Taking the long view: from e-Science humanities to humanities digital ecosystems

Anderson, Sheila; Blanke, Tobias

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems

Sheila Anderson & Tobias Blanke [*]

Abstract: *»Auf lange Sicht: Von den Geisteswissenschaften im e-Science Kontext zu einem Geisteswissenschaftlichen digitalen Ökosystem«*. In this paper we investigate the importance of research infrastructures for arts and humanities research. We seek to outline the development of a digital research infrastructure localised in the science and engineering domain and framed within the concept of e-Science. We define the primary characteristics of e-Science as big data and big structures such as the grid and high performance computing. We will attempt to demonstrate the transfer of the e-Science paradigm to the humanities and to assess what worked and what did not. We then suggest how thinking about technology and infrastructure through and within the humanities can lead to transformation and finish with a suggestion that the future for humanities research infrastructures is best framed around the emerging idea of a humanities specific digital ecosystem.

Keywords: research infrastructures, big data, digital ecosystems, humanities e-Science.

Introduction

> The word technology, which joined the Greek root, techne (an art or craft), with the suffix ology (a branch of learning), first entered the English language in the seventeenth century. At that time, in keeping with its etymology, a technology was a branch of learning, or discourse, or treatise concerned with the mechanic arts. (Marx 2010)

For this paper we have been asked to champion 'big structures' as the most suitable technical template for research infrastructures as opposed to 'lightweight web' technology. The immediate problem presented is to understand what is meant by 'technical template' in this context. Contemporary discourse about technology frequently treats it as a 'thing' – it is a device, a piece of hardware, or it is a software component or an application – it is, as Marx argues, "the material component" of the infrastructure. This narrow prism serves to distance technology from the social, economic, political and epistemic rela-

[*] Address all communications to: Sheila Anderson, Centre for e-Research, 26-29 Drury Lane, Room 301, King's College London, London WC2B 5RL, UK;
e-mail: sheila.anderson@kcl.ac.uk.
Tobias Blanke, Centre for e-Research, 26-29 Drury Lane, Room 223, King's College London, London WC2B 5RL, UK; tobias.blanke@kcl.ac.uk.

tions of which it is a part with the consequence that it may serve to give more power and authority to the purely technological than is welcome, and to fold it into an 'aura of phantom objectivity' (Marx 2010). For this paper therefore, we shall attempt to reclaim the original meaning of technology and to frame our argument as a 'discourse concerned with the mechanic arts' where the mechanic arts in this context are the digital research infrastructures and technological structures that enable and support humanities research work.

In this paper we look first to set our discussion in its historical context referencing works on the history of large technology systems. We then seek to outline the development of a digital research infrastructure localised in the science and engineering domain and framed within the concept of e-Science. We define the primary characteristics of e-Science as big data and big structures such as the grid and high performance computing. We will attempt to demonstrate the transfer of the e-Science paradigm to the humanities and to assess what worked and what did not. We then suggest how thinking about technology and infrastructure through and within the humanities can lead to transformation and innovation and enable us to re-think research infrastructure around the problems that humanities scholars are confronting as they interact with the emerging research infrastructure environments. We finish with a suggestion that the future for humanities research infrastructures is best framed around the emerging idea of a humanities specific digital ecosystem.

Historical Context: Large Technology Systems

Thomas Hughes has written extensively on the history of 'large technology systems' (LTS) (Hughes 1986, 1993). Hughes started his work by investigating the growth of electric light and power systems questioning "how the small, intercity lighting systems of the 1880s evolved into the regional power systems of the 1920s" (Hughes 1986). He was interested in why growth occurred in some cases and failure in others and for this he looked not just at the technological issues but also at the wider context in which these systems developed – the social, political, legal and other elements that acted alongside the technology as drivers of growth or causes of failure. He concluded that the key factor was interactivity and argued that the history of technology could be written only by taking a *systems* approach that considered the whole and not just the technological part. It was the interaction between the component parts (human, technological, social etc.) that was the real force behind the shape of the emerging systems and their growth or failure.

Hughes defined complex systems as "coherent structures comprised of interacting, interconnected components that ranged from relatively simple mechanisms to regional power electric supply networks" (Hughes 1993). Evolution of LTS was characterised by disruption, intervention, competition, and could result in failure, as well as success. The conceptual model for LTS sug-

gests that initial development is frequently centralised, local and homogenous, primarily developed for a particular community and often small scale, particularly in the numbers served. As others seek solutions to similar problems technology transfer occurs, for example, electricity systems are transferred and taken up in other cities or countries, or technologies are transferred from one domain to another. It is at this stage that both innovation and disruption are most likely to occur. Those drawing on existing technologies and systems are likely to have divergent practices and requirements that are particular to the social, cultural, political and financial conditions under which they operate and the act of transfer inevitably leads to change, adaption, and competition; new innovations are likely but so are competing and incompatible systems.

