

**Report on the Projects and Tools for the e-Science Scoping Study**  
**Luke Blaxill, Project Research Assistant**  
**11/08/06**

**Background**

The e-Science Scoping Study (henceforth the 'eSSS') has investigated several hundred e-science projects and tools, mainly from the hard and social science domains. From these, what we considered to be the 100 most promising and exciting were investigated further and summarised. This research is attached to this report as two appendices, and will also be deposited in the forthcoming eSSS Knowledge Base.

The projects were investigated principally to examine their methods: how scientists and social scientists are using e-science to enhance research and outputs, even if the subject of their work (which might concern topics such as aircraft maintenance, or the breeding practices of geometrid moths) had little or no relevance to Arts and Humanities researchers. The investigation of the tools used in e-science was a little more direct: it was principally a question of establishing whether existing software tools could be directly exported to Arts and Humanities research, or repurposed or copied in some way.

The majority of the projects investigated by the eSSS are listed on the National e-Science Centre's (NESC) and the National Centre for e-Social Science's (NCeSS) websites<sup>1</sup>. Unfortunately, there is no publicly accessible list of e-science software tools currently available.

**The Aim of this Report**

This document is intended to provide a general commentary on the research undertaken on these e-science methods and tools by the eSSS. It also aims to provide - in light of the discussions at the expert seminars - some general reflections on how these methods and tools might be employed in Arts and Humanities Research. To facilitate its commentary, this report will discuss the methods and tools under the following more general headings. It must be stressed – especially given that all e-

---

<sup>1</sup>NESC: [http://www.nesc.ac.uk/projects/escience\\_projects.html](http://www.nesc.ac.uk/projects/escience_projects.html) and  
NCeSS: <http://www.ncess.ac.uk/research/>

science developments can arguably be classed under a general agenda (as explained elsewhere in this report) – that these headings do not delineate a discrete breakdown of subjects, and that there is inevitably considerable overlap between them. They are:

1. Data Sharing: Data Grids, and Data Repositories
2. Data Management, Integration and Grid Middleware
3. Virtual Organisations and Virtual Research Environments
4. Grid Computing and the Access Grid
5. Visualisation and Visual Analysis
6. Training and Awareness

In investigating almost 300 e-science projects and tools, the eSSS can claim to fairly represent a cross section of what is available, and what is being developed, in the UK e-science community. However, it must be remembered that there were many projects and tools the eSSS did not have time to investigate-- in the UK, and especially overseas.

### **1. Data Sharing: Data Repositories and Data Grids**

There are a great many e-Science and e-Social Science projects which look to share data in a secure networked environment. However, for the purposes of this section these will be divided under the headings of Data Repositories, Data Grids, and Data Sharing Tools.

Data repositories are networked environments where researchers can securely and selectively share sensitive, prepublication data, by depositing it centrally. Data Repositories might be thought of as very streamlined data grids- a simple system which allows the sharing of data, and that is not so concerned with its collaborative analysis, processing, structuring, enhancement, or standardisation, as a full-scale Data Grid might be. The eSSS has examined a number of data repositories<sup>2</sup>. The ‘Grid Enabled Microarray Experiment Profile Search (GEMEPEPS) project allows biological scientists to deposit the results of experiments in the repository, and allows others to

---

<sup>2</sup> See Projects Appendix entries 5, 32, and 36

access this data via a search interface<sup>3</sup>. The obvious benefit of the GEMEPS system is that it can help prevent geographically distributed researchers from doing the same research twice. A similar project is being undertaken in the social sciences, where a central repository of Metadata is being developed via a system of voluntary depositary, in order to help social scientists make the most effective use of data held in existing archives<sup>4</sup>.

The expert seminars confirmed that this method could potentially be very useful to arts and humanities scholars. Historians, and researchers from the domain of Literary and Textual Studies, for example, are prolific list makers-- and build up huge bodies of potentially very useful data which never sees publication. In the present scholarly environment, this work will often be done many times over by distributed researchers, as wheels are continually re-invented.

