Kenneth J. Turner, Koon Leai Larry Tan, Jesse M. Blum, Guy C. Warner,
Simon B. Jones and Paul S. Lambert. Managing Data in E-Social Science,
in Robert Bestak, Laurent George, Vladimir S. Zaborovsky and Cosmin Dini (eds.),
Proc. 8th Int. Conf. on Networks, pp. 214-219, IEEE Computer Society, March 2009.

Managing Data in E-Social Science

Kenneth J. Turner Koon Leai Larry Tan Jesse M. Blum Guy C. Warner
Simon B. Jones Paul S. Lambert

Computing Science and Mathematics / Applied Social Science, University of Stirling
Stirling, Scotland, FK9 4LA
{kjt,klt,jmb,gcw,sbj}@cs.stir.ac.uk / paul.lambert@stir.ac.uk

Abstract

Grid computing is moving from its original focus on the physical sciences to other disciplines such as the social sciences. The orientation of these newer applications is on data management rather than processing. This paper describes how the DAMES project (Data Management through E-Social Science) is developing grid-based solutions for handling data in a distributed environment. The paper describes the approach being taken to meet key challenges: metadata for effective use of datasets, and data-oriented workflows for e-social science.

1. Introduction

1.1. Grid Computing

The idea of grid computing is to provide computing resources on demand much as the electricity grid provides electrical power. Grid computing has emerged as a major paradigm for sharing networked resources: processing power, data storage, scientific instruments, etc. Much of the early deployment of grids was focused on the needs of large-scale scientific experimentation. In the physical sciences, huge amounts of data are often collected and subjected to highly intensive calculations. Examples of grid computing can be found in applications such as astrophysics, biology, the earth sciences and particle physics.

However the use of grid computing is gradually moving towards other sciences and into research generally. Applications can now be found in business, economics, medicine and the performing arts. This paper reports on work being undertaken to develop grid techniques for social science applications, particularly for analysing survey data in sociology and economics. The requirements of these new applications are rather different from the physical sciences. Datasets and their processing tend to be much more modest in volume, but managing distributed data becomes a much larger issue. Particularly in social science, aspects like confidentiality and security of data become very important. Although such applications tend not to demand high-performance computing, they can exploit many useful features of grids such as the following.

Virtual organisations allow groups of individuals to work together in an *ad hoc* but controlled manner. For example, a group of social scientists in different institutions might collaborate to analyse distributed datasets for trends important to society (e.g. changing patterns of occupation).

Grid portals provide a non-technical way for users to access shared resources. Mechanisms such as single sign-on require users to authenticate only once, allowing use of resources anywhere within the grid. As an example, social scientists can share occupational resources through a portal.

Virtual resources allow access to networked resources irrespective of their format or location. Using distributed data can be a major issue for non-technical users, who have limited interest in the computing details. A portal can, for example, allow social scientists to use occupational datasets and classification schemes without having to worry about which databases, analysis packages and hosts are used.

Security mechanisms for access control, authentication and authorisation ensure appropriate use of networked resources. For example, a social scientist (or a research team) might be authorised to read but not modify a particular survey. A finer degree of control can also be exercised, e.g. allowing access to only anonymous aspects of data.

1.2. E-Social Science

Grid computing has given a strong impetus to all fields. This paper deals with its application to e-social science. However, grid computing is only one aspect of what is called the e-infrastructure. Many aspects of ICT (Information and Communication Technologies) have a role to play.

Distributed data is common in social science. Key issues include data curation, data management, distributed access, platform and location independence, confidentiality,

and access control. The importance of data management has been widely recognised. For example, the UK Data Archive (*www.data-archive.ac.uk*) is responsible for curating a vast range of data from the social sciences and the humanities. NESSTAR (*www.nesstar.com*) provides facilities to enable easy access to a wide variety of social survey datasets. Geographic and mapping datasets are maintained by EDINA (*www.edina.ac.uk*). The GEODE and DAMES projects discussed in this paper are linked with NCESS (National Centre for e-Social Science, *www.ncess.ac.uk*).

Two key themes in management of distributed data are considered in this paper: metadata (the provenance and technical characteristics of a dataset), and workflows (managing complex flows of data-oriented activities).

The authors are part of a team that previously worked on the GEODE project (Grid-Enabled Occupational Data Environment, *www.geode.stir.ac.uk*). This project exploited grid techniques for better use of occupational datasets and classification schemes [7]. The narrow focus of this project on occupation has been significantly expanded in new work on the DAMES project (Data Management through E-Social Science, *www.dames.org.uk*).

