participating in the German D-Grid initiative. The resulting attribution to one of the three proposed application profiles (Grid workflow, interactive research environment, federated archives) takes place not only according to criteria which are directly related to discipline-specific research but also to general legal and institutional conditions. The recommendation for (German) linguistics is to implement the "interactive research environment" application profile for long-term archiving, see also [W-27] – the one that TextGrid will follow. However, the WissGrid architecture for long-term archiving of resarch data and the corresponding application profiles are thoroughly generic, i.e. there are no prescriptions or limitations concerning metadata schemas, file formats, object or content models etc.

## 3.4. Semantic Web

Semantic Web technologies are heavily used in eResearch in general and in the eHumanities in particular. As mentioned eResearch is intimately linked to the Linked Open Data (LOD) movement aiming towards a web of data. In projects such as TEI and TextGrid the representation of metadata for the media objects and their collections is a major concern to enable their precise retrieval, especially for facetted search in combination with full-text search, in addition to enable sharing with third party providers and consumers.

---

[6] Since there is still no consensus about the distinctive features of a repository in contrast to other digital collections, we prefer – for the sake of simplicity – the quite generic definition given in [W-28, section 1.2].

[7] Cf. for instance the Trusted Repositories Audit & Certification: Criteria & Checklist (http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf ) or the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA, http://www.repositoryaudit.eu/).

However, in all these cases the focus is far more on the interoperable representation of metadata than inferencing that is at the core of specifications such as OWL. For this reason, also Topic Maps (ISO/IEC 13250) are often used internally to represent data.
Much of the current research on LOD is likely to be of immediate benefit to the eHumanities.

## 3.5. Architecture

### 3.5.1. Technical interoperability

Due to small research groups and low budgets that are quite common in the humanities, distributing services into a Service Oriented Architecture (SOA) with distributed resources is often the best choice for these projects. Various groups can easily reuse these services, even if they are spread far apart, and they can be reused in later projects of the same group. The more generic a service can be designed, the higher its reusabillity. Additionally, increasingly pervasive broadband connections make accessing and utilizing distributed services easy and efficient.
Distributed services today commonly fall in to one of the following architectures: Service Oriented Architecture using SOAP (SOA/SOAP), Resource Oriented Architecture (ROA) and Event Driven Architecture (EDA). All these architectures share a technical base. They mainly differ in the semantics. SOA/SOAP refers to services as a way of doing something with objects, so semantically SOA/SOAP is based on verbs. ROA refers to services using the REST (REpresentational State Transfer) idea of resources, so ROA concentrates on nouns instead of verbs. EDA, also called SOA 2.0 sometimes is based on messages exchanged with the various services and between them, so it focuses on responding to events. The discussion which of these architectures is best suited for the eHumanities is still ongoing. Because none of these approaches has significant flaws or benefits compared to the others, a final conclusion to this discussion seems unlikely. Depending on the problem to solve, a specific architecture can minimize the effort necessary to implement its solution, but a workable solution is viable with any of the architectures. Larger projects with multiple developers will most likely mix the different paradigms in the end.
While in itself less of a research challenge at present, providing guidance on these issues is clearly a CS task and an important part of CS education.
Very much still a *research challenge* both in academia and in industry is the technical cross-framework interoperability of GUI components in the web that are the basis for Web 2.0 applications (mashups).

### 3.5.2. Semantic interoperability / resource registries

Semantic interoperability permates much of eHumanities, but at times in various guises. TEI itself is, of course, a means to enable semantic interoperability of textual data, even though the necessary flexibility and extensibility of TEI makes this interoperability not quite seamless. TextGrid's baseline encoding is a way to define a minimal TEI for cross-corpus analysis and searches.
The LOD idea is a means to ensure at least the technical and syntactical interoperability of metadata descriptions and much of the work is focussed on defining micro-schemata for metadata to facilitate also their direct semantic interoperability.
A major focus of current work and research not only in the eHumanities is on not only sharing the metadata related to media resources, but also that making up the descriptions of services. The ultimate goals is less the fully-automated composition of workflows and their subsequent automatic execution, but to facilitate their manual composition, taking into account both content resources and services. This means that human-readable descriptions, possibly enriched with a recommendation system, are at least as important as machine-

