Enabling Quantitative Data Analysis through e-Infrastructures

Enabling Quantitative Data Analysis Through e-Infrastructure

Koon Leai Larry Tan
Paul S. Lambert
Ken J. Turner
Jesse Blum
Vernon Gayle
Simon B. Jones
University of Stirling
Richard O. Sinnott
University of Glasgow
Guy Warner
University of Stirling

This article discusses how quantitative data analysis in the social sciences can engage with and exploit an e-Infrastructure. We highlight how a number of activities that are central to quanti- tative data analysis, referred to as ‘ ‘data management,’ ’ can benefit from e-Infrastructural sup- port. We conclude by discussing how these issues are relevant to the Data Management through e-Social Science (DAMES) research Node, an ongoing project that aims to develop e-Infrastructural resources for quantitative data analysis in the social sciences.

Keywords: data management; quantitative data; e-Infrastructure; workflows; metadata

Authors’ Note: DAMES is an NCeSS Research Node funded by the UK ESRC under grant RES-149-25-0066. Please address correspondence to K. L. L. Tan, Department of Computing Science and Mathematics, University of Stirling, Stirling FK9 4LA, Scotland, United Kingdom; e-mail: klt@cs.stir.ac.uk.

Introduction

Quantitative Data Analysis in the Social Sciences

Quantitative data analysis represents one of the major forms of research evidence in the social sciences. A common definition of quantitative data is that it involves numerical representations of information. Quantitative data emerge from large- and small-scale social survey projects as well as from several other forms of social research, including experimen- tal designs and access to administrative data. Key activities of quantitative data analysts involve accessing appropriate social science information (e.g., downloading a copy of a major survey data set); managing and manipulating the content of the data (e.g., performing transformations and data linkage); and undertaking statistical analysis of this data. Analysis may use simple statistical summary techniques, and advanced statistical models, whose estimation is often at the forefront of statistical theory.

e-Science Background: Quantitative Data Analysis and e-Social Science

Numerous e-Social Science services have been built to support collaborative research activities related to quantitative data analysis in the social sciences. These include support for data sharing, data integration, and data analysis (e.g., Birkin et al., 2005; Grose, Crouchley, & van Ark, 2006; Lambert, Gayle, et al., 2008; Peters, Clark, Ekin, Le Blanc, & Pickles, 2007). These services feature scalable, interoperable, secure, dynamic, service-oriented environments that are designed to support current and future research requirements. In the United Kingdom, the National Centre for e-Social Science (NCeSS, **www.ncess.ac.uk)** was created by the Economic and Social Research Council (ESRC) in 2004 to coordinate the development of e-Infrastructure for the social sciences and promote its adoption in research practice. We focus in this article on the work of the Data Management through e-Social Science (DAMES, **www.dames.org.uk)** project, one of the NCeSS programme's research Nodes.

Overview of the Article

Section ''State of the Art'' discusses at a high level the roles and approaches of e-Infrastructure in general. It also covers the context of quantitative data analysis in the social sciences, expanding on selected examples of e-Social Science projects. Section ''An e-Infrastructure for Quantitative Data Analysis'' identifies the requirements and challenges for an e-Infrastructure for quantitative data analysis. It covers the strategy of the DAMES Node and how it will address these challenges. Section ''Conclusions and Future Work'' discusses expected outcomes of DAMES and future work.

State of the Art

e-Infrastructure in General

Grid computing technologies embrace a heterogeneous range of Internet resources and computing facilities related to enhanced collaboration and communication. Research communities may use these technologies on a regional, national, and global scale. ''e-Infrastructure'' describes the technologies and procedures that support research undertaken in this way.[1] There has been a great deal of investment in developing and promoting e-Science approaches for the benefit of scientific research over the last decade (Hey & Trefethen,

2004; National Science Foundation, 2006).

e-Infrastructure enables resources to be shared remotely via standard protocols, maximizing collaborative contributions toward common objectives. Large-scale and resource-intensive processes may no longer be limited by local resource constraints. Faster discovery of new drugs and a better understanding of climate change and its environmental impacts are just two examples of research enabled through e-Infrastructure (Research Councils UK e-Science Programme, 2007).

e-Social Science Examples

Grid enabled occupational data environment. The Grid Enabled Occupational Data Environment (GEODE, **www.geode.stir.ac.uk)** was an NCeSS small grant project, which sought to grid enable specialist data resources concerned with information about occupations. GEODE was motivated by problems experienced by quantitative social scientists in sharing and exploiting occupational information resources. The project identified problems with previous dissemination of this information. These issues reflected a lack of formal description of existing data, inadequate usage instructions and explanations of resources, and insufficient dissemination mechanisms (Lambert, Tan, Turner, et al., 2007; Lambert, Tan, Gayle, Prandy, & Turner, 2008). The project addressed these shortcomings by developing a portal service that allows social scientists to deposit their own occupational information and to search for other deposited data. The portal features a specific application service to address a commonly needed requirement of linking (‘‘matching’’ or ‘‘mapping’’) occupational information with the researcher’s own quantitative data. Technical details of how GEODE approached these issues are found in Tan, Gayle, Lambert, Sinnott, and Turner (2006).

