Guy C. Warner, Jesse M. Blum, Simon B. Jones, Paul S. Lambert, Kenneth J. Turner, Koon Leai Larry Tan, Alison S. F. Dawson and David N. F. Bell. A Social Science Data Fusion Tool and The DAMES Infrastructure, Phil. Trans. of The Royal Society, 368(1925):3859 3873, Oct 2010.

A Social Science Data Fusion Tool and The DAMES Infrastructure

By Guy C. Warner[1], Jesse M. Blum[1], Simon B. Jones[1], Paul S. Lambert[2], Kenneth J. Turner[1], Larry Tan[1], Alison S. F. Dawson[2] and David N. F. Bell[3]

[1]*Department of Computing Science and Mathematics, University of Stirling, UK*
[2]*Department of Applied Social Science, University of Stirling, UK*
[3]*Department of Economics, University of Stirling, UK*

The last two decades have seen substantially increased potential for quantitative social science research. This has been made possible by the significant expansion of publicly available social science datasets, the development of new analysis methodologies such as microsimulation, and increases in computing power. These rich resources do, however, bring with them substantial challenges associated with organising and using data. These processes are often referred to as 'data management'. The DAMES project (Data Management through e-Social Science) is working to support activities of data management for social science research. This paper describes the DAMES infrastructure, focusing on the data fusion process that is central to the project approach. It covers:

- the background and requirements for provision of resources by DAMES.
- the use of grid technologies to provide easy-to-use tools and user front-ends for several common social science data management tasks such as data fusion.
- the approach taken to solving problems related to data resources and metadata relevant to social science applications.
- the implementation of the architecture that has been designed to achieve this infrastructure.

Keywords: Social Science, Data Management, Infrastructure, Grid Technologies, e-Science

1. Background

Social scientists work with various forms of empirical data. Research using qualitative data may involve the analysis of textual, audio or visual information. There have been some initiatives in the storage and documentation of qualitative data, (e.g. Qualidata 2006; DRESS 2010), although in many qualitative studies the data exploited is not preserved beyond the project's lifespan. Quantitative data refers to information which can be represented through a structured numeric database such as the rectangular 'variable-by-case' matrix. Entries in the cells of a quantitative

database are numeric values which represent information about the subjects of analysis (the 'cases'). Social surveys are a major source of quantitative empirical data, though other research designs also generate quantitative information, including administrative records, 'born digital' monitoring data, and experiments. In contrast to the example of qualitative research, there is an extended tradition of storage, documentation and secondary exploitation of quantitative data resources (e.g. UKDA 2010; IPUMS 2010; Dale 2006).

This paper discusses the ways in which researchers exploit quantitative forms of social science data. A typical scenario would involve a researcher working with a social survey dataset that is itself composed of a series of related databases. For instance the UK's influential 'British Household Panel Survey' (ISER 2009) is a social survey collected over a 17 year period, currently supplied to users in the form of 173 related databases.

The structured, numeric character of quantitative data is suited to analysis using statistical software. Software is used both to support tasks that reorganise, restructure or adapt the data ('data management'), and to carry out statistical analysis of patterns within the data ('data analysis'). Many popular software packages include facilities for both data management and analysis, e.g. Stata (StataCorp 2010) and PASW/SPSS (IBM 2010). Nevertheless, the demands placed upon the analyst to exploit such software effectively are relatively high. They often require some programming skill, and this is particularly true if a researcher wishes to achieve a clear and reproducible specification of relatively complex data management tasks, such as linking disparate databases or combining external information resources with the current database (Long 2009). Programming requirements therefore present obstacles to data management for many researchers, and these may well hinder the exploitation of the wealth of existing quantitative data resources.

2. The DAMES Project and Data Management

The DAMES project (Data Management through e-Social Science) aims to support data management tasks involving quantitative social science data by providing services and resources which will improve the accessibility, and documentation, of those tasks. Firstly, accessibility is here used to refer to the degree to which complex data management tasks are readily performed by applied researchers. This can have very substantial implications for the results of analysis, because it is common for different social science analyses to reach different conclusions from the same data, due to differences in the way that the data was processed (see Lambert & Gayle 2009 for an illustrative example using data on educational institutions). It can be argued that these differences are minimised when researchers try out multiple different representations of their data, but in practical terms this is relatively uncommon, due to the perceived costs (i.e. inaccessibility) of generating those representations through suitable data management operations.

Secondly, good documentation improves the transparency and replicability of work with quantitative data (Freese 2007). Data management tasks typically involve enhancing the initial data resource such as by linking it with other relevant information from another database, or by restructuring the values within a variable ('recoding' and 'cleaning' variable values, often achieved by discovering and exploiting third party instructions).