The goals of DAMES are to support researchers in creating, managing and using distributed data – particularly in social science applications. The project is tackling a number of themes such as grid-enabling social science datasets, linking and fusion of datasets, and data resources for microsimulation and surveys. Within DAMES, the authors are developing techniques for metadata, data abstraction and workflow models for e-social science.

2. A Data Management Architecture

2.1. The GEODE Data Portal

As an example of the approach being taken for data management, figure 1 [7] shows the data portal architecture developed by the GEODE project.

Some social scientists may wish to deposit their datasets with the portal (the local data depository in figure 1). For example, the data may be fully in the public domain because its creation has been funded by a public body. This has the advantage that data management can then be left to the portal administrator. The data originator may well have little interest or skills in computing.

However, this is not always possible. For example, the data may not be fully public or the owner may prefer to have full control over it. In these situations, only a link need be created to the dataset (the external data depository in figure 1). However, the owner must then manage the data.

In either case, the portal provides guided assistance to non-specialists in creating metadata. For both local and external datasets, a data service supports platform-independent and location-independent access. This can be achieved using OGSA-DAI (Open Grid Services Architecture – Database Access and Integration, *www.ogsadai.org.uk*). A data indexing service allows datasets to be discovered uniformly using the Globus Toolkit (*www.globus.org*).

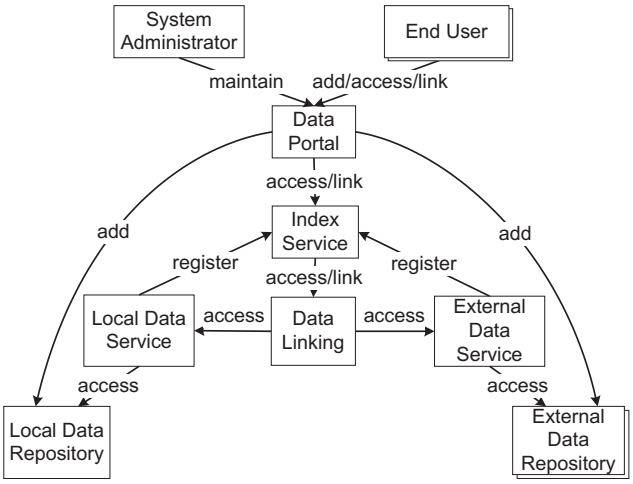

Figure 1. GEODE Portal Architecture

The data portal (GridSphere, *www.gridsphere.org*) supports users who simply require access to data, as well as users who provide datasets. The portal allows single sign-on, tied to standard security mechanisms such as GSI (Grid Security Infrastructure). This supports authentication and authorisation, ensuring appropriate access to data that may be controlled in various ways. The delegation service of Globus is used to support credential delegation, allowing users to designate data services that operate on their behalf.

2.2. Data Management for E-Social Science

The data portal is being extended in a number of ways to create an e-infrastructure on the DAMES project. Supporting more sophisticated use of metadata is described in section 3, while integrated support of data-oriented workflows is described in section 4.

The GEODE portal supported data virtualisation using OGSA-DAI. There are, however, two important limitations of this approach. The first is that OGSA-DAI is not designed for handling unstructured datasets or proprietary data files. The second is that OGSI-DAI is a rather heavyweight and complex solution for transferring entire data files between locations.

To address these issues, the DAMES infrastructure aims to support more flexible data virtualisation across a wider range of distributed datasets. iRODS (Integrated Rule-Oriented Data System, *www.irods.org*) is an appropriate solution. iRODS was recently developed as a data grid tool for

handling distributed/replicated data files, and for enacting a series of rules when file manipulation is requested. These rules can be used to retrieve the content of a distributed dataset, returning this in response to the original request. A DAMES user thus does not need to know anything about the tools that obtain the requested data. The OGSA-DAI service for GEODE then becomes just one of several (possibly grid-aware) tools for accessing remote data. The use of iRODS is also key in transferring data to its destination without going through an intermediate server.

Grid security was originally based on digital security certificates (X.509). These have generally proved unpopular with users. For example, an elaborate procedure is required to establish CAs (Certificate Authorities) and to obtain certificates from such bodies. Users are accustomed to simple login procedures using locally administered usernames and passwords. A user who has authenticated locally, should ideally be accepted through the network. This is the goal of the Shibboleth project (*shibboleth.internet2.edu*). The DAMES team at the University of Glasgow is developing Shibboleth capabilities for the data infrastructure.