The GEODE architecture is intended specifically to meet the requirements for supporting specialist occupational data. Popular standards (such as the Data Documentation Initiative, 2008; OASIS Web Services Resource Framework Specifications 1.2, 2008) and well-established middleware (Antonioletti et al., 2005; Foster, 2006) are used. There was a need to extend the data abstraction middleware Open Grid Service Architecture—Data Access and Integration (OGSA-DAI) to suit the requirements of GEODE. This was done to incorporate a metadata schema as part of each data resource, along with customized metadata management functionality. Outputs from the GEODE project include a gateway for standardized dissemination, sharing and access of occupational information resources; user-friendly support for linking survey data sets with occupational data; and an environment where researchers with the same interests can collaborate over their resources. The GEODE services are now being supported as part of the DAMES research Node.

Data management through e-Social Science. DAMES is an NCeSS research Node focused on supporting social scientists in tasks related to ‘‘data management’’ and the manipulation of social science data. Several of the Node’s activities are oriented toward quantitative data analysis. A theme known as ‘‘Grid Enabled Specialist Data Environments’’ deals with specialist data related to occupations, educational qualifications, and ethnicity. In this field, there have been many previous research efforts exploring the meaning of different types of specialist information and how it can be handled. For example, there have been evaluations of occupation-based social classifications (Ganzeboom & Treiman, 2003; Rose, Pevalin, & O’Reilly, 2005); work on the comparability of different educational qualification titles over time and between countries (Brynin, 2003; Schneider, 2008); and research on how alternative conceptual foundations for the measurement of ethnic groups can be realized in quantitative data analysis (Bosveld, Connolly, & Rendall,

2006; Lambert, 2005). Nevertheless, there have been few efforts to standardize access to and exploitation of specialist information in each area and current standards in using such data are inconsistent. The GEODE project developed a system for accessing and reviewing information resources on occupational units. In DAMES, this approach is being expanded with improved quantitative data on occupations and with new resources on educational qualifications and ethnicity.

Other research themes within the DAMES Node concern specialist data linkages associated with the analysis of social care data, e-Health records, and generic provisions related to the topic of ''data management.'' Through DAMES, interoperability across specialist data sets will be achieved to support the preparation and analysis of quantitative data. DAMES is therefore working to create an e-Infrastructure for supporting quantitative data analysis in chosen social science research domains (also see 3.2 below).

Grid enabled microeconometric data analysis. Another relevant project in e-Social Science is GEMEDA (Grid Enabled Microeconometric Data Analysis; Peters et al., 2007), which addressed the problem of data constraints in research into the economic welfare of ethnic groups within the United Kingdom. GEMEDA overcomes these constraints by using statistical data fusion techniques to combine data from various sample survey and census sources. This work requires operations of data virtualization and linkage. GEMEDA used the OGSA-DAI middleware to access and transfer remotely hosted data. In addition, metadata about the data sets was collated to support effective data linking. High-performance computing (HPC) resources provided through the UK National Grid Service (NGS, **www.ngs.ac.uk)** were used to make the statistical data fusion workload tractable, and the results were depicted visually. GEMEDA also made use of Athens (now being replaced by Shibboleth) as the trust federation for exchanging security attributes in the UK Higher Education sector. One area of particular interest was the execution of statically defined workflows in GEMEDA, demonstrating the practical application of workflows in e-Social Science.

Common themes in e-Social Science for quantitative data analysis. GEODE, GEMEDA, and DAMES, along with several other e-Science projects directed to quantitative analysis in the social sciences, share many common requirements and have often adopted similar approaches. Shared requirements include attention to resource virtualization, metadata, data integration, security, workflows, and HPC. A further common theme concerns the interface and usability aspects of e-Science services as nonfunctional but important issues.

We argue below that each of these requirements constitutes an important component of a unified e-Infrastructure for quantitative data analysis in the social sciences, and we outline below how the ongoing work of the DAMES Node should develop such an e-Infrastructure.

Data Management in Quantitative Data Analysis

An effective e-Infrastructure must engage with the practical experience of social science researchers. One enduring feature of all social science projects associated with quantitative analysis of social science data sets is that a significant component of research time involves manipulating and adjusting data after it has been accessed. These activities are often referred to as tasks of ''data management'' and are the focus of the DAMES Node.