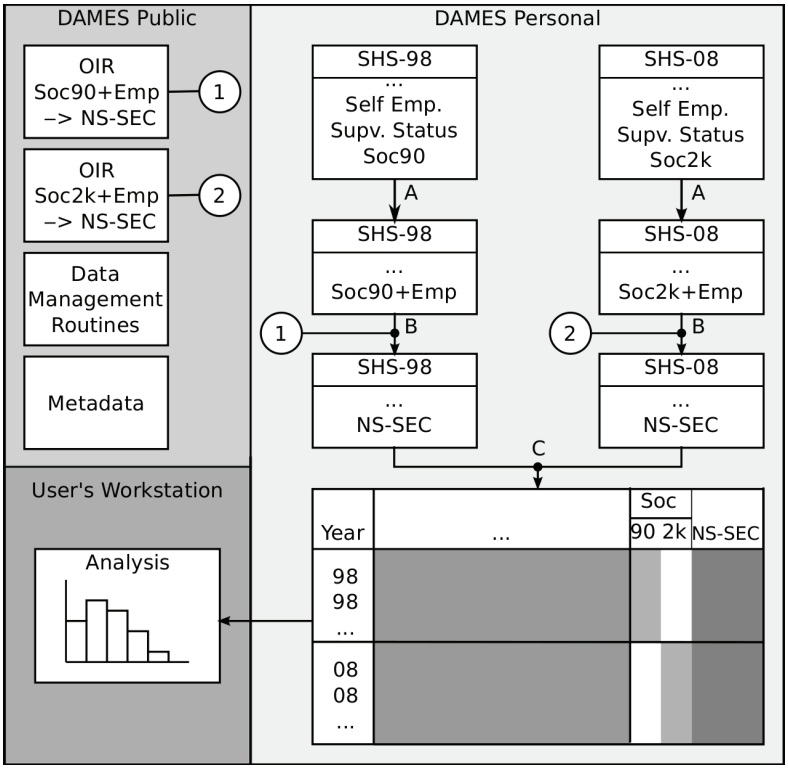

Figure 1. Scottish Heath Survey Fusion Example

The predecessor of DAMES was the GEODE project (Grid Enabled Occupational Data Environment (Tan *et al.* 2006)). Further development of the GEODE portal has been integrated within the DAMES infrastructure. GEODE supports the analysis of occupational data by providing a portal and application service environment for the curation, publication, discovery, and matching of occupational data resources. Resources are curated using the Data Documentation Initiative (DDI 2010) standard as the schema for resource metadata to facilitate resource discovery and data linking. This service provides a means and framework to share resources about occupations. GEODE has an occupational matching service that performs data linking between its resources and a user's own micro-social survey datasets. This allows users to easily enhance their survey data with additional relevant information linked to occupations.

3. Illustrative examples

Social science researchers can benefit from the DAMES project infrastructure by using it to enhance the data resources with which they are working. Their original data resource, such as a micro-social survey dataset featuring individual level details on each respondent, would typically contain a range of relevant measures that might benefit from further treatment such as processing of missing values, standardising measures, or linking with external data (i.e. data management tasks).

A typical usage scenario is described in Figure 1. It depicts the exploitation of externally provided information about occupations in order to enhance the data files used by the researcher for their own analysis. In this scenario, the user begins with two micro-social survey datasets, namely the Scottish Health Surveys of 1998 and 2008 (Joint Health Surveys Unit 2001, and Scottish Centre for Social Research 2010), which may have been downloaded via the UK Data Archive. In the example, the original survey data has detailed occupational unit group information coded according to the UK Standard Occupational Classifications from 1990 and 2000 respectively for each year. The original survey also contains data on the supervisory status and self-employment status of each respondent in their job (if they are working). These are indicated in the figure by the original variables 'Soc90', 'Soc2k', 'Self Emp' and 'Supv Status'. In the example, the user would like to carry out an analysis which compares the relative influence of occupational positions on various other measured variables (e.g. measures of health and well-being) over the two time periods.

The figure depicts several ways in which the DAMES infrastructure can assist the user in enhancing their original survey data. An expert in occupational data would know that for the UK, the NS-SEC (National Statistics Socio-Economic Classification, see Rose and Pevalin 2003) provides an appropriate tool for the analysis of occupations over both time periods. Were the researcher not to be aware of this, however, the DAMES portal includes search routines which allow them to discover the occupational information resources (OIRs) within the publicly accessible data files held on DAMES (depicted as (1) and (2) on the left pane of Figure 1).

Next, the researcher would need to manipulate and merge their two survey data files in order to link them with the occupational information resources provided through DAMES. This is a three stage process. Firstly, depicted as (A), the researcher must adapt their own measures in order to ensure they share a common coding frame to those on the DAMES resources (here, this involves using the original measures of self-employment and supervisory status to derive a new harmonised measure of employment status, referred to as 'Emp'). DAMES provides metadata on the required measures to facilitate this transformation, as well as an environment for the researcher to record the transformation they carried out. Secondly, depicted in the right pane as steps (B) and (C), the researcher can draw upon purpose-built data fusion routines to merge their own micro-data, securely, with the resources at DAMES, and with each other. In combination, these generate the new enhanced data resource, depicted at the end of the figure, which features NS-SEC for both surveys, and has combined the surveys, alongside descriptive metadata on the file matching process which itself provides a replicable log of the data recoding and fusion processes undertaken.