tractable information.[8] However, this information needs to be shared between various resource providers and even across various eHumanities Digital Ecosystems.
Currently many projects in the eHumanities domain use home-grown registries, as many of the standardized registries concentrate almost exclusively on services and are too complex for the eHumanities domain (e.g. the ebXML Registry Model defines 17 classes and brings with it its own user management). However, homegrown resource registries by themselves are difficult to federate, and the federation of resource descriptions across systems is necessary. The European SDShare specification specified in CEN CWA 15971 and in particular its third part on SDShare CWA 15971-3 [cf. also A-15] offer a RESTful low-tech solution based on Atom-feeds that is now being used in TextGrid.
At present, in spite of some first steps such as [A-14] no universally accepted metadata schema / ontology exists for eHumanities resources.
*Research Challenge:* Develop in collaboration with humanities scholars a system of federated registries for the eHumanities and define a universally accepted ontology for eHumanities resources.

### 3.5.3. Digital Ecosystems

Digital Ecosystems are "open, loosely coupled, demand-driven, domain clustered, agent-based self organized collaborative environment where species/agents form a temporary coalition (or longer term) for a specific purpose or goals, and everyone is proactive and responsive for its own benefit or profit" [W-24]. They involve various agents, human, organizational and machine-based. eHumanities systems are digital ecosystems for the humanities domain.
Digital Ecosystems focus on collaboration, interoperability and reusability of services and resources, rather than a specific realization through a given technology stack. Grid technologies are only one possible way of realization, though conceptually Digital Ecosystems are very close to the idea of semantic grids, seen as an «extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation», as [W-25] puts it with reference to Tim Berners Lee's definition of the semantic web. In fact, the resources and services used in any given Digital Ecosystem will typically be hosted on a variety of platforms, involving both Grid-based and other systems, and often involves on the user-facing part Web 2.0 applications, often leveraging standard mashups e.g. from Google, Yahoo or Facebook. In fact, a major challenge ahead will be to render key Grid services usable through simple RESTful services also for applications that do not subscribe to the Grid paradigm per se (e.g. on the model of Amazon's S3) – and vice-versa. Of course, classical repositories can also be built on cloud services, as in the case of e.g. the Fedorazon project, which uses Amazon S3 as storage layer for Fedora Commons (cf. also [A-19]).
The workflows in eHumanities Digital Ecosystems typically involve humans, and services, at times also organizations. They might be systems driven by explicit workflows executed by workflow-execution engines, but might just as well be realized as Web 2.0-style mashups.
Also, a lot can be learned for eResearch architecture from highly scalable, high availability architectures in industry, e.g. from Facebook [A-17], Google [W-18] and Amazon.
*Research challenge:* Abstract key Grid middleware services behind simple, RESTful interfaces to help opening up existing grid-based eResearch / eHumanities systems to wider collaboration across Digital Ecosystems ("hiding complexity")
*Research challenge:* A major challenge not only for eHumanities will be to reconcile the classical Grid architecture and its inherent development methodologies with the requirements of a fast moving, very loosely-coupled Web using very different development strategies, often based on mashups and reuse of simple components.

[8] This experience has been made in TextGrid and in various eGovernment projects [A-16], but is complemented by the Taverna experience in myExperiment [A-4].

## 3.6. Horizontal interaction with other fields of application

eHumanities shares not only many commonalities with eScience and eResearch, but also with other highly distributed, process-driven Digital Ecosystems in application areas such as eBusiness and eGovernment. Not only are many of the overall architectures comparable, but also many of the concrete infrastructure tools and services used, e.g. for authentication and authorization, interoperability, registries, workflow engines and basic cloud services such as storage or map-related services. For these reasons many of the challenges faced are similar.
Yet, there is little to no cross-fertilization between those fields of application. Applied CS is ideally positioned to provide these links by treating these domains as instantiations of a more generic discipline of heterogeneous distributed systems and by also seeking suitable publication venues.