A case can be made that data management tasks are ripe for support through e-Infrastructure. First, there are vast volumes of relevant quantitative data available to social scientists, and a major part of a social researcher's activities may concern identifying, linking together, and manipulating complex, related resources. Although research data are often distributed in a standardized or semi-standardized way—for example, the UK Data Archive offers access to survey data sets with standard formats and documentation (UK Data Archive, 2008)—data are made complex by heterogeneous topic coverage, the existence of many nonstandardized resources, and the sheer volume of potentially relevant resources.

Second, a significant capacity shortfall in quantitative social science research skills is recognized in many nations (Bardsley, Wiles, & Powell, 2006) and has been attributed in whole or in part to social scientists' difficulty in exploiting the moderately advanced software programming that is hitherto required for most data management tasks (Kohler & Kreuter, 2005). Third, social researchers are increasingly aware of the exciting enhancements to their analysis that might be possible with greater efforts in data management. These may include enhancing or linking related data resources (UK Data Forum, 2008) and improved standards in documentation and replicability of analysis (Dale, 2006; Freese, 2007). Taken together, these three observations on data management within quantitative social science research highlight areas where integrated collaborative resources could be effectively developed and distributed in an e-Infrastructural model.

Key data management tasks for quantitative data resources involve ''variable operationalizations'' and ''linking data.'' The former involves efforts to transform the numeric data stored on a particular measure into an effective analytical variable. Common practice involves recoding complex categorical variables into a smaller and more tractable range of different categories. The latter involves enhancing existing data with additional information drawn from a separate resource (such as the use of freely published aggregate statistical data on occupations to enhance data with details of occupational titles, an application of linking data for which services were developed in GEODE, see Lambert Tan, Turner, et al., 2007).

The potential contribution of resources for variable operationalizations and linking data can be better appreciated through use-cases. As an example, we highlight a recent analysis of intergenerational social mobility trends (Blanden, Goodman, Gregg, & Machin, 2004) that might have been improved with better practice in data management. (''Social mobility trends'' refer to patterns in the extent to which measures of parental background effect an adult's own socioeconomic attainment.) Although a popular and politically influential analysis, the findings of Blanden et al. of declining social mobility in contemporary Britain were criticized as highly misleading about longer term trends in social mobility in the United Kingdom (Ermisch & Nicoletti, 2007; Goldthorpe & Jackson, 2007; Lambert, Prandy, & Bottero, 2007).

- Linking data: Blanden et al. (2004) used data from two major UK social surveys, the birth cohort studies of 1958 and 1970. However, many other representative survey data sets also cover comparable intergenerational data. Ermisch and Nicoletti (2007) and Lambert, Prandy, et al. (2007a) linked together a wider range of other data resources to draw different conclusions on the same topic.

- Variable operationalizations: Blanden et al. (2004) measured social mobility in terms of income measures for parents and their adult children. However, many other means of assessing intergenerational mobility may be used. Goldthorpe and Jackson (2007) demonstrated that the analysis of occupational data from the same surveys gave different conclusions on long-term trends.

The use-case above is a typical illustration of how work involved in the data management of quantitative research data is typically conducted independently between projects and may not adequately capitalize on all relevant resources. The analyses by Blanden et al. (2004), Lambert et al. (2007), Ermisch and Nicoletti (2007), and Goldthorpe and Jackson (2007) shared similar features in relation to linking data and operationalizing variables. All four identified and combined related data resources, and all four undertook substantial bespoke exercises in developing and analyzing measures (of income and occupations). An infrastructural resource to enhance access to and linking of suitable data, and to support transparent variable operationalizations, could have improved the conduct of the above research. For instance, it is conceivable that a workflow model and record of the various choices in linking data and operationalizing variables could contribute to the preservation and replicability of these complex data analytical tasks. DAMES (see section ‘‘Meeting e-Infrastructure Challenges in DAMES’’ below) is developing services to support such data management tasks. These may ultimately contribute to improved practice in social science research by supporting researchers in making better use of existing data resources.

An e-Infrastructure for Quantitative Data Analysis

e-Infrastructural Requirements

We define below a list of interrelated requirements and desirable features for an effective e-Infrastructure for quantitative data analysis in the social sciences.

- Resource virtualization. Quantitative data sets should ideally have standard access interfaces abstracting from actual formats and locations. Discovery middleware is required to provide exposure and probing mechanisms for resource providers and users, respectively, with functionality to semantically query for services and resources.
- Support for the use and management of metadata resources, which will contribute to the discovery of relevant related data.
- Support for data linkage as a high-volume activity, which may involve data resources that themselves are dynamically updated. This should be an accessible service with a scalable and flexible framework to access, transport, and transform virtualized data.
- Security is also required to ensure policies over data access are upheld and to ensure resource integrity and accountability. A ‘‘content-level’’ security approach is likely to be required as a means to enforce confidentiality within the data itself (see section ‘‘Security’’).
- Researchers should be able to access, manipulate, and analyze quantitative data using procedures that build upon previous endeavors. This would involve researchers exploiting previous approaches and in turn exposing their own procedures for future researchers. A workflow approach should allow the documentation and modeling of such activities.