Although exciting in principle, it must be stressed that - given that Arts and Humanities data is often rather less standardised than scientific data - it's re-use by scholars who did not create it may be slightly problematic. There is the option to enforce standards for data input, but this would arguably be a strong deterrent to the voluntary submissions on which a repository relies. Also, there is naturally the issue that, in the arts and humanities- where there is not so much of a collaborative culture - many researchers like to keep close control of their own research and who accesses it. Overall, the current culture and complexity of much arts and humanities research militates against the data repository model, but their potential is still considerable. They are also quite simple and cheap to set up- two depositaries have been set up in the social sciences for very modest sums.<sup>5</sup>

On a simpler level, there was a clear appetite at the seminars for an efficient way of sharing large files. The Lionshare tool is a peer-to-peer networker in the style of the popular Bittorrent tool<sup>6</sup> that offers a secure, authenticated environment, and servers on which files can be stored for common use. It also includes data extraction and analysis capabilities<sup>7</sup>.

---

<sup>3</sup> *ibid*, entry 36

<sup>4</sup> *ibid*, entry 5

<sup>5</sup> *ibid*, entries 5 and 32

<sup>6</sup> <http://www.bittorrent.com/index.html>

<sup>7</sup> See Tools Appendix entry 39

The Data Grids the eSSS has investigated are rather more ambitious than the repositories<sup>8</sup>. Data Grids – despite their name - are quite often concerned with linking distributed expertise and knowledge as well as existing data. Some data grids such as the Distributed Aircraft Maintenance Environment (DAME)<sup>9</sup> and the Astrogrid<sup>10</sup> also offer grid computing facilities that enable the analysis of their data and its visualisation, within the same environment. However, other grids such as the NERC Data Grid, Discovery.net, and the Grid Enabled Occupational Data Environment (GEODE), are concerned specifically with data<sup>11</sup>. In general terms, these data grids allow participating researchers and institutions to plan, manage, share and execute complex knowledge discovery and data analysis procedures, supplied as remote services. They also allow the publication and distribution of information gathering and analysis tools as services, and allow data owners to provide access to data in the style of data repositories.

The eSSS has investigated a variety of data grids that concern a myriad of different subjects, but they all tend to share these common features. Of particular interest are the NERC Datagrid, and the Discovery.net project, as both are trying to pioneer an effective and generic model for a datagrid-- to enable them to be more easily set up in other fields. Discovery.net is creating what it describes as a ‘service-orientated computing model for knowledge discovery’<sup>12</sup> while the NERC project is more concerned with the creation of generic data and metadata standards for all data grids. It is also attempting to formulate a model through which the management of such data and metadata could be increasingly automated.

Naturally, if data grids involve numerous collaborators from different sectors and disciplines, specialist knowledge of the data cannot be assumed, hence the need for common standards. Once again though, it may be that the scientific data NERC and Discovery.net are using is more amenable to the models and standards they are creating, although both are keen to stress their general application. However, these models arguably merit further investigation , as they could potentially considerably facilitate the creation of data grids in the arts and humanities.

---

<sup>8</sup> *ibid*, entries 2, 3, 6, 10, 12, 15, 19, 20, 23, 32, 40, 46 and 52.

<sup>9</sup> *ibid* entry 40

<sup>10</sup> *ibid*, entry 46

<sup>11</sup> *ibid*, entries 19,20, and 32

<sup>12</sup> <http://www.discovery.net/> See *ibid*, entry 19

## 2. Data Management, Integration and Grid Middleware

The eSSS has investigated a number of projects and tools which concern the structuring and integration of datasets from various sources. Given that much data in the Arts and Humanities has been created at different times, to different standards, and by different institutions, a key challenge is to create systems and software that can retroactively integrate and unify this data to common formats and standards in order to make it usable and intelligible within a virtual environment such as a Data Grid or VRE.

