Developing an e-Infrastructure for Social Science

Michael Daw[1], Rob Procter[1], Yuwei Lin[1], Terry Hewitt[1], Wei Jie[1], Alex Voss[1], Kenny Baird[1], Andy Turner[2], Mark Birkin[2], Ken Miller[3], William Dutton[4], Marina Jirotka[4], Ralph Schroeder[4], Grace de la Flor[4], Pete Edwards[5], Rob Allan[6], Xiaobo Yang[6], Rob Crouchley[7]

[1]University of Manchester, [2]University of Leeds, [3]University of Essex, [4]University of Oxford, [5]University of Aberdeen, [6]Science and Technology Facilities Council, [7]Lancaster University

Email address of corresponding author: michael.daw@manchester.ac.uk

Abstract. We outline the aims and progress to date of the National Centre for e-Social Science e-Infrastructure project. We examine the challenges faced by the project, namely in ensuring outputs are appropriate to social scientists, managing the transition from research projects to service and embedding software and data within a wider infrastructural framework. We also provide pointers to related work where issues which have ramifications for this and similar initiatives are being addressed.

Introduction

The UK Economic and Social Research Council (ESRC) has funded a three year project (from January 2007) to begin building an e-Infrastructure to provide social scientists integrated access to a variety of research resources currently under development by the National Centre for e-Social Science. This e-Infrastructure is in part based on resources and software deployed in other areas of science through the UK e-Science programme which started in 2001. This paper describes the project, the aims of which are to:

- examine the feasibility of and begin to create a coherent e-Infrastructure for social science in the UK;
- provide a platform for disseminating the benefits of e-Social Science;
- leverage existing e-Social Science and e-Science investments;
- and define a roadmap for the sustainability of middleware, services and tools currently being developed by participants in the UK's National Centre for e-Social Science (NCeSS).

The paper will also focus on what are likely to be the most challenging aspects of the project, and try to answer questions such as:

- how to incorporate user requirements so that project outputs are appropriate, useful and highly usable for social scientists;
- how to move from research outputs to services for the support of social science research;

- and how to incorporate these within the wider UK e-Infrastructure as offered by the UK National Grid Service (NGS), regional services such as North West Grid (NW-GRID) and the large UK data centres such as ESDS and EDINA.

The project outputs fall into three broad categories: “research resources”, “support activities and materials”, and “investigations” into the feasibility of technologies, technical standards, sustainability and project evaluation. These are now briefly described.

Research Resources

Research resources comprise software tools, services, Grid-enabled datasets and the infrastructure on which these will reside. The project stresses the importance of enabling people to access these resources, and to be able to collaborate and interact in a rich variety of ways. The aim of the technical programme is to provide straightforward access to the e-Infrastructure core (compute resources and data), along with applications (services and tools) and demonstrators.

Social scientists today typically have less experience and expertise in the use of the Grid than researchers from disciplines more traditionally associated with large-scale computation and data analysis. Distributed systems are inherently complex and the associated middleware products are not intuitive or easy to use. We have identified several ways of accessing Grid resources and Grid-enabling statistical and other software:

1. Directly via Grid-installed middleware, e.g. using Globus via command-line interfaces or through applications (APIs);
2. Via a Web Portal, e.g., a Social Science Gateway;
3. Via a desktop application using a Web services API in any appropriate language.

The user environment consists of a portal, a service registry and workflow tools. These collectively act as “user front ends” or “clients” to middleware and applications. The portal deployed is based on the Sakai framework (Severance et al. 2007) and offers collections of tools and services appropriate to the needs of different social science virtual organisations (VOs). A VO represents a particular grouping of users with a particular set of requirements. Users will typically belong to more than one VO. Each VO has its own “worksite” on the portal that provides easy access to a collection of applications via the middleware. A number of these tools, such as Wikis, blogs and videoconference applications, will be generic to many VOs but the content managed by the tool is specific to the VO in question. Specialised applications or tools will also be available to meet the needs of any particular VO and we will encourage communities to develop their own by organising training events and providing related learning material. The project is developing these tools to be pluggable into the portal using recognised standards such as JSR-168 or WSRP and using interface technologies such as AJAX and JSF. They will be discoverable via a social science service registry, from which the VO administrator can populate the corresponding portal worksite with the tools appropriate to his/her VO. A goal of the project is to make the portal as usable as possible (a workshop is being organised to asses what is required) and to enable VOs to develop and deploy their own services.