- HPC may be required to raise the level of productivity for computationally demanding quantitative data analysis tasks.
- Usability is a nonfunctional aspect that is crucial to the uptake of the services and components to be developed, which is as important as the functional aspects.

One experience of the GEODE work was the discovery that components of the architecture for that service were applicable to other examples of quantitative social science data sets (Tan et al., 2006)0. This is because the components are generic given suitable data abstraction and metadata management. The components are also not bound to data sets from specific subdisciplines of social science but are generic e-Infrastructural requirements. Current initiatives in e-Social Science are contributing to developing these features, typically in the context of specialist research requirements. Wider ranging initiatives such as NCeSS and its e-Infrastructure project (Procter et al., 2006) coordinate approaches between initiatives and support generalizability.

Meeting e-Infrastructure Challenges in DAMES

As in the example of GEODE, many e-Social Science applications are developed only to address the aspects and requirements of particular research interests. Although these applications are specific, they exhibit common requirements and processes to a fundamental extent (as listed in 3.1). However, while well-established middleware often provides the technical capabilities for such services (e.g., using OGSA-DAI in GEODE), it does not ordinarily achieve this effect without customization or extension. We discuss below how the generic e-Infrastructural components identified in 3.1 above may be incorporated within the data management provisions of the DAMES Node. Whenever feasible, DAMES aims to use recognized standards to achieve these requirements.

Resource virtualization. Virtualization is one of the key characteristics of e-Infrastructure for realizing the vision of interoperability. Resources in a variety of formats can be accessed via standardized protocols, allowing researchers to work virtually across different formats seamlessly. Several data access functionalities have been deployed in e-Science projects (e.g., OGSA-DAI), but they may lack usability for social science researchers as they are too computing oriented.

Resource virtualization has implications for two stages common to most quantitative data analysis projects in the social sciences: accessing and reviewing data and manipulating data (or data management). The first typically involves identifying and inspecting the fundamental data that will be used in the research. The NESSTAR[2] service is one prominent existing provision in this field. However, data access often requires further processes of searching for related data, which may contribute to the intended analysis (the GEODE project was one example where a service assisting social scientists in accessing and exploiting occupational data was developed). Such latter activities are typically integrated with those of manipulating and managing the research data. Existing services tend to separate data access from data manipulation, but with suitable resource virtualization coordinated documentation of both processes is feasible.

DAMES will develop services for resource virtualization for quantitative social science data. It will let researchers define their data management activities, potentially resulting in repeatable procedures as part of an e-Infrastructural middleware. This set of data management activities is being developed according to the OGSA-DAI design pattern, configured as activities supported by the virtualized resources. The features of each activity will contribute to a suite of middleware suitable for generic quantitative data analysis activities.

Metadata. Data management metadata (such as including instructions for recoding variables) describes data manipulations for information extraction. Quantitative social science data sets usually have a considerable quantity of metadata associated with them, which provide information about the data resource, including the meaning of its variables and its provenance.

In the social sciences, metadata was traditionally recorded by data producers in an ad hoc fashion using a variety of nonstandard techniques. The adoption of metadata standards facilitates greater data discovery and access, collaboration among researchers, and data processing capabilities. Existing standards include the Data Documentation Initiative (DDI),[3] Dublin Core Element Set,[4] Statistical Data and Metadata Exchange (SDMX),[5] Metadata Encoding and Transmission Standard (METS),[6] and Common Warehouse Metamodel (CWM).[7] Some of these standards have been designed to complement one another (Gregory and Heus, 2007).

DDI version 3.0 was released in 2007. The latest release natively supports features such as improved coverage, groupings, comparisons, version control, and information processing. Such features add significant value to the data management domains of the DAMES Node, which is therefore building a DDI3 profile within all its data services.

The DDI3 framework supports data management activities as part of the data virtualization process. Within DAMES, the range of metadata covering data management activities will be defined and developed according to the design pattern of the OGSA-DAI middleware. Data virtualization can provide access to metadata resources and integrate them with wider ranging research activities. OGSA-DAI can virtualize resources but cannot incorporate metadata along with the virtualized data. Certain nongrid applications (e.g., the NESSTAR service associated with the European Data Archives) allow publishing data along with metadata schemas. However, these are not straightforwardly incorporated as services to be discovered and used by peer services. They were not developed within the paradigm of e-Infrastructure and service-oriented architecture and, therefore, have not adopted the implied standards. Data and metadata management functionality integrated with data resource virtualization is needed.

