Grid Technologies for Social Science: the SAMD Project

*by Celia Russell, Keith Cole, M. A.S. Jones, S.M. Pickles, M. Riding, K. Roy, M. Sensier**

Abstract
The Seamless Access to Multiple Datasets (SAMD) project is designed to demonstrate the benefits of Grid (e-Science) technologies for dataset manipulation and analyses in a social science context. Grid technologies run over existing internet infrastructures and offer a faster alternative to the world wide web for the transfer and analysis of large datasets. Under the SAMD project, a web-delivered social science dataset was made available for large-scale data analysis through a Grid architecture. Using an exemplar problem drawn from the UK social science community, the project demonstrates how the integration of a single sign-on environment, Grid technologies and access to high performance computational resources can significantly speed up computationally intensive queries and streamline data gathering and analysis. The approach can be generalised to virtually any kind of problem involving data retrieval and analysis. The paper also discusses how this could allow social scientists to significantly scale up their quantitative research inquiries.

Keywords:
e-Science; e-Social Science; grid technologies; time series data; multivariate analysis; Grid Security Infrastructure; digital certificates; authentication; authorisation; single sign-on.

Background
e-Science grids run on existing internet hardware but offer a more powerful infrastructure than current web technologies. They have been described as a massive extension or even the next generation of existing web services. Just as the web is designed to view html documents worldwide, grid technologies are designed to provide seamless access to large scale datasets and to share large scale computing resources, including specialised facilities and visualisation technologies.

A number of factors make quantitative social science research well suited to Grid based research strategies. Human behaviour takes place in a dense social and economic context, yet limitations in computing power and problems of data access inevitably result in models that oversimplify environmental conditions. Furthermore, many social scientists now wish to develop more complex research questions by combining datasets from more than one source, perhaps over different geographies or time periods. The datasets themselves are growing rapidly in size; the UK 2001 census datasets are estimated at over 20 gigabytes, twice the size of the previous release. Moreover, analysis of data on themes such as trade, economic governance, human development or industrialization requires research strategies that can accommodate a high degree of interdependency between what may be hundreds of disparate variables.

The Seamless Access to Multiple Datasets (SAMD) project was funded by the UK's Economic and Social Research Council (ESRC) and the Department of Trade and Industry. The Manchester Information and Associated Services (MIMAS) and the Supercomputing, Visualization and e-Science Centre at Manchester Computing in the University of Manchester were responsible for implementation. The project was designed to demonstrate the benefits of Grid technologies for dataset manipulation and analyses in a social science context. Firstly, the project showed how an existing social science dataset can be made available for large-scale data analysis via the Grid. Secondly, the project demonstrated how the development of parallelised software tools can significantly speed up computationally intensive analyses. Using an exemplar problem drawn from the UK social science community, the project showed how the integration of access to both data and high performance computational resources within a single sign-on environment enables the automation of complex workflows and can facilitate the scaling up of social science research applications. SAMD also demonstrated the successful incorporation of emerging Grid technologies into an existing social science data service.

The Problem
SAMD was based around a genuine econometric research question drawn from the academic community and is based on a typical social science databank, the National Statistics Time Series Data. MIMAS hosts a web interface to this databank, which is widely used in teaching and research. The research question examines the asymmetries in the response of UK Gross Domestic Product to interest rate changes. Sensier,

Osborn and Öcal[1] found that interest rate effects on Gross domestic product are larger when lagged growth has been high and when there has been a substantial increase in the interest rate. They did not find the opposite effect for low growth phases and decreases in the interest rate. The authors used a non-linear regression, namely the smooth transition regression model, to model this asymmetry. To help find the starting parameters, a computationally intensive 5-dimensional array search program was used. The program typically took several hours to run on a standard mainframe.

Before the existence of SAMD, the required data were collected over the web, returned to the user's PC and then sent manually to the mainframe computer for analysis. Altogether, the collection, collation and analysis of the data took around a working day. Moreover, the user had to access a variety of resources to acquire and analyze the data, all of which required authentication via multiple usernames and passwords.

The SAMD Solution

The SAMD project used Globus 2.0 (Globus is the open source software used to create Grids) and GridFTP to create a specific Grid architecture incorporating the National Statistics Time Series Data. A number of performance improvements, including parallelisation, were made to the analysis code to represent a High Performance Computing facility.

The demonstrator application transfers multiple time series from MIMAS (a national data centre based at Manchester Computing) to a high performance computing (HPC) engine for computationally intensive analysis, and then returns the results to the user. The entire operation requires only a single sign-on (grid-proxy-init) on the user's workstation. In step (1), the application on the user's workstation searches for and requests one or more time series via HTTPS with GSI authentication. In (2-3), CGI programs running on the web server verify that the user has permission to access the requested dataset. In (4-7), the requested data are extracted from the MIMAS data repository and copied to a short-lived file, and an XML "ticket" is returned to the application. Based on information supplied in the ticket, the application uses GridFTP to initiate a third-party file transfer from MIMAS to the HPC engine. Steps (1-7) are repeated for each group of time series. In (9-10), Globus mechanisms are used to launch an analysis run on the HPC engine and to retrieve the results. A housekeeping task cleans up temporary files.

