

# Grass-roots Research in Arts and Humanities e-Science in the UK

Tobias Blanke, Mark Hedges & Stuart Dunn  
Centre for e-Research  
King's College London

September 22, 2008

## Abstract

The aim of this paper is to showcase recent developments in the UK arts and humanities e-Science initiative. A specific grass-roots research agenda in arts and humanities e-Science has been developing over the past few years in the UK. In this paper, we will look at the new research questions and institutions emerging in the many e-Science grass-roots activities.

## 1 Introduction: Humanities Computing and e-Science in Arts and Humanities

Arts and Humanities e-Science has proved not only difficult to define, but also a paradox in itself. This is rooted in a language distinction, which probably hints at all language distinctions at more than a difference in the way of talking about things. The English language distinguishes between the 'humanities' and 'science' in a way that reflects a specific treatment of the division of labour in research. In German both are *Wissenschaften*, and in French both are sciences. In English, science is not just a general state of knowing; it is linked to specific scientific methods, which are mainly found in natural sciences. This makes 'e-Science in (or for) Arts and Humanities' a paradox, which is not helped by the general confusion about what e-Science, in itself, might be. The more straight-forward term e-Research however was resisted, on the grounds that Arts and Humanities e-Science is about using advanced technology to tackle Arts and Humanities research questions; technology that originates from far outside the arenas familiar to Arts and Humanities researchers, even those traditionally involved with ICT.

There has been a long-standing tradition of UK Humanities Computing with the Centre for Computing in the Humanities (CCH) at King's College London as an important and unique institution. Humanities Computing can be compared to what is known as computational sciences in disciplines like chemistry, physics, etc. Humanities Computing is centrally concerned with the common methods and techniques for analysing source materials, tackling problems and communicating results in the humanities disciplines. In this sense, e-Science in humanities is part of the tradition of

Humanities Computing in general. However, there is a lot of work to be done to see humanities e-Science as a continuation of what Humanities Computing has striven to achieve in its 50 years history rather than as an opposite or independent movement. Text Encoding is a good example. It has been at the heart of Humanities Computing for a long time. It is widely used and at the same time of high theoretical impact. Contrary to other Humanities Computing activities, it is (already) recognized as a valid addition to humanities research in general by employing computational methods. Text (en-)coding in the humanities, however, has mainly been centered around markup. The Text Encoding Initiative (TEI) standard, has been very successful as a way for humans to annotate texts and make them understandable to other humans. It remains to be seen whether the standard can have a similar impact for providing a metadata standard to exchange texts between computational agents. This however would be its main use case in more computationally oriented e-Science for humanities. It has been tried as a metadata exchange standard in the TextGrid project<sup>1</sup> in Germany as well as in the MONK project<sup>2</sup> in the US. But both of them used only a subset of the extensive TEI specification in order to achieve TEI based interoperability.

This paper is not so much a comparison of arts and humanities e-Science results with those from Humanities Computing in general. This will be the subject of a workshop at the 2008 UK e-Science All Hands conference.<sup>3</sup> In this paper, we will rather look at the new research questions emerging in the many e-Science grass-roots activities. We have done this elsewhere in more detail [1]. We amended this work with recent activities in UK arts and humanities e-Science and a more specific focus in the analysis of particular activities. The next section will look at the grass-roots activities, starting with the analysis of how UK e-Sciences made the case for arts and humanities e-Science by gathering user requirements. Afterwards, we look at the ways the data deluge was addressed, and continue with investigating performance arts activities. In the remaining two sections, we discuss arts and humanities e-Infrastructure activities at the King's College Centre for e-Research before concluding with a discussion whether e-Research is a better term to describe the often difficult rise of e-Science tools and methods in the arts and humanities.

## **2 Grass-roots: e-Science and its difficult rise in the arts and humanities**

E-Science in the arts and humanities makes the double claim that arts and humanities research can be advanced by scientific and advanced computing methods at the same time. This double claim has helped develop grass-roots activities, which included collaboration of arts and humanities researchers with scientists and computing experts. The first section will investigate whether there is a case for e-Science in the arts and humanities by looking at user requirements.