## 3.7. Software engineering

Any development for the eHumanities and indeed any eResearch discipline is "software design for empowering scientists" [A-4], i.e. it must start out from the requirements of scholars. Based on the experience of myExperiment [A-4] defines six principles to follow:

1. „Fit in, don't Force Change": fit from current research practices rather than trying to impose new workflows
2. „Jam Today and more Jam Tomorrow": enable incremental adoption of the software
3. „Just in Time and Just Enough": embrace incremental development
4. „Act Local, think Global": Start out with a small sample number of test users, then generalize
5. „Enable Users to Add Value": provide extensibility and customisation for end users
6. „Design for Network Effects"

While many of the principles are of course true more or less for any type of end-user facing software development, it is more pronounced in eHumanities than in other branches in that in many of the hard sciences in that humanities scholarship tends to be somewhat more distant to IT technology than the practice in the hard sciences where the scientists themselves often act as software developers.

### 3.7.1. Tool development

Because of the gap between CS and the humanities, traditional software development models like the Waterfall model [W-14] or the Spiral model [W-15] do not work well for eHumanities projects. The traditional models demand a complete set of project requirements before any implementation is done. But at the beginning of an eHumanities project these requirements are usually either vague or incomplete, due to the limited mutual understanding of needs and opportunities between computer scientists and humanities scholars. This results in software that is either not as powerful as it could be, is not sufficiently visible to end users, or that does not fully match the needs of its users.
Modern project management models such as SCRUM [W-16] or software development models such as eXtreme Programming [W-17] are much more effective, because these models do not demand fixed requirements, but develop requirements within the project runtime. They also enforce close cooperation between the software developers and the humanities scholars on a regular basis. Because of these regular discussions both sides develop a better understanding of each other. And because of the variable requirements this enhanced understanding can then result in more accurate and more specific requirements, which are necessary for powerful and well-suited software products.
Of course, many of the tools to be developed will be heavily specific to a given humanities discipline or even specific to individual projects. Many tools are likely to bring with them their own research challenges for CS.

## 3.7.2. Usability and User Involvement

In general, humanities research can only be partially automated, even though individual tasks can. In other words, typical eHumanities workflows involve both humans and software agents to an even higher degree than most for other eResearch activities. In addition, humanities scholars are often not particularly focussed on technological questions and are typically operating on restricted budgets, making their involvement in complex eHumanities projects more challenging that in many other settings.
In addition, humanities scholarship tends to be less collaborative than that in the hard sciences, in spite of the tendency towards collaborative efforts in recent years, or at least the growing awareness that sharing information/creating standards is vital to project development and long-term success.
The focus on collaboration and its increasing importance for the humanities is reflected by studies such as [A-12]. For example, emblem studies, a multi-disciplinary research area which serves as a good example of how individual scholars and institutions are learning that they do not always have to "reinvent the wheel" for each new project. The "three phases" have been identified in emblem studies [A-11] for the adoption of collaboration: moving from individual, unique efforts to adopting externally-generated programs to hands-on community-based development of specialized programs / standards.[9]
Given the restricted budgets, in general it is preferrable to work with easily accessible low-tech solutions whereever possible. It is not coincidental that many of the currently popular cloud-services such as Dropbox or Amazon WS are often leveraged in eResearch applications in general and the eHumanities in particular.
*Research Challenge:* Elaborate and document software engineering techniques that facilitate the «Roll-in of users» [W-21] specifically in the humanities.

## 3.7.3. Tests and Test methods

As already mentioned above, humanities scholars usually only have a non-technology focussed view of computers and software. They usually do not know how to react properly to faults or improper behaviour of the software they use. Also research projects in the humanities can extend over several years, so it is crucial that results of these researches are kept safely and are not damaged by defective code. To ensure these errors do not occur to software in production use, software tests are a very important part of the development process. Of course software tests should be an important part of any software development process, but there are specialized tests necessary, due to the architecture usually used in digital humanities projects.
Research in the humanities is normally done by single scholars or in very small groups. But often these researches involve complex operations on large amounts of data. This type of research is done most cost-effectively, when using distributed services. On one hand this reduces the cost for implementing the same solutions to specific problems in various projects again and again, on the other hand it is not necessary to purchase powerful hardware for only a few hours or days of computational work.
For distributed services the usual testing methods relying on unit- and functional tests are insufficient. Originating from the distributed nature of these services there are additional problems to deal with, namely the long runtime of the services and the parallelism of the requests. Also public exposure of the service can become a problem, because the service then has to face the various attack threats of the Internet.