The example of enhancing occupational data is not the only relevant data fusion activity available in the DAMES service, though it is a particularly common and important example (in fact, an extended description of using the GEODE service to link data with the NS-SEC social classification can be found in Lambert 2007). Other scenarios catered to by the DAMES infrastructure include linking data resources concerning educational qualifications and measures of ethnicity, and several more specific applications concerned with fusing pre-arranged datasets in the domain of social care and health research. It has been argued elsewhere by the authors that such data enhancements are very typical of quantitative data analysis in the

social sciences, but have nevertheless not been well addressed previously (Lambert *et al.* 2009).

In general, data fusion is defined as linking data resources according to a selection of deterministic and/or probabilistic criteria. A deterministic criterion is linkage of related cases (one-to-one, one-to-many or many-to-many). A probabilistic criterion involves the imputation of values or variables on shared, correlated characteristics. Software tools to undertake both deterministic and probabilistic data fusion are already available. However, these are not widely exploited by social scientists and approaches tend not to be shared. The DAMES infrastructure is therefore designed to make social science data fusion more accessible, more widespread, and better documented.

Whilst the above examples concern ways in which the DAMES infrastructure can be used to enhance data resources for the benefit of analysis, there is also a second important way in which social science researchers can use the DAMES resources. These are researchers who have themselves generated new data resources which they wish to disseminate to a wider research community. This is a surprisingly common scenario, since relevant data resources (such as new derived datasets or analytical command records) are a by-product of many empirical research projects. There is widespread willingness amongst social scientists to share derived resources for research purposes, but hitherto there have been few facilities in place to support sharing of collaborative resources in a systematically structured manner – see Treiman (2009), though compare IDEAS (2009). For such requirements, the DAMES services allow simple *pro forma* registration and uploading of relevant data resources. They also allow the construction of comparable DDI format metadata to allow effective registration and dissemination of such resources. The DDI format of metadata is introduced in §5a. Example scenarios include the derivation matrices for the new European Socio-Economic Classification scheme (Rose & Harrison 2009) which are now available via the GEODE portal, and the dissemination of command files to undertake an analysis associated with a particular research publication.

4. Designing the Architecture

(*a*) *Social Science Goals*

To improve methodological standards within the social sciences, DAMES is developing services that are intended to facilitate the accessibility and documentation of data management tasks using quantitative datasets. Accessibility is important to encourage larger numbers of researchers to exploit data resources to their full potential. It is widely argued, for instance, that many analysts shy away from linking or transforming their data in ways that would in principle be appropriate. Typically this is because they do not feel they have the requisite technical expertise to undertake the appropriate tasks (Lambert *et al.* 2009). The goal of accessibility is achieved by adopting the key principle that the services should be readily communicated, understood, and enacted by empirical researchers in the field. This requires compatibility with existing formats and terminology used in social science (such as those of mainstream quantitative data analysis packages Stata and PASW).

Another contribution of DAMES is in documentation of data and associated data resources. This involves firstly collecting suitable metadata to describe the data resources used in quantitative research, and secondly describing relevant enhancements that may have been made to that data (i.e. as a result of data management). Steps towards the former objective have been made outside DAMES, most notably in the standardised documentation requirements and approaches employed by data archives such as the UKDA. These define the preparation and storage of original materials (van den Eynden *et al.* 2009), and the growing adoption of DDI for organising metadata about such data resources (Vardigan *et al.* 2008). The contribution of DAMES is in supporting the collection of comparable levels of standardised documentation for supplementary data resources (e.g. information files that can be used to enhance the analysis of a survey dataset), and in supporting a clear record of data enhancements made in response to these resources. This has been achieved by developing metadata oriented tools that make it easy to supply, store and organise appropriate metadata. These tools also allow other researchers to search for and retrieve data. Providing metadata in this way contributes to a desirable 'virtuous data cycle' whereby the metadata that serves as research documentation is itself an intrinsic part of the day-to-day research process.

Data security is also a key goal of DAMES, though it is not discussed further in this paper. Data resource security is preserved, but DAMES should not take over this role from the data provider. The DAMES infrastructure must work with existing standards and technologies already used by the data providers.

(*b*) *Design Constraints*

The research goals for DAMES result in a series of constraints on the overall design. In understanding these constraints it is worth re-iterating that the infrastructure has a focus on pre-analysis of data, so there is a built-in assumption that the data generated will be consumed by further analysis tools.