---

<sup>1</sup><http://www.textgrid.de>

<sup>2</sup><http://www.monkproject.org>

<sup>3</sup><http://www.allhands.org.uk/2008/programme/workshop10.cfm>

## 2.1 User Requirements

User requirements play an important role in lifting e-Science systems from a field of computing research experimentations to production systems. The Arts and Humanities e-Science Scoping Survey, having conducted detailed research across seven Arts and Humanities subject domains, defined e-Science as 'the development and deployment of a networked infrastructure and culture through which resources - be they processing power, data, expertise, or person power - can be shared in a secure environment, in which new forms of collaboration can emerge, and new and advanced methodologies explored'.<sup>4</sup> The e-Science Scoping Study can be considered as an important input into the larger 'e-Uptake' research, recently commissioned by JISC.<sup>5</sup> Final results from this research into the barriers of uptake of e-Infrastructures can be expected to be published in 2009.

The e-Science Scoping Study final report was published this year and presented the 'grand challenges' for digital research in arts and humanities and how tools and methodologies from e-Science might help address these. The scoping study emphasizes that e-Science in the arts and humanities is justified in so far as a common agenda of grand challenges is developing. The e-Science Scoping Study prepared the UK arts and humanities e-Science initiative as a set of grass-roots activities to work on these research questions by suggesting to 'embed' e-Science in research practices. In the UK, even though funds were very limited, they were still distributed across as many partners as possible. This allowed e-Science to emerge from practice and led to some unforeseen approaches to deal with concrete research problems.

A general training and support services as well as a kind of dating agency were also suggested by the scoping study. This led to the funding of the Arts and Humanities e-Science Support Centre.<sup>6</sup> The main task of this centre is to encourage cross-domain and inter-disciplinary collaborations and to function as a kind of translator between the different languages of arts and humanities research and computing. The study also suggested to continue the research on the impact of e-Science tools and methodologies with further anthropological and sociological studies, in order to better place e-Science within the real world of arts and humanities research. An e-Science infrastructure in which arts and humanities research can be embedded, must therefore not just exist of technology. More important are the people and institutions that help with engagement and provide training services. It has certainly helped the success of UK e-Science that appropriate support institutions were planned in from the start. Centres of excellence and national and regional support services are an essential part of a successful cyberinfrastructure.

The larger e-Science Scoping Study was accompanied by at least two more series of workshops and experiments which looked at more specific requirements such as novel ways of establishing requirements and ways of integrating the Access Grid into Humanities research. In Oxford, three workshops entitled *User Requirements Gathering for the Humanities* were held at the Centre for the Study on Ancient Documents in order to analyse best practices in user requirement studies for e-Science solutions that

---

<sup>4</sup><http://www.ahds.ac.uk/e-science/e-science-scoping-study.htm>

<sup>5</sup><http://www.e-researchcommunity.org/projects/e-uptake/>

<sup>6</sup><http://www.ahessc.ac.uk>

work for Humanities research.<sup>7</sup> A series of workshops organised by the Humanities Research Institute in Sheffield investigated *The Access Grid in Collaborative Arts and Humanities Research*.<sup>8</sup>

As it is relatively easy to see potential benefits of Access Grids for arts and humanities research, they have attracted the attention of researchers relatively early. Arts and humanities research often takes place in highly specialised domains and subdisciplines, niche subjects with expertise spread across universities. The Access Grid and related technologies based on voice over IP can provide a cheaper alternative to face-to-face meetings. During the Sheffield workshops, the Access Grid was met with great enthusiasm but also disappointment due to obvious still unresolved technology related problems. However, a more conceptual problem has also been identified. Access Grid might be a good environment to substitute some face-to-face meetings, but lacks innovative means of collaboration, which can be especially important in arts and humanities research. The organizers would like to see a new focus on how to realise real multicast interaction, as it has been done in VNC technology or basic wiki technology. These could support those new models of collaboration, which e-Science is supposedly about.

The workshop series in Sheffield offered insights into what the specific interest of humanities researchers in dealing with advanced network technologies were. The next subsection will present what becomes necessary for these researchers when they start dealing with the new mushrooming of digital resources.

