Enabling Quantitative Data Analysis through e-Infrastructures

K. L. L Tan[1], P. S. Lambert[2], K. J. Turner[1], J. Blum[1],V. Gayle[2], S. B. Jones[1], R. O. Sinnott[3], G. Warner[1]

[1] University of Stirling, Department of Computing Science and Mathematics

[2] University of Stirling, Department of Applied Social Science

[3] University of Glasgow, National e-Science Centre (NeSC)

Abstract. This paper discusses how quantitative data analysis in the social sciences can engage with and exploit an e-Infrastructure. We highlight how a number of activities which are central to quantitative data analysis, referred to as 'data management', can benefit from e-infrastructural support. We conclude by discussing how these issues are relevant to the DAMES (Data Management through e-Social Science) 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

1 Introduction

1.1 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 emerges from large and small scale social survey projects, as well as from several other forms of social research, including experimental design 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 dataset); managing and manipulating the content of the data (e.g. performing transformations and data linkage); and undertaking statistical analysis of this data, often using both simple

statistical summary techniques and advanced statistical models, whose estimation is often at the forefront of statistical theory.

1.2 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 social sciences. These include support for data sharing, data integration, and data analysis [4][6][10][19]. These services feature scalable, interoperable, secure, dynamic, service-oriented environments that are designed to support current and future research requirements. In the UK, the National Centre for e-Social Science [20] has coordinate many services and promote their contribution to social science research practice, pursuing ongoing development to innovate and sustain these collaborations.

1.3 Overview of the Paper

Section 2 discusses at a high level the roles and approaches of e-Infrastructure in general. It also covers the context of social science, expanding on selected examples of e-Social Science projects. Section 3 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 4 discusses expected outcomes of DAMES and future work.

2 State of the Art

2.1 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 (e-Science, e-Research, e-Health, e-Social Science, etc.) use these technologies on a regional, national and global scale. ‘e-Infrastructure’ is the term to describe the technology and procedures that support research undertaken in this way [16]. 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 [22][27].

e-Infrastructures have enabled research in many domains. Projects in physics (CERN), climate studies (Earth System Grid), medicine (BIRN Biomedical Informatics Research Network), and many others, have benefited from improved quality and possibilities of collaborations. Resources can now be shared remotely via standard protocols, maximising the contribution to common objectives amongst collaborators. Large scale and resource-intensive processes no longer have the legacy of local resource constraint, as e-Infrastructure overcomes resource limitation, gaining higher throughput returns. Faster discovery of new drugs and climate predictions are examples which have demonstrated the benefits of an e-Infrastructure.

2.2 e-Social Science Examples

2.2.1 GEODE

GEODE ('Grid Enabled Occupational Data Environment', www.geode.stir.ac.uk) was an ESRC Small Grants project (2005 to 2007) 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 [13][14]. The project addressed these shortcomings by developing a portal service which allows social scientists to deposit their own occupational information, and also 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 [11].

The GEODE architecture is intended specifically to meet the requirements for supporting specialist occupational data. Standards [7][35] and well established middleware [12][24] are used. There was a need to extend the data abstraction middleware OGSA-DAI (Open Grid Service Architecture – Data Access and Integration) to suit the requirements of GEODE. This was done in order to incorporate a metadata schema using DDI (Data Documentation Initiative [7]) as part of each data resource, along with customised metadata management functionality. Outputs from the GEODE project include a gateway for standardised dissemination, sharing and access of occupational information resources; user-friendly support for linking micro-survey datasets 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 between 2008 and 2011 [6].

2.2.2 DAMES

The DAMES project is a 'research Node' focused on supporting social scientists in tasks related to 'data management' and the manipulation of social science data [6]. Several of the Node's activities are oriented towards quantitative data analysis. A theme known as ‘Grid Enabled Specialist Data Environments’ deals with specialist data related to occupations, health, 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 numerous analysis of occupation-based social classifications [15][23]; much research has focused on the comparability of different educational qualification titles over time and between countries [3][28]; in research on ethnicity, attention has often been directed towards how alternative conceptual foundations to the measurement of ethnic groups can be realised in quantitative data analysis [2][17]. Nevertheless there have been few efforts to standardise access and exploitation of specialist information in each area, and current standards in using such data are highly inconsistent. The

GEODE project developed a system for accessing and reviewing information resources on occupational units. In the DAMES project this approach is being expanded with improved data on occupations and with new resources on educational qualifications, health and ethnicity. Services for supporting analysis of specialist quantitative data will be developed.