The first and most important constraint is that new users to the system should have a minimal learning curve. This is approached by designing a simple to use portal that integrates with the user's desktop environment irrespective of its operating system. Social scientist researchers typically have favourite tools, so it is vital that the infrastructure supports these tools as transparently as is practical. Whilst this could be accomplished by developing plugins for a fixed selection of tools, this would be impracticable for every conceivable data analysis tool and require an ongoing and labour-intensive commitment. For this reason a generic, application-independent solution is needed.

A further constraint is that the infrastructure be accessible programatically. Whilst it is not viable for DAMES to create plugins for other applications, it is important to allow other projects/application developers (e.g. NeISS (2010) and Obesity e-Lab (2010)) to generate these plugins themselves. It is envisaged that other social science projects with a focus on further data analysis (beyond the pre-analysis addressed by the DAMES infrastructure) will programatically access the infrastructure for their pre-analysis. This allows for a significantly larger research impact than can be achieved by DAMES alone. In satisfying this constraint it is important (when possible) to be standards-compliant and programming language neutral. This constraint also exists in the opposite direction. The infrastructure

has to support a user who needs to access tools that exist within an external infrastructure or project. The most likely example is where a user needs to run message-passing or parallel algorithms over a very large data set. In this scenario, sending the job to a grid such as the UK's National Grid Service (NGS 2010) is the best solution. This therefore requires that the policies (security and usage) and technologies of the grid be supported.

The final constraint is that the workflows of user customised statistics modules be supported. Whilst the infrastructure does not allow typical users to construct their own routines, it is impossible (without limiting the usability of the infrastructure and therefore the impact) to provide every possible routine. Instead predefined modules are provided that the user can incorporate into a workflow with customisable parameters for each module. The fusion tool that will be introduced later is an example client for this workflow model. The workflow approach needs to be reasonably generic so as to be future-proof. A supplementary constraint is that there needs to be a list of social scientist 'power users' to create these modules. This constraint is also extended to support 'fair-share', so that no single user is able to use the entire infrastructure's resources at the expense of other users.

(*c*) *Design Principles*

The traditional approach to grid computing is job focused. By this it is meant that the focus is on "what the job does" with a dependency on "what data is required to achieve this". In DAMES the opposite approach has been taken, focusing on "what the data is" with a dependency on "what job is required to generate this". This allows a more dynamic approach to data, relevant to the many social science data resources which are regularly updated.

A second design principle comes from the definition of data management (see §2). Data management decisions are made by the social scientist, whereas the enactment of these decisions is automated. The enactment of a data management decision is thus a job management process.

When taking these two principles together, it is clear that an architecture focused on job management is not appropriate since this does not allow for the starting point of 'what the data is'. Equally, a purely web service oriented architecture is not appropriate. A key component of the architecture is the ability to transfer data (typically in the 10MB–100MB size range) from the server to the social scientist's workstation. Data transferred through web services has to be carried as attachments to SOAP (W3C, 2007) messages, since SOAP messages are written in XML. The overhead of this substantially slows down data transfers.

(*d*) *Components of The Infrastructure*

Considering the constraints and principles together led to the identification of a number of fundamental components for the infrastructure. These components fall into two categories, namely user interface components and service components. The user interface components are:

- a web portal designed to support user-configured workflows and encourage good data management

- an interface for supporting easy access to the data resources from the user's workstation.

The service components are:

- a filestore for storing user's personal resources and shared resources. This filestore also needs to provide an interface for fast, efficient and secure data transfers to the user's workstation.
- a job management system for running the pre-analysis tools
- a database for storing the DDI 2/3 metadata
- a standards-compliant interface for accessing the infrastructure.

5. Interfaces and Good Data Management

The essential feature of good data management is keeping a comprehensive, or at least sufficient, metadata record of what was performed. Within the social science community, DDI 2 and DDI 3 are commonly used meta-data standards.

(*a*) *DDI 2 and 3*

The need for scientists (and social scientists in particular) to exchange data is widely recognised. Exchange is facilitated by the existence of *metadata* — additional descriptive information attached to a data resource to enhance its interpretation and use. Metadata typically provides information about the authorship, purpose, origin and format of a data resource. The Data Documentation Initiative project (DDI 2010) has specified international XML-based standards for the content, presentation, transport, and preservation of metadata for datasets in the social and behavioural sciences (Vardigan *et al.* 2008). The DDI specification greatly increases the scope and formality of the traditional electronic 'codebook', promoting interoperability between systems and enabling the creation of compatible tools. Further benefits of DDI include facilitating the re-purposing of data, enhancing opportunities for data discovery and enabling more effective preservation due to the non-proprietary nature of DDI.

The DDI standard has now been through three versions. DDI 1 and DDI 2 were basically designed to emulate early codebooks, with DDI 1 being intended for documenting microdata surveys, and DDI 2 containing some added support for aggregate tabular data. The metadata elements in these versions of the standard cover document description, study description, data file description, variable description, and other study-related materials.