## 2.2 Managing the Data Deluge

Humanities e-Science application face challenges common to many other application domains. Everything begins with the data available. There is definitely a data deluge in humanities research data, which is the result of two complementary developments. Firstly, old analogue data in humanities is being transformed into digitized shapes. The massive digitisation project of Google Books is just one example. But the quantitative challenge is only one dimension of the challenge which is probably still manageable. Even in projects with an output of Petabytes of data such as the US Shoah Foundation Archives of testimonials of Holocaust survivors, the actual research data will be limited to Gigabytes of image and text files, which still does not compare to the size of research data produced by automatic simulations in the sciences. When it comes to humanities data, the challenges are more rooted in its inherent complexity and inconsistency, as it is mainly generated by humans directly.

A small series of workshops in 2007 investigated how to productively deal with the inherent inconsistency and fuzziness of humanities data. ReACH, workshops about *Researching e-Science Analysis of Census Holdings*, were held at UCL in London.<sup>9</sup> ReACH set out to work on available digitisations of historical census data held at the private company Ancestry.co.uk. Ancestry has built up several Terabytes of census holdings data worldwide and has digitised the censuses of England and Wales under license from the UK's National Archives. For e-Science research, these datasets pose specific challenges. First and foremost, records are incomplete due to the fact that they

---

<sup>7</sup><http://ahessc.ac.uk/files/active/0/URH-report.pdf>

<sup>8</sup><http://ahessc.ac.uk/files/active/0/AG-report.pdf>

<sup>9</sup><http://ahessc.ac.uk/files/active/0/ReACH-report.pdf>

are created by humans. This always leads to inconsistencies, but things are made worse by the fact that these are historical records of the 19th century when data acquisition standards were not yet well developed. Normally, the censuses were captured by visiting households and speaking to whoever answered the door. Say, an interview in 1851 recorded a 19-year old male named Adam in a particular household. Then, it can be the case that, in 1861, there is either no Adam anymore in the household and nobody remembers his existence, or there is an Adam, but now he is supposedly 41. In order to establish whether it is the same Adam, census historians apply heuristic methods that deliver probabilities of data matches or data links across different census holdings [3]. In order to effectively support such research, a system of (semi-)automatic matches of records would be needed to create what is known as a longitudinal database of individuals across the census. For that, systems have to be trained to incorporate modelling procedures as currently applied by historians. As the public is generally very interested in census data, it seems also feasible to use Web 2.0 techniques and social computing applications to enhance the current data resources. Many members of the public are keen genealogists and best experts on their family's history.

In computer science, technologies such as data and text mining or information retrieval deal with unstructured, potentially inconsistent data. It is therefore not surprising to find them strongly represented in arts and humanities e-Science projects. The York-based Archaeological Data Service (ADS) will be responsible for developing *Archaeotools: Data mining, faceted classification and e-Archaeology*.<sup>10</sup> Over 40000 reports of grey literature in archaeological excavations lie potentially idle, as they are hard to access. Archaeotools is an attempt to provide access to these records using automated metadata generation techniques. These will index datasets for new links in the records in terms of When, What and Where. The underlying datasets combine over one million records from the National Monuments Records of Scotland, Wales and England as well as Historic Environment Records from numerous local authorities and the ADS's own archive holdings.<sup>11</sup> The formation of the facets is supported by existing thesauri and the University of Edinburgh's geoXwalk service,<sup>12</sup> which will support geospatial information access to the data. A 3D-space will visualise facets and their links and provide access to deeper unpublished archaeological literature, whereas users will be able to ask for their own specific research interests to be represented in the indexing of these research records. This will create flexible access to resources up to now neglected in research.