Maturing systems move to a consolidation stage where one of two scenarios is reached: the emergence of a single dominant system, or a set of interoperating systems that can form networks such as the power grid or the railway system. The end result is the establishment of a service that is ubiquitous and taken for granted. In reality consolidation into a single dominant structure is rare and more often the result is a decentralised network characterised by coordination rather than control. Others (Callon 1986, Edwards 1998) have argued that a networks model better captures the state of continuous interactivity, testing, and innovation that occurs. The model of historical LTS development has arisen from an analysis of numerous cases from the 19th century onwards and we wish to argue that we can use it to help us to understand the trajectory of digital infrastructure development. Understanding this model and seeking to apply it to our thinking about the development of digital humanities research infrastructures and the technologies that we may wish to include can help to provide a framework in which we can reflect, question, and analyse their evolution.

A Universe of Digital Content

A Special Report on 'Managing Information' in the February 2010 online edition of The Economist[1] argues that information has gone from scarce to superabundant citing studies that estimate the amount of data now being generated at between 5 Exabyte's and 1,200 Exabyte's per annum. The disparity in these two figures is a function of the calculations used – the lower figure expresses only new content, whilst the latter includes projections for multiple duplications of digital content. Whichever figure is used the increase in digital data is astonishing. A significant proportion of this data is research data generated from scientific instruments, including giant telescopes, sensors, and perhaps best known, the Large Hadron Collider, or biological reference databases

[1] <http://www.economist.com/node//15557443/>.

such as Genome which organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations, or the Protein Data Bank (PDB) archive which is the single worldwide repository of information about the 3D structures of large biological molecules, both of which were formed as a consequence of collaboration across the domain. It would not be too far-fetched to argue that the availability of this data for processing and analysis is transforming scientific understanding, leading to new discoveries, and raising important new research questions about our universe and human life. Similarly, in addition to the more traditional survey and qualitative data, social scientists are increasingly turning to data generated by social media sites, retail and business transactions as the source material for their research leading to new kinds of social and economic research and methods such as webometrics (Thelwall 2009).

By contrast the humanities do not, and are unlikely to produce large volumes of digital data equivalent to the Large Hadron Collider. Instead humanities research data tends to be highly fragmented across scholarly online publications, smaller web sites and larger repositories in libraries, archives, museums, galleries, publishers and the commercial sector. Neither have the humanities managed to get the support to produce and sustain reference datasets (Perseus being one such example from the humanities) such as that provided by the legislation that established the National Center for Biotechnology Information (NCBI) as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) to host and support the Genome. But despite these differences in scale, there is a significant and growing corpus of digital content available for scholarly research: digital content initiatives such as Europeana[2] are opening up access to the mass of digital objects created by cultural heritage institutions across Europe; the Library of Congress American Memory programme provides free and open access through the Internet to written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music that document the American experience; there are numerous digitisation programmes funded at national level such as the JISC digitisation programme in the UK[3] or the TELDAP[4] programme in Taiwan, and many more smaller scale digital library, digital archive and digital scholarly publications that together form a significant resource for research across the humanities disciplines. A good argument might also be made that much of the content on the web and that generated by social media are also of value for understanding humanity.

More worrying than the volume of digital content is the creeping move towards the commodification of content, locking down behind pay walls the

[2] <http://www.europeana.eu/portal/>.
[3] <http://www.jisc.ac.uk/whatwedo/topics/digitisation.aspx>.
[4] <http://teldap.tw/en/>.

digital content that should form a key resource for humanities scholarship (Prescott 2012). For example, the British Library has partnered with a commercial company to digitise their newspaper archives locking the content away behind a pay wall, and Chadwyck Healey's Early English Books Online is only available for scholarly use through an institutional subscription, and even then most of the content is in PDF form. A frequently-quoted example of a large-scale digitisation effort that falls short of scholarly standards is Google's attempt to scan and make available Tristram Shandy (Duguid 2007). Whole pages were left out because they were considered to be misprints even though they were part of the original composition of the novel because Google did not consider it necessary to ask for more input from researchers and scholars in the field. These problems with quality and format limit the ability of the scholar to do much more than find and read, thus reducing her to the role of passive consumer and restricting opportunities for innovation and the application of computational methods and tools.

This creeping commercialisation of the digital life blood of the humanities seems to attract little opposition or protest; compared to the sciences where the fight for open data is highly vocal and ongoing (Editorial, Nature Genetics, 2012) the humanities barely raise a murmur as the commercial sector hides away digital sources behind pay walls, or strips away their options for new forms of research by virtue of providing low quality digital content in inappropriate formats. What might the reasons for this be? In comparison to the sciences and the social sciences, the humanities have yet to experiment to any meaningful extent with the transformative potential of large volumes of digital content and the application of new methods, nor have they identified the new research questions that might result from such experimentation. Humanists, even those who consider themselves of the digital persuasion, tend to be conservative in their application of technology and seemingly unwilling, except in a few exceptional cases, to put their heads above the parapet and explore and defend different forms of research practice. And even where the digital humanities has experimented with new ideas and the application of digital methods it has, by and large, failed to penetrate mainstream humanities scholarship to any substantial extent (Juola 2008, Prescott 2012). Even worse, the digital humanities and those working within it are too often seen as "… a production house, a place where the infrastructural work of digitization, marking-up texts, and producing tools to facilitate research gets done" (Trettien 2010) rather than as a space and a community which takes up the political and cultural mantle to protect the right of humanities scholars to their sources materials, and where experimentation evolves into new ideas, questions, and theories.