DDI 1 and DDI 2 were found to be insufficiently flexible. They were intended to document the endpoint of research, and the information was hard to re-use for new research. DDI 3, published in 2008, is designed to cover all stages in the life cycle of a data collection: from the formulation of research questions, to data collection, to publication and dissemination. The additional metadata included in DDI 3 covers study concepts, data collection, data processing, data distribution, data archiving, data discovery, data analysis and re-purposing.

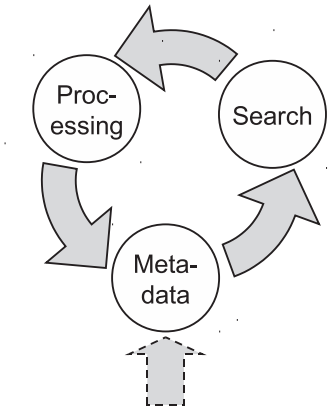

Figure 2. The Virtuous Data Cycle

The DAMES project has a significant need to capture life-cycle information about curated and generated datasets, so the infrastructure makes use of DDI 3. To illustrate this, reconsider the social science workflow example, Figure 1: Each of the survey datasets (SHS-98 and SHS-08) will have been associated with metadata when they were curated — in particular identifying the Self-employment status and Supervisor status fields. The initial derivation step, in which the Employment status field is added, identified the appropriate fields in SHS-98 and SHS-08 to combine from examination of the metadata for those datasets. This combination is arrived at by comparison with the metadata on the required variable (employment status) which is held on the DAMES system. The derivation process associates extended metadata with the resultant datasets, and may record, in the context of DDI 3, the identities of the originating datasets and the transformation process carried out. The recoding steps labelled (B), and the final fusion step (C) in the workflow, examine and generate metadata in an analogous way — so each dataset is "shadowed" by metadata that describes not only the intrinsic properties of the dataset itself, but also the relationships between the datasets.

(*b*) *The Data Cycle*

With sufficient metadata it is possible to see how sharing data now leads to the 'virtuous data cycle' as illustrated in Figure 2. Each time a new data resource with metadata is shared with other users, the 'ecosystem' of shared resources becomes richer. This means that a subsequent user searching for a resource is more likely to find a suitable one. Without the resource metadata, this search can only find resources by file name and so may miss relevant resources. The enhanced search for resources makes it more likely that users can achieve their goals and generate new results and metadata. Adding these results (and associated metadata) to the ecosystem, continues the data cycle. The fusion tool is an example of this approach. The richer the ecosystem of resources, the easier it is to find a suitable donor data resource. When the fusion tool has generated a new resource, new metadata is generated based on the instructions given to the fusion tool and the metadata of the parent donor and recipient data resources.

(c) *Interfaces*

In response to social science needs, DAMES services have been developed within an interface which reflects feedback from social scientists involved in the kinds of analysis required. As previously stated, the user front-end is designed to minimise the new skills a new user has to learn. It is assumed that the user is familiar with web browsing and network drives, but nothing more. This is achieved by the use of a portal working in conjunction with a WebDAV (IETF 2003) interface. WebDAV is a protocol that allows online resources to appear as a normal network drive on a user's workstation, and is supported by all major, current, operating systems. Hence the DAMES services can provide direct access via a network drive on the user's workstation to all permitted resources and to the outputs of the tools in the portal. This way, users can open resource and results in their (appropriate) favourite tools just as they are used to. Both modes of user interaction with the infrastructure are provided, although users and other projects can use their own modes and tools. In common with many science projects, the portal is the primary access point for using the DAMES tools. For this reason, the paper does not focus on the portal itself but rather on the fusion tool part of the portal and what it aims to achieve.

The portal is designed to enable and encourage the sharing and re-use of resources in keeping with the 'virtuous data cycle'. The fusion tool fits into the data processing category. A typical scenario is where a user wishes to fuse a public data resource with a privately held or generated one. The fusion tool initially allows the user to select or search for the donor and recipient data resources. The subsequent stages allow the user to identify common variables between the data resources, variables to be imputed by specified methods, and appropriate data fusion methods. Finally, the description of the fusion process is passed to the DAMES infrastructure for execution and generation of new metadata. The description and execution of this workflow is described in §6b. The portal also contains a curation tool that encourages users to share data resources and to generate the metadata needed for the virtuous data cycle. These tools are portlets following the JSR 168 standard (JCP 2003).

6. Service Components and Technologies

(a) *Filestore*

The role of the filestore within the DAMES infrastructure is more than to just store a user's personal data resources and shared (group or all-user) resources. It must also be able to support the following:

1. Running an automated action when a new data resource is uploaded, an existing data resource is modified or a data resource is read.
2. Fast, efficient (parallel) transfer of data. This is essential when supporting access by other projects.
3. Interfaces support multiple programming languages, notably C and Java.
4. A user account system consistent with other user interface components.