Another library and archive focussed project is based at UCL. *E-Curator: 3D colour scans for remote object identification and assessment*<sup>13</sup> will use University College London's collections and state of the art 3D colour scanner, which can revolutionize the traditional methods in museums and archives based on text and images. UCL hosts 3 museums, 10 departmental collections and half a million objects and specimens. Their Arius3D scanner is the first of its kind in Europe providing high resolution 3D geometry through the use of a laser triangulation system at a 100 micron points spacing. Colour information is captured with red, green and blue lasers. The project uses

---

<sup>10</sup><http://ads.ahds.ac.uk/project/archaeotools/>

<sup>11</sup><http://www.nesc.ac.uk/action/esi/download.cfm?index=3714>

<sup>12</sup><http://www.geoxwalk.ac.uk/>

<sup>13</sup><http://www.museums.ucl.ac.uk/research/ecurator/>

3D recording to describe artefacts as a whole - in the first instance 6 pilot projects for case studies. This method will offer yet unknown details and insights into the object's structure. Such 3D scans could then help with the identification of degraded surfaces. They would allow comparisons of whole three-dimensional objects. Of particular interest to the project are grid technologies which allow to share such objects, as they are often difficult to analyse by a single research team alone and will have to be correlated with others. Data grid technologies like SRB will be used to share and organise data. A portal will allow researchers to annotate and view the objects and share information about them. Users can e.g. zoom or rotate the 3D representations or change the lighting conditions and the colour mixture. This way, UCL hopes to work towards establishing 3D scan data as a curatorial tool. Thus, e-Curator is as much a methodological experiment as it is a technological one.

The challenge of humanities data becomes even more complex if we start talking about performance live data as it is dominant in the arts.

### 2.3 Performance Collaborations

At the Universities of Bedfordshire, Manchester and the Open University, the project *Relocating Choreographic Process (e-Dance)* on *The Impact of Grid technologies and collaborative memory on the documentation of practice-led research in dance* uses the Access Grid in the analysis and creation of distributed performances. On the one hand, new technologies can deliver enhancements of arts practices, while on the other, technology and digital world are still not able to fully mimic the analogue world, which is often seen as a limitation. Networks naturally produce delays in transmissions and Access Grids therefore are never able to fully synchronize activities across two different sites, which is to the disadvantage of performances. This delay, however, can be seen either as a problem or as an integral part of the creative process itself. Networked artworks could introduce these into their own scripts and thus acknowledge the specific characteristics of this new material for arts. Technology and creativity are not dichotomous, but are mutually dependent. But in order to support creative processes better, the Access Grid has to become more like its original promise, more like the new interface to the computing and network resources. This interface would include a whole room as an alternative to the classic desktop, as it has been envisioned in 'Group-Oriented Collaboration: The Access Grid Collaboration System' [2].<sup>14</sup>

The experiments on *Data Services for Associated Motion Capture User Categories (AMUC)* in Newcastle targeted the tracking and capturing of motions that go beyond the everyday use of human bodies.<sup>15</sup> Complex, coordinated movements produced by performing artists should benefit areas that require exact measurements of human body movement like medical engineering. Here, the complex and fuzzy data typical to arts and humanities can assist science directly. The AMUC collaboration first worked on the exact definition of user requirements regarding motion capture data in order to develop data retrieval methods for motion capture resources via grid technologies. This work included capture and storage with a Vicon advanced motion capture system, and

---

<sup>14</sup><http://www.ahessc.ac.uk/e-dance>

<sup>15</sup><http://ahessc.ac.uk/files/active/0/AMUC-report.pdf>

advanced computational methods for analysis and visualisation of such data, which comprised methods for the sketch-based retrieval of the data. The prototype retrieval tool impresses with its idea to build a sketch-based interface in order to best mimic the thinking and understanding of a very particular user group - live arts researchers including martial arts. Furthermore, it is a true e-Science project in so far as a wide range of different researchers were involved from traditional engineering and commercial experts in the field of live databases to performance artists.

Digital Music is easily available for download from the Internet. It is therefore no wonder that researchers are looking at novel ways of how to use these resources for musicology research. At Goldsmith, University of London, the project *Purcell Plus* will build upon the successful collaboration *Online Musical Recognition and Searching (OMRAS)*,<sup>16</sup> which has just achieved a second phase of funding by the EPSRC. *Purcell Plus* will investigate the impact advances in music information retrieval from the OM-RAS project can have on musicology research. *musicSpace: Using and Evaluating e-Science Design Methods and Technologies to Improve Access to Heterogeneous Music Resources for Musicology* in Southampton also intends to make online music resources available for musicology research methodologies. *musicSpace* will bring together different resources into one single user interface to avoid researchers having to carry out multiple searches on multiple resources for their research questions. Resources will be exposed as web services.