The eSSS has investigated two particularly interesting data structuring projects which use data integration tools. The first is BiodiversityWorld which is a bioinformatics cataloguing system which - using the powerful Triana tool to compose web services - draws together data from a large number of heterogeneous databases from different institutions concerning insect and plant life, in order to create a Problem Solving Environment<sup>13</sup>. The second is the Axiopé Project, which – by using a universal management system and catalogue called Catalyser - is creating a mixed media catalogue in the Neuroinformatics domain (again, with data derived from heterogeneous and distributed datasets). This catalogue will include text, images, and visualisations<sup>14</sup>. Both Catalyser and Triana are keen to stress that they are user friendly, usable outside the science domain, and can be used for grid based e-science projects and their outputs grid-enabled. While Triana and BiodiversityWorld are more concerned with linking together data held in different existing databases by web service composition, Axiopé and Catalyser are more about creating new and structured data according to a defined (but customisable) system, and both approaches might have some applicability to the Arts and Humanities domain.

Unfortunately, the vast majority of datasets have not been set up using tools such as Catalyser, so grid middleware needs to be as flexible and robust as possible if it is to fulfil its essential role as a communications layer between distributed and heterogeneous data sets (as well as applications and hardware). The Globus Toolkit<sup>15</sup> – which includes a myriad of middleware software tools is an industry staple, and

---

<sup>13</sup> BiodiversityWorld: <http://www.bdworld.org/> See ibid entry 18

Triana: <http://www.trianacode.org/> See Tools Appendix entry 1

<sup>14</sup> Axiopé: <http://www.axiope.com/> See Project Appendix entry 21

Catalyser: <http://www.axiope.com> See Tools Appendix entry 3

<sup>15</sup> <http://www.globus.org/> ibid entry 25

Ogsa-Dai provides a system of querying data via web services deployed in a grided environment<sup>16</sup>. Storage Resource Broker (SRB) is a data management program well adapted for use in the grid environment, which can also help reliably manage distributed data<sup>17</sup>. SRB had been used across a variety of platforms and is currently being implemented by the AHDS. The N/A16 Dataset Integration Project – for example - has successfully integrated 5 major gene databases using grid middleware packages<sup>18</sup>.

The most ambitious and pioneering project that the eSSS has investigated concerning data management and integration is the Spice Project, which is concerned with the idea of creating taxonomic federated architectures to create harmonised data systems conducive to use in e-science<sup>19</sup>. It draws upon a number of discrete databases (that concern biological species information) and plans to create a federated architecture which enables heterogeneous data to be made available under the same set of standards, classifications, and metadata, and that employs taxonomic coverage achieved through wide scholarly agreement. In order to achieve this, the Spice Project will attempt to create (as far as is possible) a master-taxonomy for species data from a wide scale collaborative discussion across the bio-science community- although it acknowledges that this will not be easy- as in the biological science domain (like in the Arts and Humanities) disagreements on classifications are rife.

Overall, the methods and tools investigated by the eSSS have shown that data management and integration is being approached both proactively and retroactively in the e-science community. Classifying data with taxonomies in the style of SPICE (although this would be controversial), and creating standardised datasets with tools like Catalyser, and (most obviously) adhering to standards for mark-up and presentation, can help ensure that new datasets generated are more amenable to future e-science development. However, much Arts and Humanities research involves datasets created in the past which are so heterogeneous that robust middleware to enable their querying via web services arguably represents the best solution.

---

<sup>16</sup> <http://www.ogsadai.org.uk/> ibid entry 23

<sup>17</sup> [http://www.e-science.clrc.ac.uk/web/projects/storage\\_resource\\_broker](http://www.e-science.clrc.ac.uk/web/projects/storage_resource_broker) ibid entry 27

<sup>18</sup> [http://www.nesc.ac.uk/action/projects/project\\_action.cfm?Title=65](http://www.nesc.ac.uk/action/projects/project_action.cfm?Title=65) See Projects Appendix entry 34