Confronted with the triple challenge of increasing volumes of digital content, some of which is locked away or inadequate for scholarly research; the failure of the digital humanities to fight to keep their content open; and the perceived failure of the digital humanities to fulfil the promise of transforming

humanities research practices where might we look for a way forward? We wish to argue here that instead of regarding itself as the poor relation to the sciences, picking up the crumbs left on the table, the humanities needs to stand up and make a case for a 'big' humanities that seeks to experiment, interpret and interact with large volumes of content, and that needs large research infrastructures that enable and support this work.

Defining Research Infrastructures

> When dealing with infrastructures we need to look to look to the whole array of organisational norms, practices, and institutions that accompany, make possible, and inflect the development of new technology. (Bowker 2010)

The term e-Science was created in 1999 by John Taylor, the then Director General of the United Kingdom's Office of Science and Technology. Taylor saw that many areas of science were increasingly collaborative, multidisciplinary, and working with and sharing large data volumes. What was required, Taylor argued, was a funding programme to support these new forms of research, including the necessary infrastructure components – the UK e-Science programme was born. Launched in 2001 the programme was intended to support both the development of a coordinated, shared, core infrastructure and the application of e-Science methods to research. The e-Science programme was largely technology, data and application driven assuming that the "enormous and growing capacity of computing, storage, communication and software systems – offered the opportunity not only to automate science but also to apply new methods that could revolutionise how science was performed" (Atkins et al. 2009). In this context research was done "through distributed global collaborations enabled by the internet, using very large data collections, terrascale computing resources and high performance visualisation" (Atkins et al. 2009).

The core infrastructure that arose from the e-Science programme was made up of a number of elements and technologies including data (and the curation and preservation activities that supported its creation, use, re-use and sustainability); compute network and data storage; search and navigation tools; virtual research environments; and software solutions for authorisation and authentication, middleware, and digital rights management (Pothen 2007). Key technologies included the Grid which according to Hey and Trefethan is "the infrastructure which will provide us with the ability to dynamically link together resources as an ensemble to support the execution of large-scale, resource intensive, and distributed applications" (Hey and Trefethan 2003). As the Grid has matured a number of standard technologies and web-services have emerged to support the deployment and use of the Grid.

The UK e-Science Programme was followed by a 2003 report from the National Science Foundation on Cyberinfastructure addressing the infrastructure needs of the sciences in the US, and in 2006 by the report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences'Our Cultural Commonwealth' investigating the need for infrastructure for the humanities. Interestingly the Chair of the Commission, John Unsworth, borrowed the definition of Cyberinfrastructure from the 2003 NSF report:

> The 2003 National Science Foundation report Revolutionizing Science and Engineering through Cyberinfrastructure … described Cyberinfrastructure as a 'layer of enabling hardware, algorithms, software, communications, institutions, and personnel' that lies between a layer of 'base technologies ... the integrated electro-optical components of computation, storage, and communication' and a layer of 'software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.' In other words, for the Atkins report (and for this one), Cyberinfrastructure is more than a tangible network and means of storage in digitized form, and it is not only discipline-specific software applications and project-specific data collections. It is also the more intangible layer of expertise and the best practices, standards, tools, collections and collaborative environments that can be broadly shared across communities of inquiry. (Unsworth, Our Cultural Commonwealth 2006)

2006 also saw the publication, subsequently updated in 2010, of the European Strategy Forum on Research Infrastructures Roadmap with a very similar definition.

All of these definitions identify the key characteristics as a mix of hardware, software, instrumentation, digital content, data, and archives, together with human resources, knowledge and expertise that are to be *shared* among communities of practice, and that are essentially collaborative in nature and form. At their core is the idea of *collaboration* and *sharing* between and across communities – whether sharing research data, compute power or other resources – in order to enable new forms of enquiry, and the generation and understanding of new research questions. This idea of an infrastructure based on sharing in the scientific community was at the heart of the various international e-Science programmes.

The attempts to build science Grids have recently evolved into developing open science Clouds. Hey defines the Cloud as "... the ecosystem of technologies that enable the hosting of an organisation's or individual's ICT infrastructure (hardware and software) in large data centres managed by service providers" (Hey 2010). Clouds are also used to store, manage and analyse digital content, including curation and preservation services, and to host advanced services for data analysis, data indexing, metadata extraction and so on. (Gray et al. 2005). Clouds offer the promise of seamless access to resources and services; much as we take for granted (at least in the developed world) that elec-

tric power is available at the flick of a switch, so the Cloud promises that big structures will be similarly available. Researchers will just plug in to a Cloud, which will provide storage or computation on demand. Cloud Computing is part of the larger domain of Utility Computing and if more resources are needed they are available at the click of a button.