3. Metadata for Distributed Data

3.1. Background to Metadata

Metadata describes datasets, their administration and their usage. High-quality metadata is necessary to support resource discovery and to provide users with a context for re-using data such as in data fusion.

Statistical information is collected, analysed and stored by many organisations and researchers with an interest in social data. It is estimated that the volume of UK social science data will reach around 10 terabytes by 2010 [2], although other sources consider this to be an underestimate.

Various repositories archive social science datasets. To support searching and processing across archives, metadata must first be defined for resources. In the early years of social science research, metadata was recorded by data producers in an *ad hoc* fashion using non-standard techniques.

For data resources, various standards have emerged to describe different data characteristics such as historical and contextual details, resource relationships, and procedures for archiving and access. General resource descriptions can be written using RDF (Resource Description Framework [9]). Metadata is vital for distributed data management because it is the foundation for resource discovery, access and usage. Unfortunately, the variety of metadata standards and differences in metadata quality have hindered the use of data – a challenge being addressed on DAMES.

In the social sciences and other disciplines, a number of different data standards have arisen. Their proliferation has impeded distribution and interoperability. The DDI metadata standard (Data Documentation Initiative [6]) has therefore been created for describing social science data. Metadata standards like DDI are now being used rigorously to describe existing and new data. Standardised metadata descriptions facilitate tasks such as: re-use and transformation of data, repeatability of analysis, interoperability of datasets across platforms and packages, improved data quality and relevance, and properly controlled access to data.

3.2. Metadata for Social Science

The DAMES project is using the concept of a 'data grid' as its foundation. This facilitates the federation of, and access to, data resources stored in different repositories. It can also provide analysis and processing services through the network, including re-purposing data. Metadata queries and exchange are central to this infrastructure. Social data evolves (e.g. it is updated or extended) as it is often used and transformed for new purposes (e.g. occupational data might be used to inform education strategy). Metadata must capture changes over the lifecycle of data, describing its evolution and making discovery easier.

The DAMES infrastructure is designed to register and store both data and metadata. Social scientists can describe their datasets, but are not required to curate the data fully. The infrastructure can fill in missing metadata based on the provenance, content and structure of the data. This may require querying other networked resources to retrieve associated metadata. A 'metadata wizard' can guide non-technical users in defining metadata. As data is re-used and re-purposed, metadata about the uses and any changes made to the datasets can be captured automatically, thereby preserving the integrity of the data and the original metadata.

Registering, organising and cataloguing metadata are key functions of the e-infrastructure. A vital issue is maintaining the integrity and description of datasets that are typically distributed across a network. An XML database is an obvious choice for storing DDI metadata. The preferred database is eXist-db (*exist.sourceforge.net*). However, there remain challenges of scalability, robustness and security. The AMGA metadata catalogue (Arda Metadata Grid Application, *amga.web.cern.ch/amga*) has also been investigated as an alternative. This has found popularity for storing metadata in a grid environment. It is planned to evaluate a variant of AMGA using eXist.

The DAMES project has adopted DDI version 3.0 as its metadata standard. DDI version 2.1 was successfully used on the GEODE project to describe occupational data. DDI 3 is desirable because it natively supports the data linkage and processing features needed for DAMES. For example, there is good support for longitudinal aggregated data, assessment of data comparability, resource grouping and lifecycle

management. DDI 3 supports a completely different model for thinking about survey data, and also supports different kinds of data such as summary indicator mappings. However, DDI 3 has only recently been finalised so that very few existing resources have been described using it. Mappings from DDI 2 metadata to DDI 3 will allow DAMES to use data from other sources such as GEODE.

As an example of how DDI can be used for social science metadata, the extract below [3] describes a classification scheme for occupations. The *document_description* explains who created the data, when, and on what basis it may be distributed. The *study_description* describes the study that gave rise to the data. The *file_description* defines the file(s) that contain the data. The *data_description* gives technical information about the data such as the variables it contains and how they are coded. Obvious closing tags have been omitted in the following XML:

```
<codebook>
  <document_description>
    <distribution_statement>
      <contact email="psl@stir.ac.uk">Paul Lambert
      ...
    <production_date date="2008-07-19">19th July 2008
    ...
  <study_description>
    <title>CAMSIS Scales for the UK using SOC2000
    <identifier agency="GEODE">131
    <distributor URI="http://www.camsis.stir.ac.uk">
      Cambridge Social Interaction and Stratification Scales
    <study_information>
      <summary_description>
        <production_time event="start">2000
        <nation abbr="GB">United Kingdom
    ...
  <file_description id="GB1991_SOC2000">
    <file_name id="GB1991_SOC2000">gb91soc2000.sav
  <data_description>
    <variable_group name="indexs" var="soc2000s ...">
      <concept>Index Terms
    <variable id="soc2000s" file="gb91soc2000.sav">
      <standard_category uri="http://www.geode.stir.ac.uk">
        Standard Occupational Classification 2000
    ...
  <other_material>...
```