Examples of such tools include applications that offer simulation models to support land-use and health policy-making; tools to support the use of metadata by social scientists for the annotation, querying and presentation of resources; tools to enable the editing of large datasets as well as construction and analysis of complex statistical models. Some of these

tools have already been demonstrated at the International e-Social Science Conference series (Allan, Crouchley and Daw, 2006). In addition to a portal interface we are enhancing the Grid Resources On Workstation Library (GROWL), a collection of services that hides the complexity of the underlying Grid middleware (e.g., Globus or Condor) from the user and provides a secure means of accessing the Grid from plug-ins to other traditional desktop applications such as Stata, SPSS, MatLab, etc. (Crouchley et al. 2005, Grose et al. 2006)[1]. Workflows offer an additional means of employing services and this project is developing sample demonstrations to investigate their use in the social sciences – Taverna, Pegasus and Kepler will initially be investigated. If appropriate, these will underpin use of the portal tools and GROWL to facilitate repeated similar investigations and sharing of complex procedures.

The project also aims to determine appropriate security mechanisms to meet the needs of all aspects of this emerging infrastructure (c.f. Jie et al. 2007, this volume) and to Grid-enable a number of datasets so that they may be subject to more flexible discovery, interrogation and integration ultimately through semantic services.

Further work on portal interfaces has already been completed for the UK National Grid Service portal. Because of the use of the JSR-168 Java portlet standard, this will be of immediate benefit for the e-Infrastructure. For instance, social science applications can be described using the Job Specification Description Language (JSDL) and added to a sharable repository. Applications such as the SABRE statistical analysis code are already accessible in this way.

Support Activities and Materials

To address training needs of social scientists, the project is producing a variety of support materials, including documentation, learning objects for inclusion in learning environments, and training materials, which we aim to deliver through workshops to introduce the tools and infrastructure to prospective users. The NCeSS programme, working in collaboration with the Resource Discovery for Researchers in e-Social Science project[2] (ReDReSS) is already providing a large quantity of relevant material which can be extended. In the UK, we also benefit from the National e-Science Centre's TOE training programme and two JISC-funded community engagement projects (Voss et al. 2007, this volume).

Interoperability and Licensing

In order to remove potential barriers to dissemination and to facilitate description and retrieval of the support materials to the wider community, support materials will adhere to relevant standards and specifications,(e.g., the UK LOM Core[3] or Dublin Core Educational Metadata Standards) and all produced materials will also comply with relevant accessibility guidelines. Thus, these resources will be easy to locate and run on a variety of systems and platforms. We note that the Sakai framework which is being used for the portal was originally designed as a collaborative virtual learning tool, so is well fitted to meet these requirements.

To foster an environment where a community can share support materials, legal barriers must also be overcome. In order to do this, a clear legal framework that acknowledges and protects IPR is needed. Support materials for the e-Infrastructure project will be supported by the

---

[1] Information on GROWL can be found at http://www.grids.ac.uk/GROWL

[2] http://redress.lancs.ac.uk

[3] http://zope.cetis.ac.uk/profiles/uklomcore

NCeSS IPR Framework. This combines Creative Commons[4] coverage for download and re-purpose of resources, and a version of the Jorum Deposit Licence[5] for submission. In tandem with the technical adherence to standards, this framework allows support resources to be found, used and repurposed (to reduce duplication of effort).

Investigations

The final category of project output consists of a number of investigations. Work to capture requirements for Grid-enabling datasets is devising criteria for selecting datasets that offer the most value to the community. There will be a study of the feasibility and critical success factors for providing online access to valuable datasets like national census data while appropriately dealing with the confidentiality and security issues involved.

The project will conduct a survey to assess the degree of awareness of e-Social Science to help inform policy and practice within NCeSS as part of an ongoing, continuous process of engagement with social scientists. This survey will act as a baseline to assess the future effectiveness of the UK e-Social Science programme. The workings of the project itself will also be subject to ongoing evaluation in order to maximise its effectiveness.

