

## E-Science for the Arts and Humanities: A Discussion Paper

Sheila Anderson, AHRB E-Research Expert Seminar, 28<sup>th</sup> April 2004

### Introduction

The E-science programme has arisen because of the increasing pressure in dealing with the escalating volume of data produced by scientists, both in information gathering exercises and through experiments. Scientists are working with computing scientists to find innovative ways of managing and integrating these data volumes, largely funded through what is known as the E-science programme. The programme is global with funding provided on a national basis. The arguments for the programme are that the data deluge<sup>1</sup> will be impossible to cope with using data management current methodologies.

The e-science programme is investigating tools, technologies and methodologies for automating and semi-automating essential research processes. In particular the programme is addressing:

- Sharing and integrating heterogeneous data resources that are distributed globally
- Sharing computing resources and computational power
- Sharing applications
- Developing new ways of generating knowledge from data, in particular ontologies and the semantic web
- Developing new forms of scholarly communications, in particular the idea of virtual organisations and collaborations

### Data ? Information ? Knowledge

The e-science agenda has evolved around a simple three-layer model, starting with data generation/creation, moving from 'raw' data to produce information, and on to the production of knowledge. The need is therefore to:

Automated data management	Automate storage and organisation of data and information in a globally distributed environment
Automated information generation and management	Tools to annotate data with metadata that describes the interesting features of the data and the storage and organisation of the resulting information
Automated knowledge management	Creating knowledge repositories using tools based upon logical, structural and procedural rules that can identify and express relationships within and between data and information

Through the e-science programme a number of tools (middleware) have already been produced and are being tested at a number of sites and institutions. For example, the

<sup>1</sup> Hey and Trefethen 'The data deluge: and e-science perspective' NEED REFERENCE

Storage Resource Broker (SRB) is a client-server based middleware developed to provide uniform access interface to different types of storage device. It is being tested and further developed as a means of allowing users to manage data storage and replication across a wide range of physical storage system types and locations, whilst providing a single, stable point of access to the data. The Globus Toolkit is developing extensive middleware that enables development of heterogeneous computing environments. The projects aims to enable:

- Distributed supercomputing
- Remote visualisation and virtual environments
- Collaborative environments
- Distributed supercomputing working within collaborative environments

Current work seems to be evolving around the Open Grid Services Architecture (OGSA). Essentially, OGSA brings together the concept of an Open Grid Architecture with Web Services to define a set of implementation and platform-independent protocols and standards that would enable creation, management and exchange of information among entities called Grid Services.

The Semantic Web is an extension of the current Web in which information is given well-defined meaning, enabling computers and people to work in better cooperation. The W3C Semantic Web Activity is working with others to define standards and technologies that allow data on the Web to be defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across applications. The idea is to create an environment where data can be shared and processed by automated tools as well as by people, and to move to a more expressive semantically rich Web, making explicit the particular contextual relationships that are implicit in the current Web. The goal is to enable effective information integration, management and automated services.

### **Digital Libraries / Persistent Archives**

Also crucial will be the establishment of a network of digital libraries and persistent archives undertaking the essential tasks of curating and preserving data, information and knowledge and making available the vast array of data resources.

### **Scholarly Communication**

Two key issues arise in the area of scholarly communication:

- How scholars work together and communicate throughout the research process
- The creation of 'knowledge pathways' identifying and linking different kinds of research output

The first challenge is to utilise tools and technology to not only to facilitate existing forms of communication but also to facilitate new ways of communicating –the creation of virtual collaboratories – that encourage innovation and experimentation.

The second challenge is more about the creation of 'knowledge repositories', linking research output held in different places and under different access conditions, allowing annotation and so on. In this scenario we might see e-prints linked to peer-reviewed publications linked to the research data on which it was based, linked to conference papers, linked to discussion lists etc. etc. Tools would track use and monitor the spread and production of knowledge, thereby linking back in reworking of data, and tracking reading of papers etc.