[9] "In a first phase, lasting from 1983 to about 1993, a plethora of competing standards combined with costly and primitive technology made the very process of emblem digitization difficult, cumbersome, and expensive, and essentially precluded most forms of effective collaboration. The second phase, between 1993 and 2003, has been characterized by a wave of convergence as cross-platform software standards came to be adopted, a development that enabled the first discussions about implementation of collaborative solutions to take place. We now stand at the beginning of a third phase, characterized by emergent solutions in which common standards should enable the accomplishment of goals shared by the community of emblem scholars". [A-11].

The long runtime of a service can impact its stability. Memory leaks, especially small ones, are usually no problem for a desktop application, because normally this kind of application is used only for a few hours or probably a few days. Usually the system's memory is large enough to handle smaller memory leaks in such a limited amount of time. For a distributed service, running for weeks and months, this surely can become a major problem, because the small memory leak would pile up and up until the system crashes. This would then result in data loss and researchers having to delay their work, until someone restarts the service. Load-testing tools like Tsung [W-6] and JMeter [W-7] have proven to be a good way for locating those small memory leaks. These tools send thousands of requests to the service, simulating a continuous use of the service over time. When monitoring the system's memory consumption during these tests the presence of memory leaks can be easily detected. Unfortunately these results only give small hints as to the causes of the memory leaks, so tracking them down and fixing them can be very time-consuming. Even though most modern programming languages offer memory management, memory leaks still can be a major problem, resulting from mistakes like not closing network connections or not removing objects from global lists after they are not used anymore.
The parallelism of the requests can result in concurrency problems. Again desktop applications are less affected by this problem, because there is only one user starting operations at a time. But in a publicly available service multiple users can initiate the same operation at the exact same time. Therefore shielding the various contexts of these requests against each other is very important; otherwise side-effects from one request can lead to incorrect results or even to a crash of another request. The only known way to detect concurrency problems is vast load-testing in extremely parallel environments with an extended analysis of the results of each request. Again the load-testing tools come in handy, because most of them offer the execution of web requests using various, predefined parameters and are also capable of analyzing the responses to these requests by a given logic. All of the tools mentioned above are distributable, which means they can send their requests in real parallelism from multiple computers at the same time. Again the results of the tests only give small hints about the exact location of the concurrency problem.
Being publicly available distributed services usually are exposed to all kinds of threats from the Internet like denial of service attacks or hacking. Due to the variety of the methods of attack, the targeted weak points and the chosen approach to security an automated testing for service-security is not yet available. Specially trained experts in security audits must ideally perform these tests.
In the eHumanities, in fact in all web-based services, load-tests are a well suited addition to the traditional unit- and functional tests. Not only do they reveal several weakness of code, when exceeding a certain load on the targeted system, they can also give an impression of how the service will react to a denial of service attack.
*Research Challenge:* To help finding memory leaks, develop a system to track objects in memory, grouping them by similarity, while the application is load-tested and setting up relations between requests and memory usage. A generic approach would be good, to allow the system to work for any programming language. Also it is important, that this memory profiler does not interfere with the load-test itself, and does not invalidate it's results, e.g. by consuming too much of the CPU power of the tested system or by serializing parallelism.
*Research Challenge:* Develop a system to analyze the source code of an application to find sections, accessible to multiple threads. Check the variables of these sections for possible concurrency problems. This would make finding concurrency problems easy.
*Research Challenge:* Evaluate the use of functional programming for eHumanities applications to avoid concurrency issues.

## 3.8. Incentive Systems

As for [A-1] underlines from a humanities perspective, suitable incentive systems – systems for accumulating social / academic credit – are very important for the acceptance of a branch of research in academia. Current CS incentive systems currently often generate innovative proof of concept implementations with corresponding papers, but little if any academic credit

is typically accrued by the sweat job of actually bringing those proofs of concept into a production ready state, so they too often stay in the state of the demonstrator. However, proofs of concept, while interesting in themselves from a CS point of view, are pretty much useless for anyone trying to work with them, including humanities scholars. While not a research topic in a strict sense, a successful academic collaboration between CS and humanities scholars implies a means to actually cite completed tools as a separate item for a researcher's academic credit.