Other research themes within the DAMES Node concern specialist data linkages associated with the analysis of social care data, and e-Health records, and generic provisions related to the topic of 'data management' [6]. Through DAMES, interoperability across specialist datasets will be achieved in order to support the preparation and analysis of quatitative 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).

2.2.3 GEMEDA

Another relevant project in e-Social Science is GEMEDA (Grid Enabled Microeconometric Data Analysis [10]), which addressed the problem of research data availability for the economic welfare of ethnic groups within the UK. It performs micro-econometric analysis that combines data from various sample survey and census sources. This work requires operations of data virtualisation and linkage. GEMEDA used the OGSA-DAI middleware to access and transfer remotely hosted data. In addition, metadata about the datasets was collated to support effective data linking. High performance computing technology was used to distribute econometric computation, and the results were depicted visually. GEMEDA 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 the statically defined workflows in GEMEDA, demonstrating the practical application of workflows in e-Social Science.

2.2.4 Common themes in e-Social Science for Quantitative Data Analysis

GEODE, GEMEDA and DAMES, along with many other e-Science projects directed to quantitative analysis in the social sciences, have many common requirements, and have often adopted similar approaches. Prominent shared requirements include: attention to resource virtualisation; metadata; data integration; security; workflows; and high performance computing. A further common theme concerns the interface and usability aspects of e-Science services, as non-functional but importantissues.

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

2.3 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 datasets 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 [6].

A case can be made that data management tasks are ripe for support through e-Infrastructural resources. Firstly, whilst there are vast volumes of relevant quantitative data available to social scientists. A major part of a social researcher’s activities may concern identifying, linking together and manipulating different related resources. Although research data is often distributed in a standardised or semi-standardised way (for instance, the UK Data Archive offers access to survey datasets with standard formats and documentation [31]), data is made complex by heterogeneous topic coverage, the existence of many non-standardised resources, and the sheer volume of potentially relevant resources.

Secondly, a significant capacity shortfall in quantitative social science research skills is recognised in many nations [37], 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 [38]. Thirdly, 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 [39] and improved standards in documentation and replicability of analysis [40][41]. 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 operationalisations' 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, for instance, recoding complex categorical variables into smaller and more tractable range of different categories. The latter involves enhancing existing data with additional information drawn from a separate resource. For instance, 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 [13]).

The potential contribution of resources for variable operationalisations and linking data might be appreciated through use-cases. As an example, we highlight below a recent analysis of intergenerational social mobility trends [42] 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 socio-economic attainment). Although a popular and politically influential analysis, findings by [42] of declining social

mobility in contemporary Britain were criticised as highly misleading about longer term trends in social mobility in the UK [43][44][45].

- *Linking data:* [42] used data from two major UK social surveys, the birth cohort studies of 1958 and 1970. However, many other representative survey datasets also cover comparable intergenerational data. [43] and [45] linked together a wider range of other data resources to draw different conclusions on the same topic.
- *Variable operationalisations*. [42] 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. [44] demonstrated analysis of occupational data from 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 capitalise on all relevant resources. An infrastructural resource to enhance access to and linking of suitable data, and to support transparent variable operationalisations, could have improved the conduct of the above research. The different papers above all shared similar features in their activities concerned with linking data and operationalising variables. All four analyses identified and combined related data resources, and all four undertook substantial bespoke exercises in developing and analysing measures (of income and occupations). From an e-Infrastructural perspective, it is conceivable that a workflow model and record of the various choices in linking data and operationalising variables could contribute to the preservation and replicability of these complex data analytical tasks. The DAMES Node (see section 3.2 below) is directly 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.

3 An e-Infrastructure for Quantitative Data Analysis

3.1 e-Infrastructural Requirements

We define below a list of interrelated requirements and desirable features for an effective e-Infrastructures for quantitative data analysis.

- Resource virtualisation. Quantitative datasets 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 which themselves are dynamically updated. This should be an accessible service with a scalable and flexible framework to access, transport and transform virtualised 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 data itself (see section 3.2.5).

- Researchers should be able to access, manipulate and analyse quantitative data using procedures which build upon previous endeavours. This would involves researchers exploiting previous approaches, and in turn exposing their own procedures for future researchers. A workflow approach should allow the documentation and modelling of such activities.