DDI-based metadata is relevant to e-social science in a number of ways. The DAMES project has the key objectives of linking and transforming datasets. Metadata can be used to achieve this through DDI schemas that group and compare datasets. The grouping schema relates datasets and other resources, using a compositional structure that supports sub-grouping. Groups describe relationships based on aspects such as time, geography or language. DDI supports explicit description of dataset equivalence, similarities and differences using a comparison schema. DDI-based metadata is also relevant to social science processing instructions and workflows. Being XML-based, the metadata supports queries using XQuery and XPath.

DDI 3 is significantly more complex than the previous version. DDI 2 had a single schema organised into five sections. DDI 3 has 24 schemas and 14 maintainable schemes. Fortunately, DAMES is able to subset the specification and to extend it for particular requirements. With previous versions of DDI, organisations had to make *ad hoc* modifications to the DTD. However, DDI 3 supports controlled methods for creating profiles and extending the specification. Profiles are collections of XPath expressions defining used and unused metadata fields. Profiles ensure consistent metadata usage, allow validation, and promote better sharing across networks. Since DDI schemas are XML documents, they can be extended by defining a new XML namespace and importing the relevant DDI schema. This avoids confusion with the original DDI schemas.

4. Workflows for E-Social Science

4.1. Background to Workflows

A number of workflow languages have been developed, such as Microsoft's BizTalk and IBM's WSFL (Web Services Flow Language). For workflows to support social scientists, well-established standards and techniques are preferable for reasons of interoperability. However, other interesting developments are not ruled out. Taverna (*taverna.sourceforge.net*) is a mature approach to service orchestration through its SCUFL language. JOpera (*www.jopera.ethz.ch*) was developed for orchestrating web services, though it is now being applied to the grid. Triana (*www.trianacode.org*) is a visual programming approach that has been recently extended to grid workflows.

Because of the benefits of standardisation (acceptance, training, tools), the DAMES approach to workflow modelling uses WS-BPEL (Web Services – Business Process Execution Language [1]). This was conceived for use with *web* services, but has been adapted for DAMES to support *grid* services [8]. This allows full use of WSRF (Web Services Resource Framework) – widely used in grid computing. For example, EPRs (Endpoint References) for resources can be passed among partner services, and the choice of partners can be made dynamically. Support for resources such as datasets is particularly important in social science. Compared to other grid workflow languages, BPEL offers additional flexibility in areas such as resource handling, error handling, compensation and concurrency. A BPEL-based approach has also been followed in the BPEL4Grid Workflow Engine (*mage.uni-marburg.de/trac/mage/wiki/BPEL4GridEngineOverview*) and by OMII (Open Middleware Infrastructure Institute, *sse.cs.ucl.ac.uk/projects/omiibpel*).

4.2. Workflows for E-Social Science

Social scientists often perform many similar activities at all levels of analysis, despite the diversity in each sub-discipline. For example, common activities include coding data, creating models, and using statistical techniques. Social scientists share their procedures with fellow researchers, usually by informal means such as email or web pages. Datasets and schemes are also shared in similarly informal ways, though with the additional complication that the datasets depend on particular platforms or packages. These practices result in limited sharing of expertise and data [4]. As demonstrated by the GEODE project, grid computing allows effective sharing of datasets. In extending this approach, the DAMES project is also aiming to support sharing and dissemination of procedures in the form of workflows. The idea is that social scientists can readily contribute and retrieve procedures and datasets that are of value to other researchers. This goal is similar to the work of the myExperiment project (*www.myexperiment.org*).

A social science workflow contains interacting sub-flows at macro and micro levels. Macro flows are high-level and play a direct role in analysis. They invoke partner services developed by others, e.g. for specialised analyses like data fusion or data linking. These services can be located anywhere in the network, and can be owned and operated by anyone. Micro flows are the 'building bricks' of larger procedures. They collate, (re)code and transform data prior to analysis. For example, a micro flow might collate distributed data into one stream, might handle missing data values, or might reformat data for a different analysis. Macro and micro flows are combined into a single workflow.