# 4. Summary

This report identifies a number of Computer Science research challenges ahead of us. Many are specializations of more generic challenges stemming from eScience, eResearch, data-driven science, or generic software engineering. However, many challenges are very specific to the humanities and a number will even be specific to individual humanities projects. Some of those challenges are currently attacked in existing research projects such as TextGrid, TEXTvre, Digilib, the Medieval Warfare on the Grid and many others in collaboration between humanities scholars and computer scientists. However, many challenges still wait for their champions.
The identified list is not comprehensive, but is a first step towards a full roadmap that will mature over time. However, many tasks will not in themselves be research topics, but nonetheless necessary for a fully functional eHumanities ecosystem. Incentive systems in CS will have to take also into account this need: the ecosystem needs a production-ready infrastructure and production-ready services and tools to flourish.

# 5. Bibliography

## 5.1. Books and Overview Studies


1. Willard McCarty, *Humanities Computing* (Basingstoke: Palgrave Macmillan, 2005).
2. Susan Schreibman, Raymond George Siemens, and John Unsworth, *A companion to digital humanities*, Bd. 26 (Malden, MA: Blackwell Pub., 2004), http://www.loc.gov/catdir/toc/ecip0415/2004004337.html.
3. Jacques Derrida, *De la Grammatologie* (Paris: Éditions du Minuit, 1967).
4. Ferdinand de Saussure, *Cours de linguistique générale*, hg. v. Tullio de Mauro (Payot, 1916).
5. Claude Lévi-Strauss, *Anthropologie structurale.* (Paris: Pocket, 1985).
6. Fredric Jameson, *The political unconscious: narrative as a socially symbolic act* (Ithaca N.Y.: Cornell University Press, 1982).
7. Marshal McLuhan, *Understanding media* (Routledge, 1964).
8. Willie van Peer, Jemeljan Hakemulder, and Sonia Zyngier, *Muses and measures: empirical research methods for the humanities* (Newcastle: Cambridge Scholars, 2007).
9. Wilhelm Ott, *Metrische Analysen zu Vergil* (Tübingen: Niemeyer, 1973).
10. Karl Popper, *The logic of scientific discovery* (London; New York: Routledge, 2002).
11. Elizabeth Chang, Farookh Hussain, und Tharam Dillon, Trust and reputation for service-oriented environments : technologies for building business intelligence and consumer confidence (Chichester England ;;Hoboken NJ: John Wiley & Sons Inc., 2006).
12. Andreas Aschenbrenner, *Reference Framework for Distributed Repositories – Towards an Open Repository Environment* (PhD Thesis, Göttingen 2010), http://webdoc.sub.gwdg.de/diss/2010/aschenbrenner/