The last section is now more of an outlook on how to integrate the existing grassroots activities into a larger e-Infrastructure.

## 2.4 E-Infrastructure

A more recent development in UK e-Science is the emergence of university-based e-Research institutions, such as the Oxford e-Research Centre, or the Centre for e-Research at Kings College London (CeRch). These centres are called e-Research rather than e-Science institutions to open up towards academic domains such as the arts and humanities, which have not yet been served by developments in e-Science. CeRch is heavily involved in several attempts to establish building blocks of a future e-Infrastructure for arts and humanities research. Together with CCH and in the context of its task to build an e-Infrastructure for King's College London, it will work on at least two case studies on how to link digital humanities work to the National Grid Service.

1. The first project attempts to advance the use of HPC in digital humanities. The Nineteenth Century Serials Edition (NCSE) corpus contains circa 430,000 articles that originally appeared in roughly 3,500 issues of six 19th Century periodicals. Published over a span of 84 years, materials within the corpus exist in numbered editions, and include supplements, wrapper materials and visual elements. Currently, the corpus is explored by means of a keyword classification, derived by a combination of manual and automated techniques. A key challenge in creating a digital system for managing such a corpus is to develop appropriate and innovative tools that will assist scholars in finding materials that support

---

<sup>16</sup><http://www.omras.org/>

their research, while at the same time stimulating and enabling innovative approaches to the material. One goal would be to create a 'semantic view' that would allow users of the resource to find information more intuitively. However, the advanced automated methods that could help to create such a semantic view require processing power that is currently not available to CCH researchers. CCH has implemented a simple document similarity index that would allow journals of similar contents to be represented next to each other. The program used the *lingpipe*<sup>17</sup> software to calculate similarity measures (specifically, the TF/IDF similarity measure on character n-grams) for articles within the corpus. A test using 1,350 articles, requiring a total of 910,575 ( $n * (n-1) / 2$ ) separate comparisons, was executed on a Mac Mini, which took 2 days to process 270 documents, that is to perform  $270 * 1,349 = 364,230$  comparisons. Assuming the test set was representative, a complete set of comparisons for the corpus would take more than 1,000 years. CeRch will build a Campus Grid Toolkit based on CON-DOR and a connection to the National Grid Service that will allow us reduce this time significantly.<sup>18</sup>

2. The second proposal concerns advanced data integration. A current CCH project is preparing a publication of the inscriptions from Cyrenaica and Tripolitania, Roman provinces in modern Libya, although the wider group is working with inscriptions from elsewhere around the Mediterranean, as well as with papyri. The data resources produced by these projects include corpora of texts marked up using TEI, and in particular using EpiDoc, which is an implementation of TEI developed specifically for inscriptions. As part of the Integrating Digital Papyrology project, CCH researchers have been working with copies of the Heidelberg Gesamtverzeichnis der griechischen Papyrusurkunden gypens (HGV), a database of papyrological metadata. This database contains general information on some 65,000 papyri, including bibliography, dates, and places (findspots, provenances), mostly from Roman Egypt and the environs. The proposed work will use the data integration platform OGSA-DAI<sup>19</sup> to provide an integrated view of at least two databases (with different schemas) in the field of epigraphy and one database together with an XML collection in EpiDoc.

CeRch is also a lead partner in the European Union funded project DARIAH.<sup>20</sup> DARIAH stands for 'Digital Research Infrastructure for the Arts and Humanities' and is funded to conceptualize and afterwards build a virtual bridge between different humanities and cultural heritage data resources in Europe. The project aims to improve access to the many arts and humanities resources locked away in archives, libraries and museums all over Europe. To form the initial digital infrastructure of DARIAH, data centers in France, Germany, the Netherlands and the United Kingdom joined forces. Since then the consortium has grown to 14 partners from ten countries. The partners will work out in more detail the plans for the actual construction of DARIAH, including which national/European grid infrastructures it should use. In order to (at least

---

<sup>17</sup><http://alias-i.com/lingpipe/>

<sup>18</sup><http://www.omii.ac.uk/solutions/campusgrid.jsp>

<sup>19</sup><http://www.ogsadai.org.uk/>

<sup>20</sup><http://www.dariah.eu>

partly) build a virtual bridge for European arts and humanities research data, DARIAH has identified several alternative but possibly complementary solutions. DARIAH will probably concentrate on a combination of the digital repository system Fedora together with iRODS<sup>21</sup>, the new data grid technology developed by SDSC, to support the flexible, distributed virtualised storage of files. DARIAH will commence work in September 2009.