Social scientists need to be able to create, distribute, find and execute workflows in an e-science environment. However, their focus is on performing research with greater productivity and not on the underlying computing issues. An easy-to-use method for defining workflows is therefore required. Ideally this should be graphical and platform-independent. The method should hide technical details such as the underlying languages, packages and operating system. Discovery mechanisms are required to allow non-technical users to contribute and to find workflows. It follows that workflows need metadata too. The entire lifecycle of workflows needs to be supported: conception, deployment, retrieval, execution, update, and final undeployment. For consistency, workflows as well as datasets are supported by the DAMES infrastructure.

Support for workflows has been integrated into the CRESS toolset (Communication Representation Employing Systematic Specification, *www.cs.stir.ac.uk/~kjt/research/cress.html*). CRESS is a simple graphical notation that describes services in many different domains. In an e-science context, CRESS is used to describe the orchestration of partner services in a workflow. An extract from a typical workflow is shown in figure 2. This describes how a conditional frequency analysis can be performed on occupational survey data. The workflow makes use of external partners: a *converter* service for recoding and storing data, and a *statistics* service for analysing data.

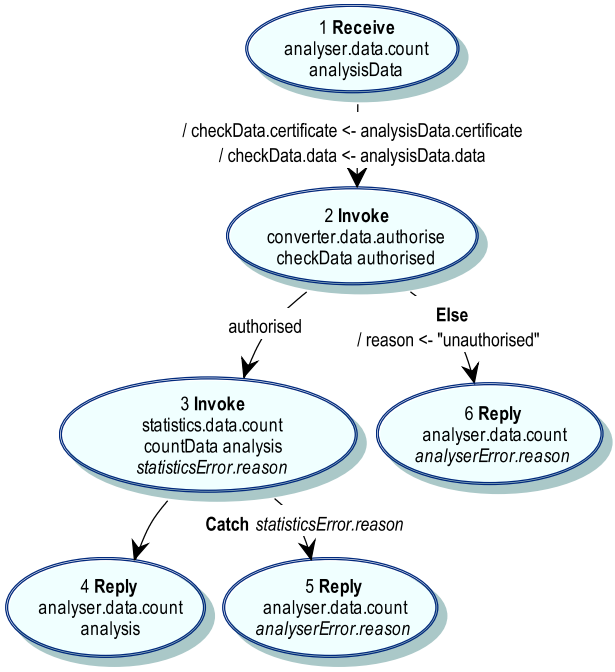

Figure 2. Example CRESS Workflow

Activities in CRESS appear in numbered ovals. As examples, activity 1 receives a service request, activity 2 invokes an external partner service, and activity 4 replies to a service request. When communicating with a service, the names of the service, port and operation are given (e.g. *analyser.data.count*). Operations are associated with parameters and results (e.g. *checkData* is sent to the *converter* service in activity 2, and an *authorised* result is obtained).

Arcs between activities show the service flow. A workflow can have alternative branches, e.g. flow proceeds from activity 2 to activity 3 if access to the data is *authorised*. A workflow can also have parallel branches and loops (not illustrated here). Assignments to variables can be made in activities or along arcs (e.g. the assignment to fields of *checkData* on the arc from activity 1 to activity 2).

CRESS workflows support error handling and compensation. For example, suppose the service invocation in activity 3 fails. The fault is caught and activity 5 is executed. If failure means that work needs to be undone, this is achieved by means of a compensation handler (not illustrated here). Error and compensation handling are supported at multiple hierarchical levels. Other aspects of CRESS diagrams support complex data definitions and the environment for executing grid services (e.g. the URIs of services).

CRESS has a mature toolset that supports automatic translation of user-drawn workflows into BPEL. Other supporting files are also generated, e.g. interface descriptions in WSDL (Web Services Description Language) and deployment files. The complexity of creating services and workflows is hidden by a simple graphical interface.

CRESS has the added advantage of being able to check correctness of workflows through automated validation and verification techniques [5]. A workflow description is automatically translated into a formal specification (LOTOS, *www.cs.stir.ac.uk/well*) that is then verified for correctness. During the development phase, this gives confidence that the workflow definition meets the user's expectations. At this point, the workflow description can then be automatically translated into an implementation and deployed into a workflow engine (ActiveBPEL, *www.activebpel.org*). The workflow can make use of external partners: plain web services (deployed using Tomcat) and grid services (deployed using the Globus Toolkit).