But despite Unsworth's use of the National Science Foundation Cyberinfrastructure definition how applicable is the e-Science paradigm to the humanities? Whilst a case can be made that the humanities has its own data (digital content) deluge which is compounded by the 'complexity deluge' (Anderson et al. 2010) inherent in the highly dispersed, multiple format, multiple media, and often highly idiosyncratic nature of the digital content, the common conception of humanities research work is that it is hermeneutic rather than experimental, rooted in narrative, rhetoric and text; that it does not seek formal laws and explanations but rather is essentially interpretive, recursive and questioning, its practices located in the deep reading and reasoning of sources. This conception of humanities research work is poetically described by Andrew Prescott who, in his inaugural lecture at King's College London in 2012, claimed that "scientists want to map the Universe; humanities scholars want to map the universe in a single poem".[5] The single poem does not require 'big data' methods.

This debate is not new. Writing in 1993 Mark Olsen argued that the reason computer aided literature studies had failed to have a significant impact on the field was because they asked only traditional questions of traditional texts and so:

> … have failed to move from a curiosity to an important and respected position in these disciplines. By contrast, quantitative social, political and economic history used computer technology to ask new questions and to develop new methods. Indeed, the computer fits nicely into a shift away from political and event based history, to the history of the social phenomena and the long term, la longue durée. (Olsen 1993)

The problem, Olsen suggests, is the failure to critically engage with theory and to locate the use of technology and data selection in a theoretical framework that would encourage "research design that exploits the strongest points of computer technology, the high speed access and analysis of large amounts of data" (Olsen 1993).

In 2005 Franco Moretti threw out a similar challenge to literary scholars inviting them to look beyond the small (around two hundred he suggested) number of literary works that most scholars work on throughout their lifetimes to consider the vast number of published works within which this literary canon sits. He argued that traditional methods of 'close reading' where the focus is on 'this word and this sentence' was stifling an understanding of the 'collective system' that is the universe of literature: "... a field this large cannot be under-

[5] Available at <http://digitalriffs.blogspot.co.uk/>.

stood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases" (Moretti 2005). He suggested instead a new quantitative approach, a new method that used graphs from quantitative history, maps from geography, and trees from evolutionary theory. He called this approach 'distant reading'.

Olsen and Moretti both make a compelling case for the use of 'big' data and quantitative and graphical methods of exploration and analysis. As we have argued above the humanities now have at their disposal significant quantities of digital information but for that big technological structures are required.

Transferring: e-Science Infrastructures and the Humanities

> *Role for arts and humanities:* Encourage and support even more participation of the arts and humanities research communities in the e-Science Programme (we saw some excellent beginnings in our review). Arts and humanities are poised to achieve large benefit from e-science methods and infrastructure as the human record becomes increasingly digitised and multimedia. (Recommendations, e-Science Review 2009)

Funded by the Arts and Humanities Research Council (AHRC), the Engineering and Physical Sciences Research Council (EPSRC), and the Joint Information Systems Committee (JISC), and managed by the AHRC ICT in Research Programme, the UK Arts and Humanities e-Science programme[6] investigated the transfer and use of large e-Science structures for arts and humanities research. At the beginning, it seemed likely that the use of these technologies would be mainly for the processing and integration of different types of humanities content. However, it soon became clear that the challenges of the underlying semantics made it very difficult to sensibly use the then existing e-Science technologies in the field (Blanke 2011); rather it was high-performance computing and the application of data analytics that proved a better fit to enhance humanities research.

The UK Arts and Humanities e-Science programme awarded grants to undertake humanities research using large e-Science structures and technologies. The projects covered a wide range of subjects in both the arts and the humanities, from dance and music to museum studies, archaeology, classics and Byzantine history, and employed a wide range of e-Science technologies.

For example, the e-Dance project[7] used Access Grid video conferencing technologies, motion tracking and other digital tools to facilitate interactive, multimedia, distributed performance, staged in more than one venue and employing a variety of traditional and digital means of expression. The use of the

[6] <http://www.ahrcict.rdg.ac.uk/activities/e-science/>.
[7] <http://projects.kmi.open.ac.uk/e-dance/welcome/>.

same tools and infrastructure was investigated for documentation of dance performances and more generally to support practice-led research in this area. The project found the use of Access Grid technologies challenging not least in synchronising distributed performance. However, these challenges prompted the investigators to incorporate the Access Grid features of delays and disturbances into their performance thus leading to the exploration of new creative practices.

The Medieval Warfare on the Grid project (MWGrid)[8] employed e-Science methods and tools to support historical research into logistics of medieval war. The battle of Manzikert (modern Malazgirt, Turkey) in 1071, between the Byzantine Empire and the Seljuk Turks, was the subject of this investigation which involved designing and building an agent-based model of this battle. Using agent-based modelling and distributed simulation the project explored military behaviour and interaction, the organisation and mobility of troops and provision required. The project sought to build infrastructure for the execution of very large multi-agent models of military logistical operations on distributed-memory parallel machines, such as those available on a computational Grid. The software engineering process to build and execute models of the kind required by MWGrid proved significantly challenging and involved a high level of collaboration between the humanities researchers and the computing scientists. This level of complexity indicated that the use of these technologies would be of value only to the limited few who could command this kind of support.