Discovery. Discovery mechanisms work alongside resource virtualization services to improve exploitation of data resources. For example, online resources may exploit discovery mechanisms to publicize themselves, promoting their uptake and potential interactions with other compatible resources. It is already possible to use grid middleware to facilitate discovery of social science resources. GEODE, for example, uses Globus MDS4 (Monitoring and Discovery System) to build its registry of virtualized data sets. MDS4 features triggers and notifications, which serve the purpose of alerting social science researchers to updates or observing the status of resources.

The discovery mechanisms for social science data resources are not trivial because of the variety of data being published in different formats and volumes across different social science disciplines. In terms of semantics, different ontologies, taxonomies, and standard schemas may be used for comparable resources. The several metadata standards for annotating social science resources (see section ''Metadata'') may have query tools that work differently. A natural question is whether it will be possible to have a discovery mechanism that supports such diversity. However, in the quantitative analysis of social science data, there are certain similarities between most data resources. For instance, almost all data resources are released in the form of relational tables or matrices, and most software formats and packages operate in a broadly similar manner to manipulate and analyze these tables. In many disciplines, similar standards for recording certain types of data (such as standard variables) are used. These similarities can fruitfully be exploited to develop a coordinated system. DAMES will develop an extensible discovery framework that allows a choice of tools as plug-ins for discovering resources across the diverse metadata implementations as well as facilitating resource discovery at all stages to maximize the exposure and dissemination of resources.

Data integration. Activities within quantitative data analysis are certainly data-centric. Virtualization, discovery, and access provide the basis for application-related activities, of which a major part is data integration. Integrating or linking data is often aimed at enhancing the value of the data. An example of interorganization data integration is linking between clinical records, patient records, disease registries, and so on to enable and support clinical trials and epidemiological studies (Sinnott, Stell, & Ajayi, 2007). Examples of intradisciplinary integration are the ''cross-walks'' (data linking) occupational data resources in GEODE and the integration of national surveys in GEMEDA.

Many existing data integration facilities are specified in terms that are not easily understood by social scientists (including the generic OGSA-DAI middleware and the specific requirements of quantitative data analysis software packages). Data integration capabilities could be abstracted so that users can more easily relate to and specify data integration. DAMES will provide a suite of tools that can specify data integration activities at a high level understood within the context of social science usable for researchers and equip them with the means to readily express, understand, and reuse data integration.

Security. Security is a common and critical requirement for much social science data. Apart from authentication and authorization, existing practices for accessing data resources in social science are particularly concerned with protecting data integrity and confidentiality. Some solutions require users to be physically present to use data that are isolated from remote access (this ''safe setting'' procedure requires each user's access activities and results to be monitored and filtered to prevent improper use of resources, see, for example, the ONS Virtual Microdata Laboratory[8]). Other solutions invoke extended (often manual) techniques that anonymize information to prevent identification of an individual from the records.

Identification of users and authorizing appropriate access satisfy most requirements for protecting data from improper use. e-Infrastructure has the technologies and vision to meet this requirement, with complex security measures including security attribute assertions, credential repositories, delegation, configuration of policies, security, and trust federations.

These technologies are well established and already widely used. Shibboleth[9] is an example of these technologies in action, federating security among numerous organizations. DAMES is using Shibboleth for several reasons. First, the infrastructure can manage the trust federation and security attributes interexchange between members. Second, Shibboleth already has a set of established procedures and software with acceptable performance. Third, many potential DAMES users are part of Shibboleth participating organizations. Shibboleth offers a seamless authentication and authorization framework among potential users of DAMES.

Surprisingly, there has not been as much development to support the requirement of preventing potential compromise of data by authorized entities, an issue that is especially relevant when permitting authorized access to remote resources. Addressing this challenge will influence the resources that are currently accessed and shared, improving the trust and involvement of data providers, and therefore bringing new possibilities for remote collaboration. The VANGUARD (Virtual ANonymisation Grid for Unified Access of Remote Data) project (Sinnott, Ajayi, Stell, & Young, 2008) has demonstrated the possibility of enabling such collaborations under similar tight security constraints. DAMES is evaluating existing approaches and techniques for data confidentiality, such as anonymization data algorithms, to determine the feasibility of incorporating these.

Social science workflows. A workflow comprises two or more existing services combined in a specified fashion, resulting in a new service. There are well-known workflow specification standards such as Business Process Execution Language (BPEL)[10] and Web Service Choreography Interface (WSCI), of which the most widely used is BPEL.