## 5.2. Web Sites


1. “Digital humanities - Wikipedia, the free encyclopedia,” http://en.wikipedia.org/wiki/Digital_humanities.
2. "Publications of the association for literary and linguistic computing", http://www.allc.org/content/pubs/index.html
3. "List of Digital Humanities annual conferences", http://www.allc.org/content/conf/index.html
4. "Abstracts of all papers of the Digital Humanities conference 09," http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf
5. “Humanities - Wikipedia, the free encyclopedia,” Mai 29, 2010, http://en.wikipedia.org/wiki/Humanities.
6. "Tsung load testing framework", http://tsung.erlang-projects.org/
7. "Apache JMeter", http://jakarta.apache.org/jmeter/
8. “e-Science - Wikipedia, the free encyclopedia,” o. J., http://en.wikipedia.org/wiki/E-Science.
9. “The Humanities Go Google - Technology - The Chronicle of Higher Education,” Mai 28, 2010, http://chronicle.com/article/The-Humanities-Go-Google/65713/?sid=wc&utm_source=wc&utm_medium=en.
10. “Thesaurus Linguae Graecae - General Information,” *Thesaurus Linguae Graecae - Project History*, October 20, 2009, http://www.tlg.uci.edu/about/history.php.
11. T. Berners-Lee, “Tim Berners-Lee on the next Web | Video on TED.com,” Februar 2009, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html.
12. T. Berners-Lee, “Linked Data - Design Issues,” July 27, 2006, http://www.w3.org/DesignIssues/LinkedData.html.
13. Hans Rosling, “Hans Rosling shows the best stats you've ever seen | Video on TED.com,” 2007-06, http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html.
14. "Waterfall model - Wikipedia, the free encyclopedia", http://en.wikipedia.org/wiki/Waterfall_model
15. "Spiral model - Wikipedia, the free encyclopedia", http://en.wikipedia.org/wiki/Spiral_model
16. "Scrum - Wikipedia, the free encyclopedia", http://de.wikipedia.org/wiki/Scrum
17. "Extreme Programming - Wikipedia, the free encyclopedia", http://de.wikipedia.org/wiki/Extreme_Programming
18. David De Roure, “Replacing the Paper: the six Rs of the e-Research Record | e-Research,” *Replacing the Paper: the six Rs of the e-Research Record*, August 2009, http://blog.openwetware.org/deroure/?p=56.
19. OAI, “Open Archives Initiative Protocol - Object Exchange and Reuse,” *Open Archives Initiative: Object Reuse and Exchange*, October 17, 2008, http://www.openarchives.org/ore/.
20. Eric Drexeler, “The Data Explosion and the Scientific Method,” *The Data Explosion and the Scientific Method*, Oktober 25, 2008, http://metamodern.com/2008/10/25/the-data-explosion-and-the-scientific-method/.
21. David De Roure, “Semantic Grid and Sensor Grid: Insights into the e-Research Ecosystem,” *Semantic Grid and Sensor Grid: Insights into the e-Research Ecosystem*, April 15, 2010, http://www.myexperiment.org/packs/109.html.
22. Mac Slocum, “Data science democratized - O'Reilly Radar,” *Data science democratized. With new tools arriving, data science may soon be in the hands of non-programmers*, Juli 1, 2010, http://radar.oreilly.com/2010/07/data-science-democratized.html.
23. ThatCamp, “Manifesto for the Digital Humanities,” Mai 2010, http://dl.dropbox.com/u/1744854/DH%20MANIFESTO.pdf.
24. Elizabeth Chang and Marc Wilhelm Küster, “IEEE-DEST 2011 ----- Daejeon, Korea,” IEEE DEST: Background and Objectives, Juni 2010, http://dest2011.debii.curtin.edu.au/.
25. David De Roure, “Semantic Grid Document Store and Bibliography,” Semantic Grid Vision, 2001, http://www.semanticgrid.org/vision.html.

26. WissGrid Report “Generische Langzeitarchivierungsarchitektur für D-Grid”, http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.1-LZA-Architektur-v1.1.pdf
27. WissGrid-Spezifikation: Grid-Repository, http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.5.2-grid-repository-spezifikation.pdf
28. Rachel Heery and Sheila Anderson, “Digital Repositories Review” (UKOLN, ahds), 2005, http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf
29. Turkell, William J. “A few Arguments for Humanistic fabrication” [Webblog entry] *Digital History Hacks: Methodology for the Infinite Archive* (2005-08). 21. November 2008. Online 20. August 2010.