King's College London's LaQuAT (Linking and Querying of Ancient Texts)[9] project explored issues of integration and diversity of representation when information is gathered in research databases from different domains and for different purposes in the humanities. LaQuAT used OMII-UK's OGSA-DAI software (Jackson 2009), a de facto standard in e-Science for integrating heterogeneous databases. While technically successful, the project raised important questions regarding the ability of researchers to investigate such integrated datasets. Running queries across datasets required a great deal of understanding about the semantics of the data at a fine-grained level. These semantics were for the most part left implicit in the underlying databases, and LaQuAT concluded that integrating humanities research material was more problematic than initially envisaged and will require researchers to make the connections themselves, including decisions on how they are expressed and how to understand and explore the data more effectively. Data integration in the humanities therefore requires larger structures that join up technology possibilities with human interaction to realise the potential of the technology.

[8] <http://www.cs.bham.ac.uk/research/projects/mwgrid/>.
[9] <http://laquat.cerch.kcl.ac.uk/>.

While data integration remains difficult, the arts and humanities e-Science experiments have shown that other big structures such as high performance compute clusters can solve specific and exceptional problems with humanities data. An e-Science experiment at King's was concerned with enabling connections between humanities data sets using predictive technologies. In the HiTHeR (High-Throughput Computing for Humanities e-Research)[10] project, it could be shown how the computational needs for document analysis in Humanities can be served using clusters of high-performance computing machines (Blanke 2011). For this project HiTHeR worked with the Nineteenth-Century Serials Edition (NCSE) collection in the UK, a corpus containing circa 430,000 articles that originally appeared in approximately 3,500 issues of six 19th Century periodicals.[11] The NCSE is a free, online scholarly edition of nineteenth-century periodicals and newspapers created as a collaboration between Birkbeck College, University of London, King's College London, the British Library and Olive Software, and was funded from January 2005 to December 2007 by the Arts and Humanities Research Council in the UK.

Published over a span of 84 years, materials within the corpus exist in numbered editions, and include supplements, wrapper materials and visual elements. Using high-performance computing, HiTHeR was able to create a browsing interface, for which articles in the NCSE are related by the content they have in common. This is a typical classification task known from many information retrieval and text mining applications. The challenge is that on a stand-alone server our benchmarks indicated that a complete set of comparisons for the NCSE corpus would take more than 1,000 years. We therefore developed a high-performance computing infrastructure at King's College to deliver such tasks in a reasonable amount of time.

A recent workshop led by Geoffrey Rockwell also found that HPC had utility for the humanities. In the report of the workshop Rockwell suggests:

> There is a gap between research in the Humanities and Canadian high-performance computing (HPC) facilities, but it is not what we thought it was. We used to think humanists didn't need supercomputing – they were happy with a wordprocessor, email and the Web. Now it is clear that humanists have large multimedia datasets and big questions to ask of the history of human culture. Then we used to think the gap was primarily between facilities set up for queued batch programs and practices in the Humanities of asking questions repeatedly of 'always-on' web services. Though there is still some truth to that gap, many HPC facilities have begun to support 'portal' or 'cloud' facilities that are always-on and can thus support Humanities practices. The gap now is really one of research culture and support. On the one hand we have to find ways of training and preparing humanities research teams to be able to imagine using existing HPC facilities, and on the other we have to develop the

[10] <http://www.arts-humanities.net/projects/high_throughput_humanities_e_research_hither>.
[11] <http://www.ncse.ac.uk/index.html>.

> ability of HPC consortia to be able to reach out and support humanists. (Rockwell 2010)

If we want to understand the use of big structures in the humanities, we will therefore have to move beyond technologies and also beyond the simple reuse of existing e-Science applications. Rather, we need to consider the complex challenges from the processing of big data in such a way that humans can make sense of it, and the nature of the trust collaborations and understandings that will be essential to our work.

We are at the very beginning of understanding what a humanities research infrastructure is and could be, and what technologies are best suited for supporting and enhancing our research practices and processes. We may be minded therefore, to ask wider questions of the relationship of humanities communities to the digital space and to the big structures within it: what are the tensions and contested areas that are emerging as infrastructure becomes digital and scholars increasingly engage with big data questions and methods? What does it mean for humanities research practices, for the relationships and collaborations, and the norms, values and accepted conventions that bind individuals together in shared communities of practice? What is to be gained from working with and through these new research infrastructures and equally important, what might be lost?