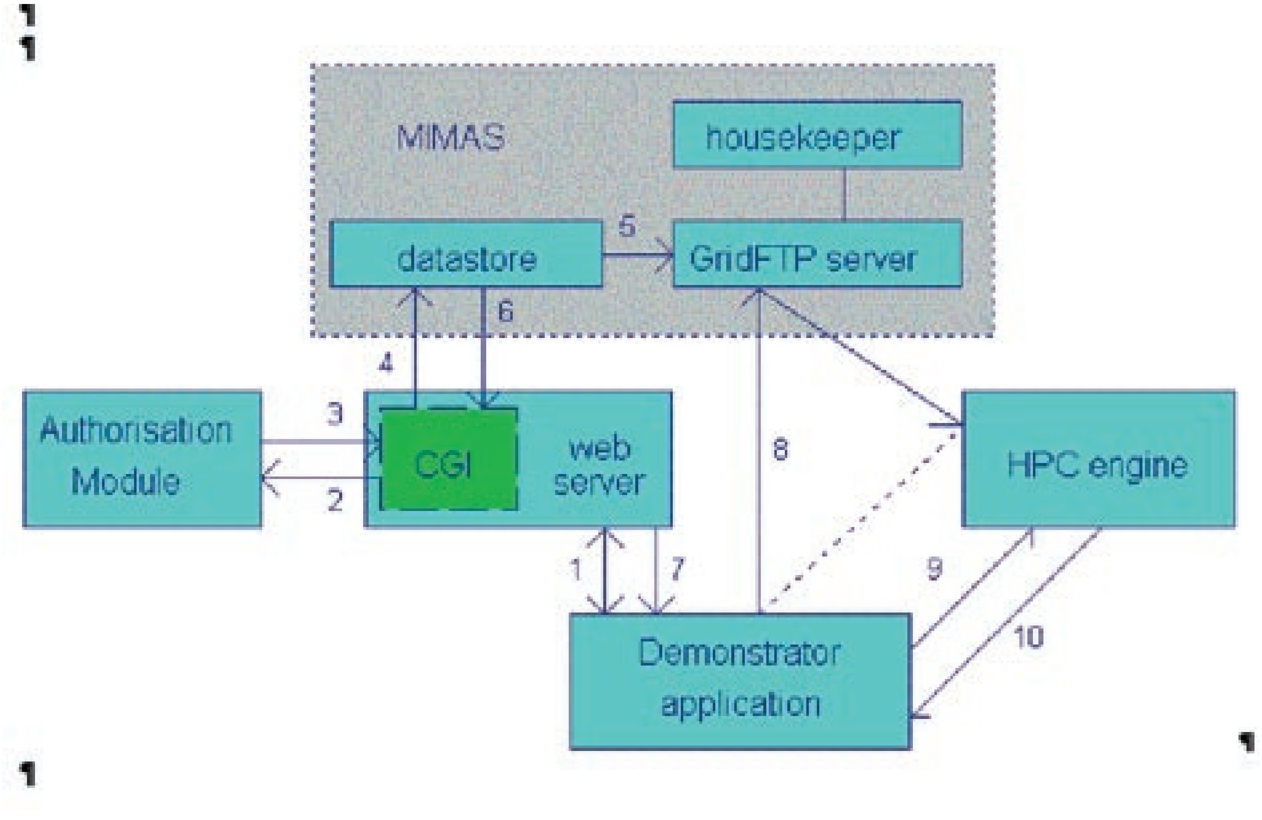

Figure 1 SAMD Architecture

A graphical user interface allows the user to search the databank, extract the data and transfer the resulting time series directly from the databank to an high performance computing facility via Grid FTP. The results are then returned to the user. The process is further streamlined by a security model which gives the user access to all the resources to which they have permission through a single sign on procedure. As a result of these advances, the data collection and analysis, which had previously taken around a working day, is reduced to a matter of minutes.

The project team also developed a command line shell script that runs the essential steps of data retrieval, transfer and analysis. The scripts enable the query to be scheduled to rerun automatically (whenever the datasets are updated, for example). The shell scripts also demonstrate that the principles of the project can be reapplied to other applications without the need to develop a specialised graphical interface.