## 5.3. Articles

1. John Unsworth, "The Humanist: ‘Dances with Wolves’ or ‘Bowls Alone’?,” in *Scholarly Tribes and Tribulations: How Tradition and Technology Are Driving Disciplinary Change*, 2003, http://www.arl.org/bm~doc/unsworth.pdf.
2. Gregory Crane, Alison Babeu, and David Bamman, “eScience and the Humanities,” *International Journal on Digital Libraries* 7, 2007: 117-122.
3. Marc Wilhelm Küster, Christoph Ludwig, and Andreas Aschenbrenner, “TextGrid as a Digital Ecosystem,” in *DEST 2007*, hg. v. Elizabeth Chang, 2007.
4. David De Roure and Carole Goble, “Software Design for Empowering Scientists,” *IEEE Software* 26, Nr. 1 (1, 2009): 88-95.
5. Christine L. Borgman, “Scholarship in the Digital Age: Blurring the Boundaries between the Sciences and the Arts,” in *Digital Humanities 2009: Conference Abstracts* (held during the Digital Humanities 2009, University of Maryland, College Park, 2009), xvi, http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf.
6. T. Blanke, M. Hedges, and S. Dunn, “Arts and humanities e-science—Current practices and future challenges,” *Future Generation Computer Systems* 25, Nr. 4 (4, 2009): 474-480.
7. A. B. M Russel and Asad I Khan, “Towards dynamic data grid framework for eResearch,” in *Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54*, ACSW Frontiers '06 (Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2006), 9–16, http://portal.acm.org/citation.cfm?id=1151828.1151830.
8. Roberto Busa, “From Punched Cards to Treebanks. 60 Years of Computational Linguistics,” in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories* (gehalten auf der Eighth International Workshop on Treebanks and Linguistic Theories, Milano: Ente per il Diritto allo Studio Universitario dell’Università Cattolica, 2009), 1-2, http://tlt8.unicatt.it/allegati/Busa_abstract_TLT8.pdf.
9. Gregory Crane u. a., “Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries,” in *Research and Advanced Technology for Digital Libraries*, hg. v. Julio Gonzalo u. a., Bd. 4172, Lecture Notes in Computer Science (Springer Berlin / Heidelberg, 2006), 353-366, http://dx.doi.org/10.1007/11863878_30.
10. Peter Shilingsburg, “How Literary Works Exist: Convenient Scholarly Editions,” *Digital humanities quarterly* 3, Nr. 3 (Summer 2009), http://digitalhumanities.org/dhq/vol/3/3/000054/000054.html.
11. David Graham, “Three Phases of Emblem Digitization: The First Twenty Years, the Next Five,” in Digital Collections and the Management Knowledge: Renaissance Emblem Literature as a Case Study for the Digitization of Rare Texts and Images, DigiCULT Publications, 2004, 13-18.
12. Annamaria Carusi and Torsten Reimer, “Virtual Research Environment Collaborative Landscape Study.” Report issued by the JISC in January 2010. http://www.jisc.ac.uk/publications/reports/2010/vrelandscapestudy.aspx.
13. Christoph Ludwig and Marc Wilhelm Küster, “Digital ecosystems of eHumanities resources and services,” in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, 2008, 476-481.

14. Andreas Aschenbrenner u. a., “Open ehumanities digital ecosystems and the role of resource registries,” in *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies* (gehalten auf der 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST), Istanbul, Turkey, 2009), 745-750, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5276672.
15. Graham Moore and Marc Wilhelm Küster, “Protocol for the Syndication of Semantic Descriptions,” in *Subject-centric Computing. Fourth International Conference on Topic Maps Research and Applications, TMRA 2008*, 2008.
16. Marc Wilhelm Küster, Graham Moore, and Christoph Ludwig, “Semantic Registries,” in Berliner XML Tage 2007, hg. v. Robert Tolksdorf and Johann-Christoph Freytag (Berlin, 2007), 21-36.
17. Jason Sobel, “Building facebook,” in Proceedings of the 1st ACM symposium on Cloud computing - SoCC '10 (gehalten auf der the 1st ACM symposium, Indianapolis, Indiana, USA, 2010), 87, http://portal.acm.org/citation.cfm?doid=1807128.1807142.
18. Asit K. Mishra u. a., “Towards characterizing cloud backend workloads,” ACM SIGMETRICS Performance Evaluation Review 37, Nr. 4 (3, 2010): 34.
19. Andreas Aschenbrenner, Tobias Blanke, and Mark Hedges, “Synergies between Grid and Repository Technologies - A Methodical Mapping,” in 2008 IEEE Fourth International Conference on eScience (gehalten auf der 2008 IEEE Fourth International Conference on eScience (eScience), Indianapolis, IN, USA, 2008), 778-781, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4736898.
20. Andreas Aschenbrenner, Tobias Blanke, Marc W. Küster, Wolfgang Pempe, "Towards an Open Repository Environment," Journal of Digital Information, Vol 11, No 1 (2010), http://journals.tdl.org/jodi/article/view/758

## 5.4. Reports and Lectures

1. Standing Committee for the Humanities (SCH). Position paper 2007. (European Science Foundation, 2007).
2. Willard McCarty, “Attending from and to the machine” (Inaugural lecture, Kings College London, Februar 2, 2010).