The Digging into Data Challenge research programme was established to address how "big data" changes the research landscape for the humanities and social sciences now that large volumes of digital content are available to scholars in the humanities and social sciences, and what new, computationally based research methods might be applied.[12] Dan Cohen, Tim Hitchcock and Geoffrey Rockwell received funding under the 2009 call for proposals to bring together three online resources: the Old Bailey Online, Zotero and TAPoR to experiment with the application of data mining and statistical analysis to a large corpus of complex texts and information (127 million words of trial accounts) using analytical tools from TAPoR like Voyeur, information management tools like Zotero, and the Canadian HPC facilities. The project has resulted in an infrastructure that allows users to engage with the Old Bailey Online using these tools[13].

The White Paper (Cohen et al. 2011) written by the project team demonstrates the success of the approach and provides examples of new insights into the data produced by the application of data analytical techniques and tools. Clearly the use of 'big data' techniques and tools (and the use of high performance computing) has added a layer of understanding, new questions, and interesting insights that would not be possible for humans alone to achieve. However, at a seminar at King's College London in February 2012 Hitchcock

[12] <http://www.diggingintodata.org/>.
[13] <http://criminalintent.org/>.

reflected on this work and expressed concern at what he perceived as the tension between the original vision behind the construction of the Old Bailey Online and the use of 'big data' methods. He explained that Old Bailey Online was driven by an intellectual and political agenda to understand the underclass and to give a voice to the unheard. Despite the success of the data mining project and the useful results it provided, Hitchcock suggested that the application of data analytics fundamentally damages the relationship with the underclass and removes the voice of the individual. The danger with big data and data analytics, Hitchcock argued, is that we lose the power of individual stories and narratives rooted in the personal to impersonal and positivist statistics and graphs.

Hitchcock's concerns return us to the debate between 'big' humanities and 'small' humanities, and between the lure of close reading and the individual story against the perceived sterility of distant reading and big data analysis. Olsen, however, confronts this issue head on and urges his readers to think of this not as one against the other, but to recognise the value of both; the one to give us insight into a single text or a single individual, the other to give us the wider societal context in which the single text or individual exists. He argues that texts are amenable to both qualitative and quantitative analysis and to combine both sets the individual narrative within and alongside an understanding of the wider context in which that narrative is situated. This combination, he suggests, can provide startling insights that would not otherwise have surfaced (Olsen 1993).

In the March 2012 issue of *Perspectives on History*, a publication of the American History Association (AHA), the Executive Director of the AHA, James Grossman, argued that historians must and should engage with what he termed big history and big data:

> Whether or not we have a facility with numbers, we are good at asking questions and analyzing evidence that by its nature generates many variables at once. And because we look for stories – for ways of synthesizing diverse strands into narrative themes – we usually look for interactions among variables that to other eyes might not seem related. By casting our insights into the form of narratives, we also make them more accessible than multivariate regression analyses could ever be – and arguably more amenable to uncertainty and ambiguity. (Grossman 2012)

Grossman suggests that as well as collaboration between historian and computer scientist historians would do well to collaborate with statisticians. What Grossman is arguing for is an interaction *between* the scholarship of big history and big data and the narrative form that can mesh together the insights arising from each. But perhaps even more important is to recognise that the humanities can bring to bear its own methods of enquiry to humanise the use of big data and big structures and the outputs that emerge from their use.

Except for a small minority the humanities do not have a tradition of dealing with machine algorithms, with the graphs produced from statistical analysis,

and the maps, trees and other forms of visual representation that arise from big data analysis. As a consequence many are wary about engaging with what is seen as a scientific paradigm based on reason and objectivity that runs counter to the epistemology of the humanities. The solution, Drucker argues, is to regard machine algorithms and the visual forms produced from big data analysis as interpretive objects in and of themselves, to think of them as visual signs on a flat surface that require the application of the hermeneutic method so that "the forms that are generally used for the presentation of information can be understood and read as culturally coded expressions of knowledge with their own epistemological assumptions and historical lineage" (Drucker 2010). This bringing to bear of a humanities sensibility – to create meaning from the patterns, to interpret the algorithms, and to foreground the complexity and uncertainty to be found in the visual expressions arising from big data analysis – can surely serve to entwine big data methods and outputs with the more familiar methods and outputs from hermeneutic enquiry.

Big data and the big structures required to use it are, we would argue, a key element of digital scholarship and should take their rightful place in the evolution of humanities research infrastructures, but only as long as we remain mindful at all times of the particularities of humanities research (and the differences between humanities disciplines) and seek to question our assumptions and practices. As Grossman argues if we wish to employ big data techniques and technologies in our scholarship we must seek to understand the implications of our work and to develop different forms of large structures based on communities and collaboration that enable us to contextualise and question what is we are doing, how we are doing it, and why.