<sup>19</sup> <http://www.sp2000.nies.go.jp/> ibid entry 26

### 3. Virtual Organisations and Virtual Research Environments

Virtual Research Environments (VREs) are far simpler than data grids but also seek to provide a framework of resources to support the underlying processes of research, such as access to and analysis of existing resources. The JISC has a major VRE programme, which aims to “develop a common framework and its associated standards and to encourage others to work within this framework to develop and populate VREs with applications, services and resources appropriate to their needs”<sup>20</sup>. Perhaps the best known of the JISC VREs is the Silchester Project which enables those present at an archaeological dig to make new data gathered from the site available online in order to access distributed expertise to exchange ideas and interpretations through the VRE<sup>21</sup>. In essence, it plans to use the VRE to synchronise the processes of gathering information, co-ordinating expertise, and managing bodies of data. The VRE on Political Discourse 1500-1800 is doing something similar in the area of teaching and learning, where the universities of East Anglia and Hull have set up a collaborative course where lectures, seminars, and discussions involving experts and students from both institutions can take place over the VRE<sup>22</sup>.

The nodes involved in VREs, or indeed in Data grids, are usually individuals, but can also be institutions and virtual organisations- or complex collaborations between all three. The eSSS has investigated a number of quite innovative interdisciplinary and multi-institutional projects which involve actors from many branches of the academic, corporate and political worlds, who either work towards a common goal, or trade expertise with each other. The aforementioned DAME project involves the many actors who are involved in maintaining aircraft, for the common goal of safety<sup>23</sup>. The famous e-DiaMoND project brings together medical image experts, computer scientists, and clinicians from all over the country to forward the breast imaging process<sup>24</sup>, and the Biomedical Research and Informatics (BRIDGES) project involves institutional collaborators that have little understanding of each other, but who might nevertheless benefit from the mutual sharing of expertise<sup>25</sup>.

---

<sup>20</sup> [http://www.jisc.ac.uk/programme\\_vre.html](http://www.jisc.ac.uk/programme_vre.html)

<sup>21</sup> <http://www.silchester.reading.ac.uk/vre/> See Project Appendix, entry 49

<sup>22</sup> <http://www.uea.ac.uk/his/research/projects/vre/> See *ibid* entry 51

<sup>23</sup> See *ibid* entry 40

<sup>24</sup> <http://www.ediamond.ox.ac.uk/> See *ibid*, entry 53

<sup>25</sup> <http://www.brc.dcs.gla.ac.uk/projects/bridges/> See *ibid*, entry 27

There has already been a great deal of interest and progress in the domain of VREs in the arts and humanities- thanks in part to the JISC initiative. There are also some interesting projects seeking to encourage and widen access to VREs. Oxford is working on the 'Building a Virtual Research Environment in the Humanities' (The BVREH Project) which aims to investigate the demands and potential uses for VREs in the Arts and Humanities through surveys and expert meetings<sup>26</sup>, and the forthcoming GROWL VRE toolkit will include a user-friendly client interface and software application model to enable VREs to be more easily and generically created<sup>27</sup>.

#### **4. Grid Computing and the Access Grid**

It has often been assumed that the computational grid has little application in the arts and humanities. However, the expert seminars revealed that there was a genuine demand for easily accessible large computational processing power for various tasks- such as the analysis of very large linguistic corpuses, or to empower existing software tools, for example. Naturally, grid computing is just one way of accessing processing power, so a key question is whether it represents the most appropriate way of delivering it for the arts and humanities.

The eSSS has investigated a variety of projects which use the computational grid, or relate specifically to it<sup>28</sup>. The Particle Physics Grid is an example of an enormous multi million pound grid computing infrastructure where 20 institutions are involved<sup>29</sup>, whereas the Scotgrid is much smaller and less expensive- simply involving the universities of Dundee, Edinburgh, and Glasgow<sup>30</sup>. The Scotgrid project sees itself as a pioneer of a small and flexible model for a computational grid, and is conducting a self-assessment study, looking at (in particular) how it's availability has impacted upon the work of researchers. In the same way as the GEODE<sup>31</sup> and

---

<sup>26</sup> <http://bvreh.humanities.ox.ac.uk/> See ibid entry 50

<sup>27</sup> <http://tyne.dl.ac.uk/GROWL/> See ibid entry 47, and Tools appendix entry 30

<sup>28</sup> ibid entries 10, 30, 33, 45, 46

<sup>29</sup> <http://www.gridpp.ac.uk/> See ibid entry 30

<sup>30</sup> <http://www.scotgrid.ac.uk/> See ibid entry 33

<sup>31</sup> <http://www.geode.stir.ac.uk/> See ibid entry 32

DAME<sup>32</sup> projects, the Scotgrid also acts as a datagrid at the same time. It might represent a good study for the arts and humanities - as would the Grid Enabled Mico Econometric Data Analysis Project- a demonstrator which is examining how 'the seamless interfacing of distributed data and computational power via the grid' might be applied to social science research practices- in this case, Micro-Econometric Analysis<sup>33</sup>.