## Summary

Vast quantities of data from a wide variety of sources, held in wide variety of places will need to be properly managed and made available, annotated with suitable metadata and structured in such a way that encourages use and re-use.

Scholars will want to search for distributed sources of diverse data types to discover, integrate and use resources that meet their needs and can offer them new insights.

They will want to analyse and visualise those data using applications suited to their purposes, and may want access to significant amounts of computational power in order to do so.

They will want not only to access the resources themselves, but also to read papers, publications, conference papers etc. based upon those resources. They may also want to follow the trail to find other related research outputs.

They will want (or perhaps need, rather than want?) to communicate with fellow scholars across the globe in new and different ways, and to collaborate in new and different ways.

## E-research in the arts and humanities

### Some questions:

What might the arts and humanities gain from the e-science agenda? What are its needs? What might it contribute?

### Some issues:

*Data Volume:* Arguably similar pressures exist within the arts and humanities. The volume of data created, whilst not quite as mind-boggling as the volumes coming out of the sciences, is still significant, and growing rapidly. Moreover, the nature of the data created within the arts and humanities tends to be complex and increasingly multi-media bringing together images, text, statistical tables, sound and moving image into an integrated multi-media collection.

*Incomplete and fuzzy data:* Much data arising from the humanities are incomplete and 'fuzzy'.

*Cross-disciplinarity:* Increasingly, scholars within the humanities are using data more traditionally used and created in other disciplines – for example, historians are looking to use images, sound and moving image resources in addition to the more traditional documentary sources. Scholars will need to learn new methods and new applications if they are to successfully engage with these different kinds of materials, and the data will also need to be annotated with the appropriate metadata.

*Practice based research in the arts:* For practice based research the output is the performance, or the artwork. How might not only the final performance or artwork be captured, but also the process of research involved in creating the output. Tools exist in the sciences to capture the research and experiment process – might we re-purpose these for the arts?

*Scholarly communication:* Much work in the arts and humanities is individualistic and rooted within discipline scholarship traditions. The challenge is to retain the best of this way of working whilst enabling new forms of communication and inter-disciplinarity.

*Management of rights and security issues:* Many resources are created or distributed by the commercial sector, or have complicated rights attached to their use, requiring a secure environment for their management and access.

### **What might an Arts and Humanities E-research agenda look like?**

1. Establish AHDS + partner as an arts and humanities test-bed – for example
  - link to the grid using Globus Toolkit and OGSA
  - implement SRB for distributed access and management of resources
  - expand the concept of the knowledge repository by investigating linking all types of research output – in conjunction with RDN hubs
  - use as a test-bed for other initiatives
  - integrate the large storage and transfer demands of the AHDS archive into the GRID
2. Establish a metadata/ontology research and application programme –for example
  - establish a thesaurus management system on the GRID and populate it with existing thesauri addressing the A&H area (e.g. Getty Art & Architecture thesaurus) so that this can be used to index and populate metadata entries for the field
  - integrate thesauri and topic maps with those from other domains and move towards Semantic Web Ontology technologies (DAML-OIL etc)
  - translate existing metadata from the AHDS to a format common to the GRID to allow cross domain searching
3. Set up a network of scholars to share expertise, knowledge, methods etc. and to encourage/investigate new forms of scholarly communications – include computing and information scientists in this network – through Methods Centre?
4. Create an authentication and authorisation tool using web services technologies on top of the existing Globus layers, and a User Profile that will be robust, scalable, link to both the GRID and web services, and meet the needs of the AHDS community for User Profiling to support Rights Access – seamless access to commercial resources
5. Establish/fund 2-3 exemplar research projects using the cross-domain knowledge, thesaurus and metadata mechanisms from A&H research teams in order to demonstrate the effectiveness of the infrastructure established
6. Investigate automatic capture of research processes in practice based arts.
7. Present the expertise of the A&H archive in establishing collections for the purpose of research to other disciplines using the GRID