Consolidation? The Move to Digital Eco-Systems

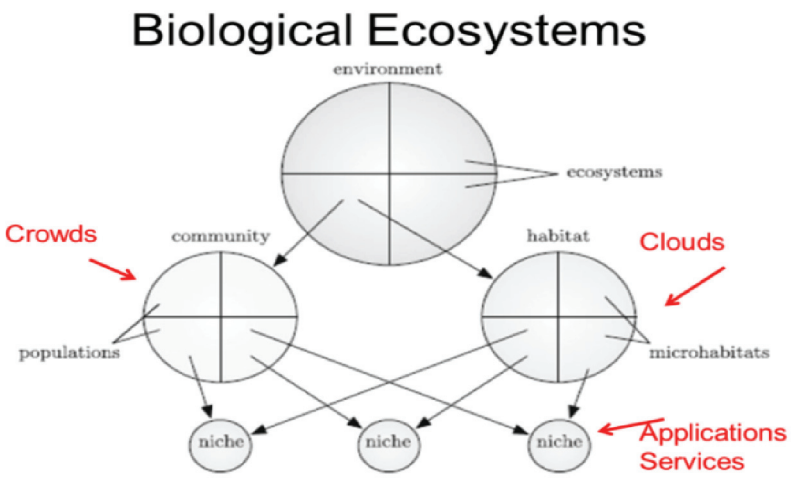
> Digital Ecosystems transcend the traditional, rigorously defined, collaborative environments from centralised, distributed or hybrid models into an open, flexible, domain cluster, demand-driven, interactive environment. (Boley and Chang 2007)

The metaphor of the digital ecosystem is taken from the biological world in order to explain the intrinsic interaction between communities and computing platforms. A natural environment consists of ecosystems, which in turn have habitats and communities inhabiting them. The biological derivation of digital ecosystem, however, only takes us so far. As with many concepts in computing, the uses of the concept determine better what it is about. Here, digital ecosystems are an emerging new concept of infrastructures that recognises the need for a flexible combination of humans, machines, content and things to work together on a common task. In the call for papers for the inaugural IEEE International Digital Ecosystems Technologies conference digital ecosystems are defined as "agents-based, loosely coupled, domain-specific [...] communi-

ties which offer cost-effective digital services and value-creating activities" where value is created by "making connections through collective intelligence and promoting collaboration"[14].

In this definition, digital ecosystems are derived from communities rather than technologies. As open systems, digital research ecosystems will rely on communities and community involvement in a scenario where anyone can participate. The digital ecosystem is not for the specialist few but is instead about increased participation, sharing and building a social network of people, things, content and so on. Here is where the original comparison with biological ecosystems makes sense. Looking at the relationship between components of the digital ecosystem and the biological ecosystem in Figure 1 (taken from Briscoe et al. 2011), we can easily map communities in the biological ecosystem to domain research crowds while the biological habitats are the digital platforms our research crowds work on. Together communities (research domains) and habitats (platforms) build niches which for digital ecosystems are applications and services.

Figure 1: From Biological Ecosystems to Digital Ecosystems



To us digital ecosystems represent best how big structures for humanities research will look like. This can be explained using an example from our ongoing current research on the European Holocaust Research Infrastructure (EHRI).[15] EHRI serves the Holocaust research community with a platform (habitat) that

[14] <http://www.ieee-dest.curtin.edu.au/2007/060607%20-%20Call%20for%20Papers.pdf>.
[15] <http://www.ehri-project.eu>.

enables the integration of Holocaust material. It provides online access and integration of dispersed sources from archives and libraries relating to the Holocaust, and by encouraging collaborative research through the development of tools. The digital Holocaust research communities and platforms are enabled by the larger DARIAH digital ecosystem[16] dedicated to digital research in the arts and humanities, as described in (Blanke et al. 2011b).

Some of the largest digital data sets and computational infrastructures for humanities are linked to preserving the memory of the Holocaust (Unsworth 2006). Some are held in central copy archives such as the ones at US Holocaust Memorial or YadVashem in Israel. But, most of data does not come from such central observatories but from many smaller archives distributed across Europe and the rest of the world. The aim of EHRI is to bring these datasets together into a unified observatory: a Cloud of Holocaust research material. Research communities can then build their own views on these data sets and accumulate them for their research 'niches'. The Holocaust research community is not homogeneous but split up into divergent research; research into victims' materials or testimonials stands next to more traditional archival research into perpetrators. This is a digital ecosystem because it has emerged from the community both by involving the researchers in the selection and description of the material, and by amending archival metadata with research specific information to enhance its representation. By providing platforms around such communities we establish an effective large structure for humanities research, and in the enabled niches for each research interest the long tail of humanities research is addressed as much as the larger scale.

For the humanities, digital ecosystems are where we start to bring together the graphs, maps and trees with the individual narrative, and the universe of digital content with the universe of the poem.

References

[Anonymous]. 2012. 'Your data are not a product', *Nature Genetics*, vol. 44: 357, doi:10.1038/ng.2244.

Anderson, S., T. Blanke, and S. Dunn. 2010. Methodological commons: arts and humanities e-Science fundamentals. *Philosophical Transactions of the Royal Society A* 368 (1925): 3779-96.

Atkins, D. et al. 2010. Report of the International Panel for the 2009 Review of the UK Research Councils e-Science Programme, <http://www.epsrc.ac.uk/SiteCollectionDocuments/Publications/reports/RCUKe-ScienceReviewReport.pdf>.