Participants at the expert seminars were particularly interested in the easy use of the computational grid from a desktop environment. The Grid Enabled Desktop Environments (GRENADE) project and toolkit aims to tightly integrate grid computing capabilities within a users desktop environment. Finally, there are a number of existing software tools- such as the Armadillo text mining tool used in the History domain<sup>34</sup>, and the Sabre statistical modelling tool in social and political science<sup>35</sup>, which would benefit from grid enabling that would enable them to analyse greater volumes of data, more rapidly. A working example of this is the SABRE in R project which is using OGSA-DAI to grid enable this tool, in order to attempt analysis of a new scale and scope.<sup>36</sup>

The Access Grid is technically an e-science tool, but is becoming so widespread it might arguably be classified as a methodology in itself<sup>37</sup>. In the Visual and Performing Arts, use of the access grid is already quite advanced, but the eSSS did examine a number of projects which looked at how the technology might be customised for use (especially) in the arts-- such as the Wimbledon School of Art's involvement in the MARCEL group<sup>38</sup>, where they are creating a 'permanent very high bandwidth interactive network dedicated to artistic, educational and cultural experimentation' over the access grid. There is also the 'New Technologies, New Applications' Access Grid study which is investigating the use of access grid nodes in field research and training in the social sciences<sup>39</sup>.

---

<sup>32</sup> <http://www.cs.york.ac.uk/dame> See ibid entry 40

<sup>33</sup> <http://www.ncess.ac.uk/research/pdp/#gameda> See ibid entry 10

<sup>34</sup> <http://www.hrionline.ac.uk/armadillo/index.html> See Tools Appendix entry 40

<sup>35</sup> <http://sabre.lancs.ac.uk/> See ibid, entry 2

<sup>36</sup> <http://www.ncess.ac.uk/research/pdp/#ogsa> See Project Appendix entry 4

<sup>37</sup> <http://www.accessgrid.org/> See ibid entry 8

<sup>38</sup> <http://www.icfar.co.uk/pdf/wimbledon.pdf> See ibid entry 54

<sup>39</sup>

<http://www.accessgrid.surrey.ac.uk/Projects/NewApplicationsNewTechnologies/tabid/55/Default.aspx>  
See ibid entry 8

## 5. Visualisation and Visual Analysis

E-science might have a considerable impact on existing processes in this area as it provides the potential to increasingly distribute visualisations and simulations. Distributed computational power via the grid can be employed to speed up existing generation processes, and can enable more ambitious visualisations to be attempted. The ease at which data can be transferred in real time across the data grid also enables simulations to be updated and altered while a visualisation is running. Therefore, developments fall under two main headings- high performance visualisation, and distributed visualisation and simulation.

The eSSS has investigated a number of projects which relate to the former<sup>40</sup>. The RealityGrid was created for modelling scientific matter, and is using the computational grid to create a 'highly flexible and robust computing infrastructure for supporting visual modelling'<sup>41</sup>. It is attempting to formulate a grid enabled visualisation package that might be exportable to other domains. In the same manner, the Eviz project is attempting to pioneer some general purpose software tools which will manage high performance visualisation automatically, as well as a generic model for a networked supercomputing environment for High Performance Visualisation over the grid<sup>42</sup>. The Jigsaw Project takes a different approach and instead looks at the data that is to be visualised, which – in distributed visualisations – might be incomplete, changing, or anomalous. It is attempting to formulate algorithms that are able to handle poor data streams without disrupting the whole visualisation process<sup>43</sup>.