The workflow environment for DAMES is shown in figure 3. This meets the needs of social science researchers for defining, discovering and using workflows. In the figure, rectangles are activities offered to users, rounded rectangles are networked resources (within or outside DAMES), and arrows are activities that impinge on workflows.

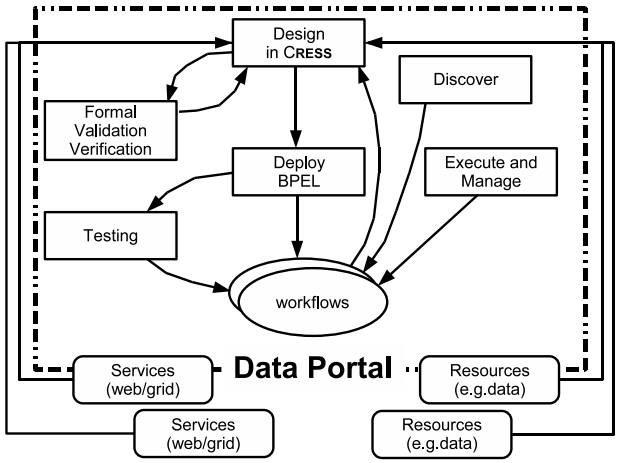

Figure 3. DAMES Workflow Environment

5. Current Status and Future Work

The DAMES project is work in progress, so evaluation results from users are not yet available. However, positive user acceptance of the previous GEODE results are an encouraging sign. An e-infrastructure is being created with extended support for metadata, workflows, distributed data, and security in e-social science.

The new capabilities for metadata are built on previous work by GEODE. The project is developing DDI profiles for a variety of social science data (e.g. for health, microsimulation and surveys). Tools are being created to translate existing resource metadata into the new DDI profiles. Queries over metadata are supported, as well as discovering procedures to link and transform datasets.

The DAMES workflow tools for e-social science are relatively mature. Using a simple graphical notation, services can be integrated from multiple sources into a data-oriented workflow. Workflow descriptions are automatically validated, verified and implemented.

Acknowledgements

The DAMES project is funded by the UK Economic and Social Research Council (grant RES-149-25-1066). The authors thank their colleagues in Social Science at the University of Stirling and in Computing Science at the University of Glasgow for their fruitful collaboration.

References

[1] A. Arkin *et al.*, editors. *Web Services Business Process Execution Language*. Version 2.0. Organization for The Advancement of Structured Information Standards, Billerica, Massachusetts, Apr. 2007.

[2] T. Hey and A. Trefethen. The data deluge: An e-science perspective. *Grid Computing*, pp. 809–824. John Wiley, 2003.

[3] P. S. Lambert *et al.*. Data curation standards and social science occupational information resources. *Int. J. of Digital Curation*, 2(1):73–91, July 2007.

[4] K. L. L. Tan, P. S. Lambert, V. Gayle, and K. J. Turner. Enabling quantitative data analysis on cyberinfrastructures and grids. *Proc. Int. Conf. on e-Social Science*, pp. III.20–III.31, Ann Arbor, Michigan, USA, Oct. 2007.

[5] K. L. L. Tan and K. J. Turner. Automated analysis and implementation of composed grid services. *Proc. 3rd South-East European Workshop on Formal Methods*, pp. 51–64. Thessaloniki, Greece, Nov. 2007.

[6] W. Thomas, A. Gregory, J. Gager, I.-L. Kuo, A. Wackerow, and C. Nelson. Data documentation initiative. Technical Report Version 3.0, DDI Alliance, Michigan, Apr. 2008.

[7] K. J. Turner *et al.*. Grid computing for virtual organizations: An E-social science case study. *Encyclopaedia of Networked and Virtual Organizations*, pp. 643–651. IGI Global, Hershey, Pennsylvania, Feb. 2008.

[8] K. J. Turner and K. L. L. Tan. Graphical composition of grid services. *Rapid Introduction of Software Engineering Techniques*, number 4401 in Lecture Notes in Computer Science, pp. 1–17. Springer, Berlin, Germany, May 2007.

[9] World Wide Web Consortium. *Resource Description Framework (RDF)*. Version 2.0. World Wide Web Consortium, Geneva, Switzerland, Feb. 2004.