Blanke, T., and M. Hedges. 2010. Scholarly Primitives: Building institutional infrastructure for humanities e-Science. Future Generation Computer Systems, doi: <10.1016/j.future.2011.06.006>.

[16] <http://www.dariah.eu>.

Blanke, T., M. Bryant, M. Hedges, A. Aschenbrenner, and M. Priddy. 2011. Preparing DARIAH, IEEE International Conference on eScience, 2011: 158-65.

Boley, H., and E. Chang. 2007. Digital Ecosystems: principles and semantics. IEEE International Conference on Digital Ecosystems and Technologies, Cairns, Australia. NRC 48813.

Bowker G., P. Edwards, S. Jackson, and C. Knobel. 2010. The Long Now of Cyberinfrastructure. In *World Wide Research: Reshaping the Sciences and Humanities*, ed. W. H. Dutton and P. W. Jeffreys, 40-4. MIT Press.

Briscoe J., S. Sadedin, and W. De Wilde. 2011. Digital Ecosystems: Ecosystem-Oriented Architectures. *Natural Computing* 10 (3): 1143-94.

Callon, M. 1986. Society in the Making: the study of technology as tool for sociological analysis. In *The Social Construction of Technological Systems*, ed. W. E. Bijker, T. Pinch and T. P. Hughes. MIT Press.

Chang, E., and E. Damiani. 2006 IEEE Digital Ecosystems Conference – Call for papers, <http://www.ieee-dest.curtin.edu.au/2007/>.

Cohen, D. et al. 2011. Data Mining with Criminal Intent. Final White Paper. Downloaded from: <http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf>.

Drucker, J. 2010 Graphesis: Visual Knowledge Production and Representation. *Poetess Archive Journal* 2 (1): 1-50.

Duguid, P. 2007. Inheritance and loss? A brief survey of Google Books. *First Monday* 12.

Edwards, P. 1998. Y2K: Millennial Reflections on Computers as Infrastructure. *History and Technology* 15: 7-29.

Edwards, P., S. Jackson, G. Bowker, and C. Knobel. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" <http://deepblue.lib.umich.edu/handle/2027.42/49353>.

European Strategy Forum on Research Infrastructures. 2011. Strategy Report on Research Infrastructures Roadmap 2010. European Commission <http://ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf>.

Gray, J., A. Lui, D. Szalay, D. DeWitt, and G. Heber. 2005. Scientific data management in the coming decade. *SIGMOD Record* 34 (4): 35-41.

Grossman, J. 2012. Big Data: An Opportunity for Historians? Perspectives on History, American Historical Association, March, 2012, <http://www.historians.org/perspectives/issues/2012/1203/Big-Data_An-Opportunity-for-Historians.cfm>.

Hey, T., R. Barga, and S. Parastatidis. 2010. Research Platforms in the Cloud. In *World Wide Research: Reshaping the Sciences and Humanities*, ed. W. H. Dutton and P. W. Jeffreys, 67-71. MIT Press.

Hey, T., and A. Trefethan. 2003. The data deluge: an e-Science perspective. In *Grid Computing: making the global infrastructure reality*, ed. F. Berman, G. Fox and T. Hey, 809-24. Chichester UK, Wiley.

Hughes, T. 1986. The Seamless Web: Technology, Science, Etcetera, Etcetera. *Social Studies of Science* 16: 281, doi: 10.1177/030631278601600 2004.

Hughes, T. 1993. *Networks of Power: Electrification in Western Society, 1880-1930*, Baltimore, Maryland: John Hopkins University Press.

Jackson, Mike et al. 2009. *Building Bridges between Islands of Data - An Investigation into Distributed Data Management in the Humanities*, 33-9. Oxford.

Juola, P. 2008. Killer applications in the digital humanities. *Literary and Linguistic Computing*, 23 (1): 73-83.
Marx, Leo. 2010. Technology the Emergence of a Hazardous Concept. *Technology and Culture* 51 (3): 561-77.
Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.
Olsen, M. 1993. Signs, Symbols and Discourses: A New Direction for Computer-Aided Literature Studies. *Computers and the Humanities* 27: 309-14.
Pothen, P. 2007. Developing the UK's e-infrastructure for science and innovation. Report of the OSI e-Infrastructure Working Group.
Prescott, A. 2012. Consumers, creators, or commentators? Problems of audience and mission in the digital humanities. *Arts and Humanities in Higher Education* 11: 61, doi: 10.1177/1474022211428215.
Rockwell, G., and M. Meredith-Lobay. 2010. Mind the Gap Draft Report, downloaded from: <https://docs.google.com/View?id=dhbw7427_4hnbkr8cd>.
Thelwall, M. 2009. *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. San Rafael, CA: Morgan & Claypool.
Trettian, W. 2010. 'Digital Humanities vs the Digital Humanist.' Blog post April 2010, <http://blog.whitneyannetrettien.com/search/label/digital%20humanities>.
Unsworth, J. 2006. Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences <http://www.acls.org/programs/Default.aspx?id=644>.