Distributed visualisation projects - where the underlying simulation can be updated in real time – have been appearing in recent years. These projects are based on a data grid model as they allow for complex multidisciplinary collaborations between distributed professionals and institutions in the style of the DAME and e-DiaMoND projects<sup>44</sup>. The gVIZ project<sup>45</sup> is particularly interesting as it is attempting to grid-enable the existing and popular visualisation system IRIS explorer<sup>46</sup>, and is

---

<sup>40</sup> See *ibid* entries 24, 41, 43, and 44

<sup>41</sup> <http://www.realitygrid.org/> See *ibid* entry 24

<sup>42</sup> <http://www.eviz.org/> See *ibid* entry 44

<sup>43</sup> [http://www.nesc.ac.uk/action/projects/project\\_action.cfm?title=122](http://www.nesc.ac.uk/action/projects/project_action.cfm?title=122) See *ibid* entry 41

<sup>44</sup> See the DAME and e-DiaMoND Projects, pp. 3-4

<sup>45</sup> <http://www.comp.leeds.ac.uk/gviz> See Projects Appendix entry 35

<sup>46</sup> [http://www.nag.co.uk/welcome\\_iec.asp](http://www.nag.co.uk/welcome_iec.asp)

developing a library of computational steering to enable real time modifications of remote simulations from the desktop. In theory, gViz will enable the collaborative generation of a visualisation, and collaborative real time manipulation of the underlying simulation. The Financial Information Grid (FINGRID) is a methodologically similar social science project attempting real time distributed simulations, within an environment where these can be analysed<sup>47</sup>. It is a 24 node grid that enables constantly changing financial figures, market trends, and financial news from different nodes to be simulated and analysed in a central virtual research environment.

Distributed visual and video analysis is a related area which the eSSS has also examined. The VIDGRID project is examining the key difficulties associated with distributed video analysis over the grid, and, from this, is attempting to create a general set of parameters and requirements for work amongst distributed researchers 'in light of the way they handle, catalogue, share, examine, present, and disseminate materials'<sup>48</sup>. A related social science project is also looking at how semantic annotation can be used to enhance distributed video analysis practices<sup>49</sup>.

## **6. Training and Awareness**

A theme that continually resurfaced at the seminars - was that knowledge of e-science potentials and outputs was low in the arts and humanities. In many ways, the social sciences faced a similar problem, and they have set up a number of training programmes that have been noted by the eSSS. The Fastrack training scheme has created training materials aimed at social scientists with little prior knowledge of e-science or the grid<sup>50</sup>. The Training and Awareness Raising Resource Discovery for Researchers (ReDReSS) programme has a similar aim, and has created a discovery portal which includes guides, training materials and exemplars, and is also running a programme of workshops and a roadshow<sup>51</sup>.

---

<sup>47</sup> <http://www.computing.surrey.ac.uk/grid/fingrid/> See ibid entry 2

<sup>48</sup> <http://www.kcl.ac.uk/depsta/pse/mancen/witrg/vidgrid.html> See ibid entry 1

<sup>49</sup> <http://www.ncess.ac.uk/research/small-grants/#semanticannotation> See ibid entry 9

<sup>50</sup> [http://www.jisc.ac.uk/index.cfm?name=project\\_fasttrack&src=alpha](http://www.jisc.ac.uk/index.cfm?name=project_fasttrack&src=alpha) See ibid entry 42

<sup>51</sup> <http://redress.lancs.ac.uk/> See ibid entry 11

One particularly interesting undertaking is the Entangled Data- Knowledge and Community Making in e-Social Science project, which is due to finish by the end of 2006<sup>52</sup>. This is an almost anthropological study of why researchers do or do not collaborate using shared digital data sources. It will investigate groups- some of who already use e-science technology to collaborate, and some who don't - in order to establish more clearly how and how far e-science methodology is useful in the social science domain, and the priority areas for future development.

Luke Blaxill

11/06/08

---

<sup>52</sup> [http://www.nesc.ac.uk/action/projects/project\\_action.cfm?title=259](http://www.nesc.ac.uk/action/projects/project_action.cfm?title=259) See ibid entry 7