Future Research Applications
By substantially decreasing the time required to locate, transfer data and perform a complex analysis, SAMD showed how a Grid approach could allow the social scientist to significantly scale up their research. The demonstration problem used in the SAMD project will now be extended to include the effects of international influences (such as US monetary policy decisions) on business cycles in Germany, France Italy and the UK. In addition to the National Statistics Time Series Data, this new model will incorporate data from the OECD's Main Economic Indicators and the IMF's International Financial Statistics. The use of Grid technologies allows the models to be rerun automatically whenever the databases are updated.

A Grid strategy also enables users to include a much larger number of data points in a computationally intensive analysis. For instance, a current cluster analysis of the Samples of Anonymised Records (a dataset derived from the 1991 and 2001 UK Census) uses just 1% of the available data. Using a Grid approach allows the entire dataset to be used. More generally, using the Grid permits researchers to attempt more computationally intensive analyses (including non-statistical modelling methods), or drop assumptions to create a model that more closely reflects the complexities of the real world.

The SAMD architecture also encourages the cross–analysis of multiple datasets, an issue of increasing importance as researchers develop more complex multi-level models, or investigate the interdependencies between economies, societies and the environment. Although SAMD was built around a single databank and computing facility, the architecture can be simply extended to include multiple databanks and computing resources. A single sign-on procedure gives the user access to any resource to which they have an entitlement wherever it is physically located. As a result, databanks hosted by a range of data providers can be accessed and searched seamlessly as an apparently single resource.

Project Outcomes
SAMD was based on a particular substantive problem, but its approach and methods can be generalised to virtually any kind of social science research involving data retrieval and analysis. One aim of the project was to reduce the technical barriers that have slowed the diffusion of Grid technologies into social science applications. This was achieved by developing features such as the desktop graphical user interface and transparent access system. These generic elements can be modified and redeployed in future social science Grid applications

The SAMD web site (http://www.sve.man.ac.uk/Research/AtoZ/SAMD) contains an overview of the project and a page from which various resources developed in the project can be downloaded. These resources include the patches to mod_ssl, shell scripts, presentations and handouts.

SAMD was the first of the ESRC e-Social science pilot projects to be successfully completed. As such, it provides a concrete example of the benefits e-Science could offer in a social science context. Consequent to the project completion in 2002, the ESRC developed a broader e-Social science strategy with the establishment of a National Centre for e-Social Science at Manchester. This £7.5 million programme will stimulate the acceptance of Grid technologies in social science applications. The five elements of the ESRC programme are presented in Appendix 1

Appendix 1
The ESRC e-Social Science Strategy
As part of the broad cross council e-Science programme, ESRC has been developing an e-Science strategy to stimulate adoption and use by social scientists of new and emerging Grid-enabled computing and data infrastructure, both in quantitative and qualitative research. The ESRC e-Science programme currently consists of the following five integrated components:

1) **The National Centre for e-Social Science (NCeSS)**
NCeSS is the key component of the ESRC e-Science strategy. The NCeSS will have a distributed structure, consisting of a co-ordinating Hub based at the University of Manchester in collaboration with the UK Data Archive at the University of Essex, and a set of research-based Nodes distributed across the UK. There is an overall budget of £4.5 million for the Nodes, which will be commissioned in 2004 and begin work in April 2005. See (http://www.ncess.ac.uk/).

2) **Pilot demonstrator projects**
A programme of 11 small-scale pilot projects began in autumn 2003. These projects are exploring the potential application of Grid technologies within the social sciences.).

3) **Scoping studie**s
The ESRC commissioned four scoping studies aimed at identifying the key issues that the NCeSS should address through its research programme. These scoping studies are available via the ESRC website at http://www.esrc.ac.uk/esrccontent/researchfunding/esciencecentre.asp

4) **A training and awareness programme**
Co-funded by the Joint Information Systems Committee (JISC) and the ESRC, this programme will highlight the potential of e-Science within the social science community and develop materials and courses to equip researchers to exploit Grid technologies. See (http://www.jisc.ac.uk/index.cfm?name=circular_2_03)

4) **Establishing a Network of Access Grid Nodes for UK Social Scienc**e
ESRC is establishing a network of Access Grid Nodes (AGNs) for the UK social science community. This network will facilitate active participation in remote group-to-group interactions for large-scale technical collaborations, e.g., distributed meetings, collaborative teamwork sessions, seminars and training.

* Paper presented at the IASSIST Conference, Madison, May 2004, by Celia Russell (MIMAS, Manchester Computing, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL United Kingdom). Contact: celia.russell@man.ac.uk (Web-site: URL: http://www.esds.ac.uk/international).**)**

Footnotes
[1] Sensier, M., Osborn D.R. and Öcal N. (2002) ‘Asymmetric Interest Rate Effects for the UK Real Economy’, with *Centre for Growth and Business Cycle Research Discussion Paper Series*, University of Manchester, No. 10. Forthcoming in *Oxford Bulletin of Economics and Statistics*, September 2002.