Workflow environments have been developed for domains such as bioinformatics—Taverna (Turi, Missier, Goble, DeRoure, & Oinn, 2007) and Open Middleware Infrastructure Institute (OMII-BPEL) (Emmerich, Butchart, Chen, Wassermann, & Price, 2005) for scientific workflows. Taverna introduces a workflow language Simple Conceptual Unified Flow (SCUFL) and enactment workbench specifically for designing and executing bioinformatics workflows. OMII-BPEL extends a BPEL implementation to support large-scale scientific workflows. OMII-BPEL is made available as middleware, with a workbench environment for designing and monitoring. The P-Grade portal (Kacsuk & Sipos, 2005) allows users to graphically build, execute, monitor, and manage workflows via a portal interface. An advantage of P-Grade is the support of a wide range of grid middleware, including legacy code. However, its workflow specification is not standards based.

In general, a workflow is built using constructs such as iteration, callouts to peer services, and assignments. These support the basic workflow requirements, which may or may not be applicable to social science research. In quantitative data analysis in the social sciences, higher level workflow building blocks may be identified. These would include analysis functions usually performed by researchers as well as data access, manipulation, and integration. A comprehensive set of constructs oriented toward social science activities would allow the building of workflows, potentially contributed by users themselves. As the pool of documented workflows expands, an increasingly productive library of facilities is developed (a process known as ‘‘workflow proliferation’’).

DAMES will develop services for capturing social science workflows applicable to quantitative data analysis, whereby researchers can reuse existing workflows and also

proactively contribute to the workflow pool. DAMES currently favors BPEL as it is an established and widely adopted standard with several implementations, including OMII-BPEL that can support large-scale workflows. Workflow constructs could be developed as extensions to BPEL. However, the design of P-Grade will be considered in the development of the workflow framework.

High-performance computing. One of the main motivations for the initial vision of the grid was to be able to perform intensive and large-scale computations, by pooling (heterogeneous) resources that may be distributed. Tasks may be completed at a fraction of the resources (time, cost, hardware, etc.) normally consumed when running them locally. HPC is attractive if there are parallel components within computations. HPC has been used in areas such as physics, medicine, astronomy, and social science, to name a few. For example, the SabreR project undertaken at CQeSS included a grid-based implementation of HPC for computationally intensive calculations in social science applications (Grose et al., 2006).

The UK NGS has a large number of high-throughput hardware and software resources, augmented with a support framework to ensure stability. DAMES will develop within its framework the capability for high-throughput computation, making use of existing services such as the NGS.

Interfaces and usability. It is necessary to establish how social science researchers view and interact with the e-Social Science environment. Within the remit of DAMES services, social scientists from different backgrounds are likely to have a range of expectations from the environment. DAMES currently favors a hybrid of desktop and portal environments and services to cater for the spectrum of users, whereby adaptation is minimized for experienced researchers and flexible accessibility provided for others.

Conclusions and Future Work

e-Social Science services have proven able to support quantitative data analysis and bring new possibilities to the way collaborations are achieved. We have seen in many projects, such as GEODE, practical examples of e-Social Science that enable remote and complex collaborations, improve research productivity, and contribute to the effectiveness of resource dissemination and sharing. Nevertheless, there is room for improving the coordination of services by moving toward an e-Infrastructure that underpins quantitative data analysis.

We have discussed and identified key development areas, derived from the experience of previous and ongoing projects, for e-Infrastructure to support quantitative data analysis: virtualization with integrated functionality for data and metadata management; metadata and discovery mechanisms appropriate to quantitative data analysis; security mechanisms applicable to social science; and proliferation of new services through workflows potentially contributed by peer researchers. Existing middleware can be extended to achieve these goals.

We have also elaborated how DAMES will address these challenges with its development of capabilities for supporting data management tasks associated with quantitative data analysis in the social sciences. The use of existing standards in DAMES, such as generic

middleware, is an important strategy for compatibility with related e-Infrastructure resources.

As the work of the DAMES Node continues, its tools and services will be subject to iterative evaluation to ensure that their development is driven by the needs of users. For example, new requirements might emerge for incorporating new techniques to anonymize data for confidentiality, new data and metadata management activities, and new workflow constructs These and other potential developments are likely to emerge and be transformed with iterative feedback from social science users. Such feedback should continue to shape the character of e-Infrastructural developments and the wider conduct of quantitative data analysis in the social sciences.