The technology chosen to satisfy all of these requirements is the integrated Rule Oriented Data System (DICE 2009). iRODS is the successor to the Storage Resource Broker (DICE 2009). Whilst many filestore technologies support requirements 2 to 4 the support of requirement 1 by iRODS was the main reason for choosing it. A key component of iRODS is the rule system. The majority of actions in the filestore, such as putting a new data resource into the filestore, will trigger the execution of a rule. These rules are basic workflows that allow multiple 'microservices' to be run, depending on particular conditions being met. iRODS conditions run defined rules, for example after a new file is uploaded. A microservice is a modular unit of server-embedded code that may access information about the user, the current state of the filestore, and the actual data being transferred. The main problem with microservices is that they are created in the C programming language and require the server to be restarted when any changes are made. Within DAMES this difficulty has been circumvented by using the C application programming interface (API) to Perl (a scripting language that does not require pre-compilation) so as to dynamically load units of Perl code from within the microservice. An example of a rule could be making an additional replica of the data resource, based on who is transferring the file and how much disk space they are already using.

Access to iRODS is provided by APIs in C, Java and PHP. The Java API is called Jargon and is deliberately used in a similar way to standard Java file access. By default, iRODS also provides a series of command-line Unix clients called the icommands that use the C API. To support WebDAV access, an external tool called Davis (ARCS 2010) is deployed.

(*b*) *Job Management*

A key feature in the design of the infrastructure behind the user interfaces is the treatment of social science tasks in terms of grid jobs. By treating all tasks (such as running a method or accessing a remote data resource) as grid jobs, it is possible to ensure not just a fair share of computing resources but also access to remote computing resources such as the NGS.

The job management system chosen for the DAMES infrastructure is Condor (2010). Good customisability and support for DAGs (Directed Acyclic Graphs) were the prime reasons for choosing Condor. Whilst it is possible to store the master definition of the workflow in terms of Condor submit files, it is better to use a standards complaint language. Job Submission Description Language (JSDL) is an XML schema from the Open Grid Forum (OGF 2010) that is already in use by the NGS portal. The only disadvantage of using JSDL is that it is designed for a single job and not for a workflow. The initial plan was to run an executable that would generate and submit an appropriate DAG. However it was found that easily defining the parameters for this executable became unmanageable. An alternate solution was arrived at by extending JSDL to directly support workflows.

(i) *JFDL (Job Flow Description Language)*

As described above, DAMES services require interactions between numerous jobs to complete their processing. For example, a service for fusing two datasets accessed from different external databases might require jobs for staging in each

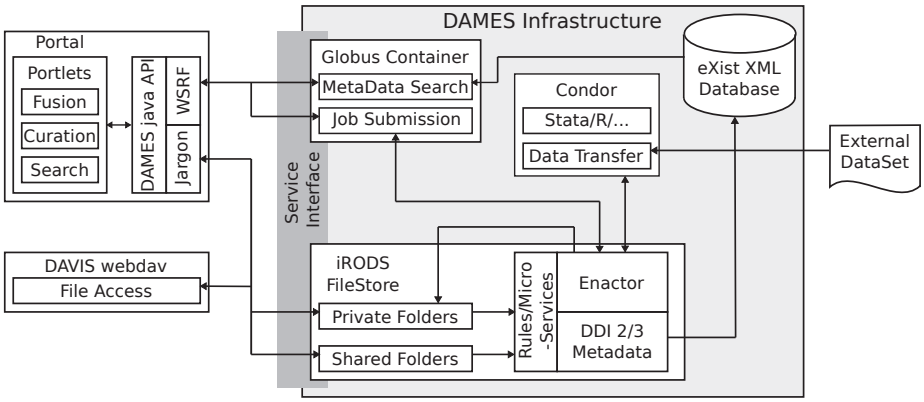

Figure 3. The Architecture of The Infrastructure

dataset, imputing variables, mapping variables, and fusing the data. Some of these jobs could run in parallel, while others might depend on the results of earlier jobs. The solution adopted was to describe service job flows by embedding elements from the new, purpose-designed JFDL (2009). JSDL describes service provision issues and data staging, while JFDL describes the relationships between datasets and the transformational tasks needed to realise DAMES services.

A significant benefit of this approach to workflow description is that when DAMES services succeed in outputting datasets created from JSDL/JFDL inputs, the JSDL/JFDL instructions are metadata describing how the output datasets were created. The DAMES infrastructure uses XSLT (a method of transforming one form of XML into another form) to create DDI 3 metadata from JSDL/JFDL instances along with DDI 3 data describing user and job profiles. Following translation, the DDI metadata is stored in the metadata database for future searches by researchers. In addition, the JSDL/JFDL file is stored so that the job can be re-run if necessary.