- High performance computing may be required to raise the level of productivity for computationally-demanding quantitative data analysis tasks.

- Usability is a non-functional 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 datasets [11]. This is because the components are generic, providing data abstraction and metadata management. The components are also not bound to datasets from specific sub-disciplines of social science. The features listed above are therefore identified, emergent from [11], as 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 Hub [20] and e-Infrastructure project [48] co-ordinate approaches between initiatives and support generalisability.

3.2 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. Though these applications are specific, they exhibit common requirements and processes to a fundamental extent (as listed in 3.1). However, whilst 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 customisation 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 is inclined towards using recognised standards to achieve these requirements, in order to develop

approaches with bridge common requirements over many different applications of quantitative data analysis in the social sciences.

3.2.1 Resource Virtualisation

Virtualisation is one of the key characteristics of the e-Infrastructure underpinning the vision of interoperability. Resources in a variety of formats can be accessed via standardised protocols allowing researchers to work virtually across different formats seamlessly. Whilst data access functionalities exist in e-Infrastructure (e.g. OGSA-DAI), their usability are not fully appropriate for social science researchers as they are inclined to computer science.

Resource virtualisation 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 which will be used in the research. The NESSTAR service is one prominent existing provision in this field [21]. 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 service assisted 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 virtualisation coordinated documentation of both process is feasible.

DAMES will develop a set of quantitative data management. 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

virtualised resources. The features of each activity will contribute to a suite of middleware suitable for generic quantitative data analysis activities.

3.2.2 Metadata

Data management metadata, including instructions for recoding variables, describes data manipulations for information extraction. Quantitative social science datasets usually have a considerable quantity of metadata associated with them. They illustrate the properties of the numerical values in the data in the form of 'variable' and 'category' labels, and information about the context and provenance of the data resource.

Metadata standards have been designed to describe studies, datasets and other resources. Metadata standards define sets of elements called schemas. The meaning of the elements gives the semantics of the metadata. Metadata standards can be syntax independent or dependent. According to [33], most modern standards are syntax-dependent and are using SGML (Standard Generalized Mark-up Language) or the XML (Extensible Mark-up Language).

In the social sciences, metadata was traditionally recorded by data producers in an ad hoc fashion using a variety of non-standard techniques. Take-up of metadata standards facilitates greater data discovery and access, collaboration among researchers, and data processing capabilities. Existing standards include DDI (Data Documentation Initiative [7]), Dublin Core Element Set [9], SDMX (Statistical Data and Metadata Exchange [29]), METS (Metadata Encoding and Transmission Standard [18]) and CWM (Common Warehouse Metamodel [5]). Some of these standards have been designed to complement each other [8].

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 virtualisation process. The range of metadata covering data management activities will be defined and developed according to the design pattern of the OGSA-DAI middleware. Data virtualisation can provide access to metadata resources and integrate them with wider-ranging research activities. OGSA-DAI can virtualise resources, but cannot incorporate metadata along with the virtualised data. Certain non-grid applications (for instance the NESSTAR service associated with the European Data Archives [21]) 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 in 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 virtualisation is needed.

3.2.3 Discovery

Primarily, resource virtualisation works alongside discovery mechanisms to provide realistic use. Resources that come online may publicise themselves to peers for potential interactions. It is already possible to use existing middleware to discover social science resources quickly. GEODE, for example, uses Globus MDS4 (Monitoring and Discovery System) to build its registry of virtualised datasets. 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 trival because of the variety of data being published in, different formats, and volumes across different social science disciplines. In terms of semantics, different metadata schemes with different ontologies, taxonomies and standard schemas are used. The several metadata standards for

annotating social science resources (see section 3.2.2) 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 comparabilities 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 analyse these tables. In many disciplines, similar standards for recording certain types of data (such as standard variables) are employed. These comparabilities 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 maximise the exposure and dissemination of resources.

3.2.4 Data Integration

Activities within quantitative data analysis are certainly data-centric. Virtualisation, 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 inter-organisation data integration is linking between clinical records, patient records, disease registries, etc. to enable and support clinical trials and epidemiological studies [34]. Example of intra-disciplinary integration are the ‘cross-walks’ (data linking) occupational data resources in GEODE, and the integration of national surveys in GEMEDA.

Like resource virtualisation the data integration procedures are specified in computing terms which many social scientists cannot relate to contextually. These capabilities could be abstracted so that users can more easily relate to and specify data integration.