These e-Infrastructure initiatives have more or less directly emerged from the grass-roots activities in so far as they close gaps and attempt to bring together different tools and methods into one framework. They can be seen as a major emerging trend for future arts and humanities e-Science.

### **3 Conclusion and Future Trends: e-Science and e-Research**

The e-Science initiative in the UK has sparked enthusiasm and desire in the arts and humanities community to work together with scientists and computing researchers to solve challenges posed by the new digital resources available in arts and humanities. The activities within the UK's arts and humanities e-Science community demonstrate the specific needs that have to be addressed to make e-Science work within these disciplines. We could see how e-Science in the arts and humanities has matured towards the development of concrete systems and embedded infrastructures that systematically investigate the use of e-Science for research.

Arts and Humanities e-Science has never limited itself to more traditional ideas of e-Science, which link e-Science mostly to the application of certain advanced network technologies in supporting sciences. These technologies like the grid have their place but they might generate the impression that the solution is already there and only needs to find the right application and those willing to give the money. From our experience also in the wider world of e-Science, e-Science will fail if it deems itself as just an application of technologies. Rightfully, it will then be perceived as an invasion of some technology know-it-alls, which know the solution without knowing the problem. This perception might be even worse in arts and humanities. Their development and success was at least in the past seldomly linked to advances in computing. Men and women are still vastly superior to machines when it comes to discussing history, analysing concepts or revolutionizing arts.

As described, there is a trend in the UK to rebrand existing e-Science centres and projects into e-Research activities. This is justified as it recognizes the growing importance of new communities for e-Science. Social sciences and humanities have played a major role in the e-Science programme in the UK. In so far, e-Research in the arts and humanities might be a better term. However, as demonstrated, many of the successful grass-roots activities in arts and humanities e-Science originated from newly formed productive partnerships with sciences. This meant not only to work together on a common problem but also the adoption of a new different viewpoint in humanities research regarding old questions. For example, looking at the way computer science

---

<sup>21</sup><http://irods.sdsc.edu/>

tries to simulate the understanding of meaning in music and texts helped the research practice of textual studies and musicology. Discovering the statistics behind mining technologies means understanding what a more mathematically oriented methodology can deliver and what it will fail at. This enhanced perspective on research practices and methodologies is also part of the theme of e-Science which will hopefully not get lost in the rebranding as e-Research.

## 4 Acknowledgments

We would like to thank all the principal investigators of the projects discussed here for their collaboration and support for our work. It would take too much space to list all of them here, a detailed description of their work and their reports can be found on the AHeSSC website,<sup>22</sup> but their research has been the foundation for the success of the Arts and Humanities e-Science Initiative. Furthermore, AHRC, EPSRC and JISC have shown a remarkable will to experimentation in their commitments to arts and humanities e-Science and to cross-council collaboration.

## References

- [1] Tobias Blanke, Mark Hedges, and Stuart Dunn. E-science in the arts and humanities - from early experimentation to systematic investigation. In *E-SCIENCE '07: Proceedings of the Third IEEE International Conference on e-Science and Grid Computing*, Washington, DC, USA, 2006. IEEE Computer Society.
- [2] I. Foster and C. Kesselman. *The Grid 2: Blueprint for a New Computing Infrastructure*. Morgan-Kaufmann, second edition, 2004.
- [3] Kevin Schuerer and Matthew Woollard. National sample from the 1881 census of great britain 5sample. working documentation v1.1, 2002.

---

<sup>22</sup><http://www.ahessc.ac.uk/projects/>