(*c*) *DDI 2/3 Metadata Database and the Service Interface*

Metadata and services use technologies in a fairly standard way. The metadata is defined by an XML schema. Searching the metadata for a suitable data resource will need to respect and understand this schema (as opposed to converting it to tables and relationships). The obvious choices for searching metadata are therefore the XPath and XQuery XML search languages. The technology used for the metadata database also needs to support easy uploading and downloading of the metadata. The eXist database program (2009) supports XPath and XQuery searches, accessing the stored XML files as if they were a web page. For this reason, an unmodified eXist database is suitable for the metadata database. Within the grid community, a commonly used approach for providing the service interface is the Globus Toolkit (2009) that implements WSRF (OASIS 2010) standards.

7. Architecture

Now that the components of the infrastructure and the reasons for them have been explained, the architecture of the infrastructure is simple to understand. The

design is built around the iRODS filestore as the central component. In particular a microservice within iRODS and a series of rules are collectively called the 'enactor' since they enact the data management decisions. The microservice:

1. transfers any data resources, such as the donor and recipient data resources needed in the fusion process, into a temporary working directory. If one of these resources is in turn dynamic and described by a JSDL, a recursive call to the enactor is made.
2. converts JSDL/JFDL into a series of Condor submit files and a Condor DAG
3. submits the DAG and waits for it to complete
4. transfers the results back into the user's private space in the filestore
5. updates a relational database with the current status of the job (not shown in Figure 3).

The rules in the enactor also handle:

- synchronising any uploaded metadata into the eXist database
- converting JSDL/JFDL files into DDI 3 descriptions of jobs and generating HTML summaries.

The iRODS filestore also provides the (potentially distributed) storage of the DAMES public resources and a users personal files as was shown in Figure 1.

The Condor jobs that are created by the enactor fall into two categories. The first category of jobs is running statistics applications on the server as a node of a workflow. These statistics applications are run as conventional grid jobs and hence must be executable as non-interactive commands.

The Globus container contains two primary services. One service provides a wrapper to the queries sent to the eXist metadata database, while the other provides an interface to manually submit jobs (as opposed to dynamic file job submission) and monitor their status. To simplify development of client tools, a Java API was created to handle all calls to Globus services and to wrap the Jargon API (effectively masking unsupported authenticated mechanisms). The fusion tool uses this API to find the available data resource. Once the JSDL/JFDL file has been created, this is stored in the user's private space in the filestore prior to job submission.

8. Conclusion

The DAMES project is addressing the needs of social science researchers for support in exploiting the substantial body of publicly available social science datasets. There are significant challenges in effective use of these datasets, which can lead to under-utilisation of these important resources. In a recent report (ESRC 2010), it is observed that academics "make less than full use" of datasets to which they have access, and identifies a "call for additional efforts to bring quantitative research methods into closer alignment with institutional and state-of-the-art standards".

The DAMES infrastructure is being developed during the period 2009–2011. Parts are using internationally recognized standards and making innovative usage

of externally developed components. As these parts of the infrastructure become available, they are made accessible via the DAMES portal.

The success of the infrastructure in supporting social science data management will ultimately hinge upon uptake by non-specialist researchers. A series of outreach events is taking place during 2010 and 2011 to promote the DAMES resources and to train non-specialist researchers in better standards for data management. Through these developments and activities the DAMES project is addressing a need clearly identified in the report quoted above.

The DAMES project is supported by the UK ESRC, grant reference RES-149-25-1066.

References

ARCS, Australian Research Collaboration Service, Davis WebDAV interface to iRODS See https://projects.arc.org.au/trac/davis/wiki/WikiStart, last updated 25th November 2009.

Condor project from the University of Wisconsin. See http://www.cs.wisc.edu/condor/, consulted 11th January 2010.

Dale, A. 2006 Quality Issues with Survey Research. *Int. J. Soc. Res. Meth.* **9(2)**, pp. 143-158.

DAMES, Data Management through e-Social Science. See http://www.dames.org.uk, last updated September 2009.

DDI, The Data Documentation Initiative. See http://www.ddialliance.org, consulted 11th January 2010.

DICE, Data Intensive Cyber Environments. Integrated Rule Oriented Data System. See https://www.irods.org, last updated 6th November 2009.

DICE, Data Intensive Cyber Environments. Storage Resource Broker. See http://www.sdsc.edu/srb/index.php/Main_Page, last updated 8th October 2009.

DRESS, Digital Records for e-Social Science. See http://web.mac.com/andy.crabtree/NCeSS_Digital_Records_Node/Welcome.html, consulted 11th January 2010.

ESRC, Economic and Social Research Council , International Benchmarking Review of UK Sociology. See http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/Sociology, edited by ESRC, BSA, HaPS, March 2010

eXist Open Source Native XML Database. See http://exist.sourceforge.net/, last updated 2nd December 2009.

Freese, J. 2007 Replication Standards for Quantitative Social Science: Why Not Sociology? *Soc. Meth. Res.* **36(2)**, pp. 153-171.