Notes

1. See Joint Information Systems Committee e-Infrastructure Programme. Retrieved April, 23, 2006, from http://www.jisc.ac.uk/whatwedo/programmes/programme_einfrastructure.aspx
2. See http://www.nesstar.com/about/background.html
3. See http://www.ddialliance.org
4. See http://www.dublin.core.org
5. See http://www.sdmx.org
6. See http://www.loc.gov/standards/mets
7. See http://www.omg.org/technology/documents/modeling_spec_catalog.htm
8. See http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/business-data/vml/index.html
9. See http://shibboleth.internet2.edu/about.html
10. See http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html

References

Antonioletti, M., Atkinson, M. P., Baxter, R., Borley, A., Chue Hong, N. P., Collins, B., et al. (2005). The design and implementation of grid database services in OGSA-DAI. *Concurrency and Computation: Practice and Experience, 17,* 357-376.

Bardsley, N., Wiles, R., & Powell, J. L. (2006). *A consultation to identify the research needs in research methods in the UK social sciences.* Southampton: National Centre for Research Methods, University of Southampton.

Birkin, M., Clarke, M., Chen, H., Dew, P., Keen, J., Rees, P., et al. (2005). MoSeS: Modelling and simulation for e-Social Science, University of Leeds. *Proceedings of the 4th UK e-Science All Hands Meeting, Nottingham.*

Blanden, J., Goodman, A., Gregg, P., & Machin, S. (2004). Changes in intergenerational mobility in Britain. In M. Corak (Ed.), *Generational income mobility in North America and Europe* (pp. 147-189). Cambridge: Cambridge University Press.

Bosveld, K., Connolly, H., & Rendall, M. S. (2006). *A guide to comparing 1991 and 2001 Census ethnic group data.* London: Office for National Statistics.

Brynin, M. (2003). Using CASMIN: The effect of education on wages in Britain and Germany. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in Cross-National Comparison* (pp. 327-344). New York: Kluwer Academic.

Dale, A. (2006). Quality issues with survey research. *International Journal of Social Research Methodology, 9,* 143-158.

Emmerich, W., Butchart, B., Chen, L. Wassermann, B., & Price, S. (2005). Grid service orchestration using the Business Process Execution Language (BPEL). *Journal of Grid Computing, 3,* 283-304.

Ermisch, J., & Nicoletti, C. (2007). Intergenerational earnings mobility: Changes across cohorts in Britain. *The B.E. Journal of Economic Analysis and Policy, 7,* 1-37.

Foster, I. (2006). Globus toolkit version 4: Software for service-oriented systems (pp. 2-13). *Proceedings of the IFIP International Conference on Network and Parallel Computing, LNCS 3779, Springer-Verlag.*

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods and Research, 36,* 153-172.

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnick & C. Wolf (Eds.), *Advances in Cross-National Comparison* (pp. 159-193). New York: Kluwer Academic Press.

Goldthorpe, J. H., & Jackson, M. (2007). Intergenerational class mobility in contemporary Britain: Political concerns and empirical findings. *British Journal of Sociology, 58,* 525-546.

Gregory, A., & Heus, P. (2007). DDI and SDMX: Complementary, not competing, standards, open data foundation. Retrieved October 16, 2008, from http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf

Grose, D., Crouchley, R., van Ark, T. (2006). SabreR: Grid-enabling the analysis of multi-process random effect response data in R. *Proceedings of the e-Social Science Conference, Manchester, June 28-30.* Retrieved September 29, 2008, from http://www.ncess.ac.uk/events/conference/2006/papers/papers/GroseSabreR.pdf

Hey, T., & Trefethen, A. (2004). UK e-Science programme: Next generation grid applications. *International Journal of High Performance Computing Applications, 18,* 285-291.

Kacsuk, P., & Sipos, G. (2005). Multi-grid, multi-user workflows in the P-GRADE portal. *Journal of Grid Computing, 3,* 221-238.

Kohler, U., & Kreuter, F. (2005). *Data analysis using Stata.* College Station, TX: Stata Press.

Lambert, P. S. (2005). Ethnicity and the comparative analysis of contemporary survey data. In J. H. P. Hoffmeyer-Zlotnick & J. Harkness (Eds.), *Methodological aspects in cross-national research* (pp. 259-277). Manheim: ZUMA-Nachrichten Spezial 11.

Lambert, P. S., Prandy, K., & Bottero, W. (2007). By slow degrees: Two centuries of social reproduction and mobility in Britain. *Sociological Research Online, 12*(1).

Lambert, P. S., Tan, K. L. L., Turner, K. J., Gayle, V., Sinnott, R. O., & Prandy, K. (2007). Data curation standards and social science occupational information resources. *International Journal of Digital Curation, 2,* 73-91.

Lambert, P. S., Tan, K. L. L., Prandy, K., Gayle, V., & Bergman, M. M. (2008). The importance of specificity in occupation-based social classifications. In Robert M. Blackburn (Ed), *International Journal of Sociology and Social Policy*, 28, 179-192. Emerald.