The DAMES project 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.

3.2.5 Security

Security is a common and critical requirement for much social science data. Apart from authentication and authorisation, existing practices for accessing data resources in social science are particularly concerned with protecting data integrity and confidentiality. There are solutions that require users to be physically present to use data that are isolated from remote access. This 'safe lab' procedure requires each user's access activities and results to be monitored and filtered to prevent improper use of resources [46]. Procedures of such a nature can be supplemented by techniques that anonymise the information to prevent, for instance, identification of an individual from the records.

Identification of users and authorising appropriate access satisfy the requirement for protecting data from improper use. e-Infrastructures have 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 [30] is an example of these technologies in action, federating security amongst numerous organisations. DAMES is using Shibboleth for several reasons. Firstly the infrastructure can manage the trust federation and security attributes interexchange between members. Secondly, Shibboleth already has a set of established procedures and software with acceptable performance. Thirdly, many potential DAMES users (e.g. social science researchers) are part of Shibboleth participating organisations. Shibboleth offers a seamless authentication and authorisation 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 *authorised* entities, challenge especially relevant whilst permitting authorised access to remote resources. Breaking through this barrier will be a major step, and will influence in a practical way the resources which are currently accessed and shared, bringing new possibilities for remote collaboration. The VANGUARD project [36] has demonstrated the possibility of enabling such collaborations under similar tight security constraints. We believe that this is an important aspect that will influence the trust and involvement of data providers in providing their assets via the e-Infrastructure. DAMES is evaluating existing approaches and techniques for data confidentiality, such as anonymisation data algorithms, and determine the feasibility of incorporating these.

3.2.6 Social Science Workflow

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 BPEL (Business Process Execution Language [47]) and WSCI (Web Service Choreography Interface), of which the most widely used is BPEL.

Workflows environments have been developed in e-Infrastructures for domains such as bioinformatics (Taverna [32]) and for scientific workflows (OMII-BPEL [25]). Taverna introduces a workflow language SCUFL (Simple Conceptual Unified Flow) 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. P-Grade [26] allows user to graphically build, execute, monitor and manage workflows via a portal interface. An advantage of P-Grade is the support of wide range of grid middleware, including legacy code. However the workflow specification is not based on standards.

In general, a workflow is built using constructs such as iteration, call-outs 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 towards social science activities would allow for building workflows, potentially contributed by users themselves.

E-Infrastructures can support workflow proliferations. DAMES will develop services for capturing social science workflows applicable to quantitative data analysis, whereby researchers can reuse existing workflows and also pro-actively contribute to the workflow pool. DAMES is inclined towards BPEL as it is an established and widely adopted standard with several implementations, including OMII-BPEL which 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.

3.2.7 High Performance Computing (HPC)

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. This allows completing tasks in 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 Sabre-R project undertaken at CQeSS included a grid-based implementation of high performance computing for computationally intensive calculations in social science applications [4].

The UK NGS (National Grid Service) has a large number of high-throughput hardware and software resources, augmented with a support framework to ensure stability. Well-developed middleware is used to access and submit high-performance requirements. These middleware such as Globus and OMII-UK are widely used and supported. This virtualises computational pool and cluster environments such as Condor, PBS (Portable Batch System), LSF (Load Sharing Facility). DAMES will develop in its framework the capability for high throughout computation, making use of existing services such as the NGS.

3.2.8 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 propose a hybrid of desktop and portal environment to cater for the spectrum of users where adaptation is minimised for the experienced, and flexible assessibility for others.

4 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, some practical examples of e-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 and moving towards 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. This will help to create e-Infrastructures for better use in quantitative data analysis: virtualisation 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. Subject to successful service delivery over the period of 2008-2011, these DAMES resources will constitute an e-Infrastructure for social science research. The inclination to use existing standards in DAMES, such as generic middleware, is an important strategy for compatibility with related e-Infrastructural resources.

Certainly the activities of the DAMES Node are by no means exhaustive in the vision of e-Infrastructure for quantitative data analysis. Iterative evaluations will be carried out to improve existing developments, driven by the needs of users. For example, new requirement might emerge for incorporating new techniques to anonymse data for confidentiality, new data and metadata management activities, and new workflow constructs. Visualisation is a substantial area that will bring value to different stages of quantitative data analysis, such as rendering results in visual formats. These and other potential developments are likely to emerge and be transformed with iterative feedback from social science users which should critically shape the future character of e-Infrastructural developments for quantitative researchers in the social sciences.