Globus Toolkit from the Globus Alliance. See http://www.globus.org, last updated 22 October 2009.

IBM's Predictive Analytics Software from SPSS. See http://www.spss.com/statistics, consulted 11th January 2010.

IDEAS at Research Papers in Economics (RePEc) project. See http://ideas.repec.org, last updated 31st December 2009.

IETF, Internet Engineering Task Force Web Distributed Authoring and Versioning (WebDAV) Working Group. See http://ftp.ics.uci.edu/pub/ietf/webdav/, last updated 3rd July 2003.

IPUMS, Integrated Public Use Microdata Series from the University of Minnesota. See http://www.ipums.org, consulted 11th January 2010.

ISER, Institute for Social and Economic Research at the University of Essex 2009 British Household Panel Survey: Waves 1-17, 1991-2008 [computer file], 6th Edition. Colchester, Essex: UK Data Archive [distributor], May 2009, SN 5151.

JCP, Java Community Process JSR 168 Portlet Specification, final release 27th October 2003. See http://www.jcp.org/en/jsr/summary?id=168.

JFDL, Job Flow Description Language from the DAMES project June 2009. See http://wwww.dames.org.uk/schemas/jfdl/2009/06.

Joint Health Surveys Unit of Social and Community Planning Research and University College London, Scottish Health Survey, 1998 [computer file]. Colchester, Essex: UK Data Archive [distributor], July 2001. SN: 4379.

Lambert, P. S. 2007. An illustrative guide: Using GEODE to link data from SOC-2000 to NS-SEC and other occupation-based social classifications, Edition 1.1. Stirling: GEODE Project Technical Paper No. 2, University of Stirling. See http://www.geode.stir.ac.uk.

Lambert, P. S. & Gayle, V. 2009 Data management and standardisation: A methodological comment on using results from the UK Research Assessment Exercise 2008. Stirling, University of Stirling: Technical Paper 2008-3 of the Data Management through e-Social Science Research Node.

Lambert, P. S., Gayle, V., Bowes, A., Blum, J.M., Jones, S.B., Sinnott, R.O., Tan, K.L.L., Turner, K.J. & Warner, G.C. Standards setting when standardising categorical data. Paper presented to the *Fifth International Conference on e-Social Science, Cologne, Germany, 24-26 June 2009*. See http://www.dames.org.uk/publications.html.

Long, J. S. 2009 *The Workflow of Data Analysis Using Stata*. Boca Raton: CRC Press.

NGS, National Grid Service. See http://www.ngs.ac.uk, last updated 8th January 2010.

NeISS, National e-Infrastructure for Social Simulation project. See http://www.geog.leeds.ac.uk/projects/neiss/about.php, consulted 11th January 2010.

OASIS, Advancing Open Standards for the Information Society Web Services Resource Framework. See http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf, consulted 11th January 2010.

Obesity e-Lab project at the University of Manchester. See https://www.nibhi.org.uk/obesityelab/default.aspx, consulted 11th January 2010.

OGF, Open Grid Forum Job Submission Description Language Working Group. See http://forge.gridforum.org/projects/jsdl-wg, consulted 11th January 2010.

Qualidata from the Economic and Social Data Service. See http://www.esds.ac.uk/qualidata, last updated 6th September 2006.

Rose, D., & Pevalin, D. J. (Eds.). 2003. *A Researcher's Guide to the National Statistics Socio-economic Classification*. London: Sage.

Rose, D. & Harrison, E. 2009 *Social Class in Europe: An introduction to the European Socio-economic Classification*. London: Routledge.

Scottish Centre for Social Research, University College London. Department of Epidemiology and Public Health and Medical Research Council. Social and Public Health Sciences Unit, Scottish Health Survey, 2008 [computer file]. Colchester, Essex: UK Data Archive [distributor], March 2010. SN: 6383.

StataCorp. Stata data analysis and statistical software. See http://www.stata.com, consulted 11th January 2010.

Tan, K. L. L., Gayle, V., Lambert, P. S., Sinnott, R. O. & Turner, K. J. GEODE – Sharing occupational data through the grid. *Proceedings of Fifth UK e-Science All Hands Meeting, September 2006*, pp. 534–541.

Treiman, D. J. 2009 *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco: Jossey-Bass.

UKDA, UK Data Archive. See http://www.data-archive.ac.uk, consulted 11th January 2010.

van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. 2009 *Managing and Sharing Data: A best practice guide for researchers*. Colchester: UK Data Archive, University of Essex.

Vardigan, M., Heus, P. & Thomas, W. 2008 Data Documentation Initiative: Towards a Standard for the Social Sciences. *Int. J. Dig Cur.* **3(1)**, pp. 107-113.

W3C, SOAP Version 1.2. See http://www.w3.org/TR/2007/REC-soap12-part1-20070427, last updated 27th April 2007.