Lambert, P. S., Gayle, V., Tan, K. L. L., Blum, J., Bowes, A., Jones, S., et al. (2008). Grid enabled specialist data environments: Forward planning for GE*DE services for specialist data on occupations, educational qualifications, and ethnicity. Retrieved December 12, 2008, from http://www.dames.org.uk/docs/tech_papers/DAMES_tp2008-1.pdf

National Science Foundation. (2006). Cyber-infrastructure: A special report. Retrieved December, 2008, from http http://www.nsf.gov/news/special_reports/cyber/index.jsp

Peters, S., Clark, K., Ekin, P., Le Blanc, A., & Pickles, S. (2007). Grid enabling empirical economics: A micro-data application. *Computational Economics, 30,* 349-370.

Procter, R., Batty, M., Birkin, M., Crouchley, R., Dutton, W., Edwards, P., et al. (2006). The National Centre for e-Social Science. *Proceedings of the UK e-Science All Hands Meeting, Nottingham.*

Research Councils UK e-Science Programme. (2007). Highlights from the UK e-Science programme. Retrieved July 13, 2008, from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/escihighlights2.pdf

Rose, D., Pevalin, D., & O' Reilly, K. (2005). The NS-SEC: Origins, development and use. Basingstoke: Palgrave Macmillan.

Schneider, S. L. (Ed.). (2008). *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries.* Mannheim: MZES.

Sinnott, R. O., Ajayi, O., Stell, A., & Young, A. (2008). Towards a virtual anonymisation grid for unified access to remote clinical data. *Proceedings of the 6th International HealthGrid Conference, Chicago.*

Sinnott, R. O., Stell, A., & Ajayi, O. (2007). Supporting grid-based clinical trials in Scotland. *Health Informatics Journal*, Special Issue on Integrated Health Records.

Tan, K. L. L., Gayle, V., Lambert, P. S., Sinnott, R. O., & Turner, K. J. (2006). GEODE—sharing occupational data through the grid (pp 534-541). *Proceedings of the 5th UK e-Science All Hands Meeting, ISBN 0-9553988-0-0, Nottingham*.

Turi, D., Missier, P., Goble, C., DeRoure, D., & Oinn, T. (2007). Taverna workflows: Syntax and semantics. *Proceedings from the 3rd IEEE International Conference on e-Science and Grid Computing, Bangalore, India*.

UK Data Archive. (2008). Retrieved July 16, 2008, from http://www.data-archive.ac.uk/

UK Data Forum. (2008). The national strategy for data resources for research in the social sciences. Warwick: University of Warwick. Retrieved October 18, 2008, from http://www2.warwick.ac.uk/fac/soc/nds/

Koon Leai Larry Tan holds a BSc (Hons.) in software systems engineering. He is currently pursuing a PhD in computing science at the University of Stirling. His research interests are focused on specification, verification, and validation of grid service orchestration. He may be reached at klt@cs.stir.ac.uk.

Paul S. Lambert is a lecturer in sociology. His research interests cover the measurement and analysis of social stratification and inequalities. His methodological interests cover the manipulation of complex social survey data. He may be reached at paul.lambert@stir.ac.uk.

Ken J. Turner is a professor of computing science at the University of Stirling. His research interests include grid computing, services, workflows, and applications in social science. He may be reached at kjt@cs.stir.ac.uk.

Jesse Blum is a graduate researcher in computing science at the University of Stirling. His research interests focus on adaptable programming models and compositional services. In DAMES, he focuses on metadata standards for social science data and integration. He may be reached at jmb@cs.stir.ac.uk.

Vernon Gayle is a professor of sociology at the University of Stirling. His research interests include the sociology of youth, education, young people, and social stratification. His methodological interests include analysis of large-scale survey data sets, longitudinal data, and e-Social Science. He may be reached at vernon.gayle@stir.ac.uk.

Simon B. Jones has been a lecturer in computing science at the University of Stirling since 1983. He has a Bachelors degree in Physics from the University of York and MSc and PhD in Computer Science from the University of Newcastle upon Tyne. His research interests are object-oriented methods and web and grid services. He may be reached at sbj@cs.stir.ac.uk.

Richard O. Sinnott is the Technical Director of the National e-Science Centre at the University of Glasgow, Scotland. He organizes and administrates both UK wide work and local Glasgow University activities associated with e-Science/e-Research. He has over 100 publications across a range of computing science and application-oriented fields, most recently in the area of grid security and usability especially in the life sciences. He may be reached at r.sinnott@nesc.gla.ac.uk.

Guy Warner has worked in e-Science for several years. He is currently a member of the DAMES project at the University of Stirling. He was in the training team at the National e-Science Centre in Edinburgh. His background is in applied mathematics and computer programming/systems administration. He may be reached at gcw@cs.stir.ac.uk.