Acknowledgements

DAMES is an NCeSS Research Node funded by ESRC under grant RES-149-25-0066.

References

[1]Boslaugh, S. *An intermediate guide to SPSS programming: Using syntax for data management*. London: Sage, 2005

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

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

[4]Collaboratory for Quantitative e-Social Science, http://e-science.lancs.ac.uk/cqess/, Oct 2006.

[5]Catalog of OMG Modeling and Metadata Specifications website, http://www.omg.org/technology/documents/modeling_spec_catalog.htm, July 2008.

[6]Data Management through e-Social Science (DAMES - An ESRC Research Node for the National Centre for e-Social Science), http://www.dames.org.uk/, November 2008.

[7]The Data Documentation Initiative website, http://www.ddialliance.org/, July 2008.

[8]Gregory, A. and Heus P. DDI and SDMX: Complementary, not competing, standards, Open Data Foundation, http://www.opendatafoundation.org/papers/DDI_and_ SDMX.pdf, July, 2007.

[9]The Dublin Core Metadata Initiative website, http://dublincore.org/, July 2008.

[10]Peters S., et al. Grid Enabled Data Fusion for Calculating Poverty Measures. In Simon J. Cox, editor, Proc. 5th UK e-Science All Hands Meeting, ISBN 0-9553988-0-0, Nottingham, September 2006.

[11]Tan, K. L. L., et al. GEODE – Sharing Occupational Data Through The Grid. In Simon J. Cox, editor, Proc. 5th UK e-Science All Hands Meeting, pp 534-541, ISBN 0-9553988-0-0, Nottingham, September 2006.

[12]Foster, I. Globus Toolkit Version 4: Software for service-oriented systems. IFIP International Conference on Network and Parallel Computing, LNCS 3779, pp 2-13, 2006. Springer-Verlag

[13]Lambert, P. S., et al. Data curation standards and social science occupational information resources, *International Journal of Digital Curation,* 2(1): 73-91, 2007.

[14]Lambert, P. S., et al. The importance of specificity in occupation-based social classifications. In Robert M. Blackburn, pp 179-192, International Journal of Sociology and Social Policy, volume 28(5/6), Emerald, 2008.

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

[16]Joint Information Systems Committee e-Infrastructure Programme, http://www.jisc.ac.uk/whatwedo/programmes/programme_einfrastructure.aspx, April 2006.

[17]Lambert, P. S. Ethnicity and the comparative analysis of contemporary survey data. In Hoffmeyer-Zlotnick, J. H. P. & Harkness, J. (Eds.), *Methodological Aspects in Cross-National Research,* pp. 259-277. Manheim: ZUMA-Nachrichten Spezial 11.

[18]Metadata Encoding and Transmission Standard website, http://www.loc.gov/standards/mets/, July 2008.

[19]Birkin, M., et al. MoSeS: Modelling and simulation for e-Social Science, University of Leeds. Proc. 4th UK e-Science All Hands Meeting, September 2005.

[20]National Centre for e-Social Science. http://www.ncess.ac.uk/, July 2008.

[21]Nesstar background, http://www.nesstar.com/about/background.html, July 2008.

[22]National Science Foundation, Cyber-Infrastructure: A Special Report, http://www.nsf.gov/news/special_reports/cyber/, March 2005.

[23]Rose, D., Pevalin, D., & O'Reilly, K. (2005). The NS-SEC: Origins, Development and Use. Basingstoke: Palgrave Macmillan.

[24]Antonioletti, M., et al. The design and implementation of grid database services in OGSA-DAI. Concurrency and Computation: Practice and Experience, 17(2-4): 357-376, February 2005.

[25]Emmerich, W., et al. Grid service orchestration using the Business Process Execution Language (BPEL), Journal of Grid Computing, volume 3, pages 283-304, 2005

[26]Kacsuk, P, and Sipos, G.: Multi-Grid, Multi-user workflows in the P-GRADE Portal, Journal of Grid Computing, 3(3-4): 221-238, Springer Publishers, pp. 221-238, 2005.

[27]Research Councils UK e-Science Programme, http://www.rcuk.ac.uk/escience/, July 2008.

[28]Schneider, S.L. (Ed.), *The International Standard Classification of Education (ISCED-97). An Evaluation of Content and Criterion Validity for 15 European Countries*. Mannheim: MZES, 2008.