A study to examine project technical standards will facilitate ongoing review of how the disparate technical elements feeding into the e-Infrastructure can start to converge. The study will take into account and feed into other activities in this area, such as the e-Research Tools and Resources Interoperability project (eReSS), the Joint Information Systems Committee (JISC) e-Framework, Open Source Software (OSS) Watch and the NGS in order to provide a cohesive approach to standards adoption by this and related communities.

Finally, the project is investigating issues related to sustainability, which will examine the steps needed to taken in order to provide highly effective and widespread support for many aspects of social science research (c.f. Voss et al. 2007a, this volume).

Challenges

User Requirements

We are aiming for research resources produced by the project to be appropriate to the needs of social scientists, useful in increasing the productivity and effectiveness of social science research, and incorporate highly usable user interfaces to increase and reinforce take-up. Realising these aims will enable the project to have a significant impact on research practice and begin to realise the potential for e-Social Science. However, whilst desirable, these aims have significant resource and methodological implications that will be hard to meet in the context of this project, which has relatively meagre resources in relation to its ambition.

Although usability focus groups and requirements gathering exercises are a part of the work of this project (for example, in the design of the portal user interface and in the assessment of which datasets to Grid-enable), employing a coherent user-centred design methodology will be challenging given the relatively small parcels of work focussed on developing existing (rather than new) applications. Investigations are underway to ensure the project retains a consistent focus on user needs through the project's internal evaluation exercise, which

4 http://creativecommons.org/

5 http://www.jorum.ac.uk/docs/word/JORUM_Deposit_Licence_11_07_05.doc

includes focus groups of project team leaders. The necessity of placing the user at the centre of the development process will also feed into investigations into how to sustain project objectives.

From Research to Service

Much of the project's programme of work involves building on outputs from the first phase of the NCeSS research programme, which yielded many promising e-Social Science applications. One of the challenges for this project is make the transition for tools that were research outputs to prototype services that can support more effective social science research.

This shift in emphasis has implications for academics involved in the original research when engagement in software development to create robust tools for others may not offer the same recognition as more traditional research outputs, such as academic papers. Additionally, the investment required to develop robust and reliable software is high and may be beyond what this project can achieve. (In this respect, this challenge is similar to the comprehensive user engagement challenge described above.) This is one of the reasons why work to investigate the sustainability of project outputs and objectives is so important to enable this project to become a platform for future work. There must also be management of expectations of what this project can, or cannot, achieve that is commensurate with the resources available.

The issue of "single sign-on" for services implies that resources on the Grid need to be closely integrated. This is a middleware problem and is being tackled by computer scientists around the world. The adoption of Shibboleth as an authentication technology in the UK may be of benefit here (Jie et al. 2007, this volume). A second issue is the "client problem" – how does a researcher connect to the Grid and to related technologies such as collaboration tools? Typically this requires a sophisticated level of expertise, installation of complex (and sometimes unreliable) software with complex dependencies, and punching holes in firewalls which are widely used as the primary protection for key resources in UK academic institutions. All this is difficult for a non-specialist.

Two issues require longer-term investigation:

1. how data will be moved from the source onto a server or "gateway" for manipulation prior to running the computational model on the Grid;
2. and how more flexible client access can be provided from desktop PCs.

Current solutions to issue (1) are: (i) obtain data on CD or DVD as is now the case and upload it to the server. Probably FTP can be used for this purpose, for instance an FTP server could be added to the list of GROWL services. SRB[6] or WebDav[7] are other options. GROWL already has an SRB client interface but SRB is not so widely used in social science as it is on other data grids. It is also possible to download data directly from the Web if the URL of the resource is known and the user has, e.g., a valid Athens or Shibboleth account. (ii) We could also implement other solutions being adopted in related projects such as Grid Enabling MIMAS Services (GEMS)[8], which uses OGSA-DAI[9] middleware.

---

[6] Storage Resource Broker, widely used for Grid data management http://www.sdsc.edu/srb

[7] http://www.webdav.org/

[8] http://pascal.mvc.mcc.ac.uk:9080/gems

[9] http://www.ogsadai.org.uk/

As noted above, current solutions to issue (2) are: (i) a Web interface such as the portal; (ii) GROWL client toolkit shell commands; (iii) GROWL plug-in to R, Stata, etc. running on the desktop. The latter would require some software installation and is currently far from trivial with Windows systems. Appropriate security has to be implemented for accessing the server, for instance in the UK, data services will be protected using Shibboleth authentication services,,but on the computational Grid, X.509 certificates are used as is commonplace in other internet applications using the TLS (formerly SSL) standard. Other work is addressing how to combine these technologies and the e-Infrastrucuture standards strand will assess outcomes.

Embedding within an Infrastructural Framework

The final challenge relates to embedding the e-Infrastructure and associated tools and services within a wider infrastructural framework such as offered by the NGS. Whilst it is desirable for sustaining the project outcomes to embed the infrastructure within a centralised national Grid service, the social science community is likely to present particular challenges to a model that has to date largely been focussed on the provision of Grid services for the natural sciences.

The challenge for this project is to articulate social science requirements in order to ensure that NGS service provision is appropriate to this community. If successful, we have an opportunity to shape aspects of the NGS in ways which are likely to have ramifications beyond the domain of social science and to improve the NGS for all its users.

Looking at activities elsewhere in the world, the US TeraGrid runs a successful Science Gateways programme that aims to support communities in delivering access to advanced IT infrastructures. Many of the gateways built under this programme are portal-based and experiences gained in the US will provide valuable input for the e-Infrastructure portal development. There is also a Virtual Institute for Computing in Humanities, Arts and Social Science (CHASS) run as a cyberinfrastructure activity. The Humanities, Arts and Social Sciences Research Group of the Open Grid Forum also works on widening uptake of e-Infrastructures and will provide an ideal platform for the development of relevant technical standards required by the e-social science community. We believe that the UK is in a good position to lead the way in this research domain.

Acknowledgements

This project is funded by the UK ESRC under grant number RES-149-25-1063.

References

R.J. Allan, R. Crouchley and M. Daw (2006) e-Collaboration Workshop: Access Grid, Portals and other VREs for the Social Sciences. *Workshop Report, ARIADNE Issue 49* (October 2006) http://www.ariadne.ac.uk/issue49/e-collab-rpt/

Crouchley, R., van Ark, T., Pritchard, J., Kewley, J., Allan, R., Hayes, M. and Morris, L. (2005) Putting Social Science Applications on the Grid. *Proceedings of the 1st International Conference on e-Social Science*. Manchester, UK.

Grose, D., Crouchley, R., van Ark, T., Kewley, J., Allan, R., Braimah, A. and Hayes, M. (2006). SabreR: Grid-enabling the analysis of multi-process random effect response data in R. *Proceedings of the 2nd International Conference on e-Social Science*. Manchester, UK.

Jie,W., Daw, M., Procter, R. and Voss, A. (2007). Towards Shibboleth-based Security in the e-Infrastructure for Social Sciences. *Proceedings of the 3rd International Conference on e-Social Science*, Ann Arbour, Michigan, US.

Severance, C., Hardin, J., Golden, G., Crouchley, R., Fish, A., Finholt, T., Kirschner, B., Eng, J. and Allan, R. (2007) Using the Sakai Collaborative Toolkit in e-Research Applications. *Concurrency and Computation: Practice and Experience* 19(12).
DOI: 10.1002/cpe.1115

Voss, A., Mascord, M., Fraser, M., Jirotka, M., Procter R., Halfpenny, P., Fergusson, D., Atkinson, M., Dunn, S., Blanke, T., Hughes, L. and Anderson, S. (2007): e-Infrastructure Development and Community Engagement. *Proceedings of the 3rd International Conference on e-Social Science*, Ann Arbour, Michigan, US.

Voss, A., Procter, R., Hewitt, T., Asgari-Targhi, M., Daw, M., Baun, C. and Gentzsch, W. (2007a). Sustainability of e-Infrastructures (for the Social Sciences). *Proceedings of the 3rd International Conference on e-Social Science*, Ann Arbour, Michigan, US.