[29]Statistical Data and Metadata Exchange website, http://www.sdmx.org/, July 2008.

[30]Shibboleth Web Site, http://shibboleth.internet2.edu/about.html, 2008.

[31]UK Data Archive, http://www.data-archive.ac.uk/, July 2008.

[32]Turi, D., Missier, P., Goble, C., DeRoure, D., and Oinn, T. Taverna Workflows: Syntax and Semantics, 3rd IEEE International Conference on e-Science and Grid Computing, Bangalore, India, December 2007.

[33]National Information Standards Organization, Understanding Metadata,, http://www.niso.org/publications/press/UnderstandingMetadata.pdf, 2004.

[34]Sinnott, R. O., Stell, A., Ajayi, O.. Supporting Grid-based Clinical Trials in Scotland, Health Informatics Journal, Special Issue on Integrated Health Records, November 2007.

[35]OASIS Web Services Resource Framework Specifications 1.2, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf, May 2006.

[36]Sinnott, R. O., et al. Towards a virtual anonymisation grid for unified access to remote clinical data, Proceedings of the 6th International HealthGrid conference, Chicago, USA, June 2008.

[37]Bardsley, N., Wiles, R., Powell, J. L. A Consultation to identity the research needs in research methods in the UK social sciences. Southampton: National Centre for Research Methods, University of Southampton.

[38]Kohler, U., & Kreuter, F. Data analysis using Stata. College Station, Texas: Stata Press.

[39]UK Data Forum. The national strategy for data resources for research in the social sciences. Warwick: University of Warwick, http://www2.warwick.ac.uk/fac/soc/nds/, June 2007.

[40]Dale, A. Quality issues with survey research. International Journal of Social Research Methodology, 9(2), 143-158, 2006.

[41]Freese, J. Replication standards for quantitative social science: why not sociology? Sociological Methods and Research, 36(2), 2007.

[42]Blanden, J., et al. Changes in generational mobility in Britain. In M. Corak (Ed.), Generational income mobility in North America and Europe. Cambridge: Cambridge University Press, 2004.

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

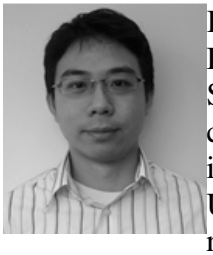
[44]Goldthorpe, J. H., & Jackson, M. Intergenerational class mobility in contemporary Britain: political concerns and empirical findings. British Journal of Sociology, 58(4), 525-546, 2007.

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

[46]ONS Virtual Microdata Laboratory, http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/business-data/vml/index.html, November 2007.

[47]OASIS, Web Service Business Process Execution Language version 2.0, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html, April 2007.

[48]Procter, R., et al. 2006. 'The National Centre for e-Social Science', in Cox, S.J. (ed.) Proceedings of the UK eScience All Hands Meeting, Edinburgh, 2006.



Koon Leai Larry Tan holds a B.Sc. (Hons.) in Software Systems Engineering. He is currently pursuing a Ph.D. in Computing Science at the University of Stirling. His research interests are focused on specification, verification and validation of grid service orchestration. Email: klt@cs.stir.ac.uk

Ken 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.

Email: kjt@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 datasets, longitudinal data and e-social science.

Email: vernon.gayle@stir.ac.uk

Prof. Sinnott is the Technical Director of the National e-Science Centre at the University of Glasgow, Scotland. He organises 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.

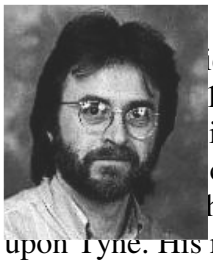
Email: r.sinnott@nesc.gla.ac.uk

Paul Lambert is a lecturer in sociology. His research interests cover the meaurement and analysis of social stratification and inequalities. His methodological interests cover the manipulation of complex social survey data.

Email: paul.lambert@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.

Email: jmb@cs.stir.ac.uk



has been a Lecturer in ience at the Unversity of 1983. He has a Bachelors ics from the University of c and PhD in Computer he University of Newcastle upon Tyne. His research interests are object-oriented methods and web and grid services.

Email: sbj@cs.stir.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. Was in the training team at the National e-Science Centre in Edinburgh. His background is in Applied Mathematics and Computer Programming/Systems administration.

Email: gcw@cs.stir.ac.uk