K. L. L. Tan, V. Gayle, P. S. Lambert, R. O. Sinnott and K. J. Turner. GEODE - Sharing Occupational Data Through The Grid, in S. J. Cox, editor, Proc. 5th. UK e-Science All Hands Meeting, National e-Science Centre, Edinburgh, pp. 534-541, Sep. 2006, ISBN 0-9553988-0-0.

GEODE – Sharing Occupational Data Through The Grid

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

1. Department of Applied Social Science, University of Stirling
2. Department of Computing Science and Mathematics, University of Stirling
3. National e-Science Centre, University of Glasgow

Abstract

The ESRC funded Grid Enabled Occupational Data Environment (GEODE) project is conceived to facilitate and virtualise occupational data access through a grid environment. Through GEODE it is planned that occupational data users from the social sciences can access curated datasets, share micro datasets, and perform statistical analysis within a secure virtual community. The Michigan Data Documentation Initiative (DDI) is used to annotate the datasets with social science specific metadata to provide for better semantics and indexes. GEODE uses the Globus Toolkit and the Open Grid Service Architecture – Data Access and Integration (OGSA-DAI) software as the Grid middleware to provide data access and integration. Users access and use occupational data sets through a GEODE web portal. This portal interfaces with the Grid infrastructure and provides user-oriented data searches and services. The processing of CAMSIS (Cambridge Social Interaction and Stratification) measures is used as an illustrative example of how GEODE provides services for linking occupational information.

This paper provides an overview of the GEODE work and lessons learned in applying Grid technologies in this domain.

1. Introduction

1.1 Current occupational data utilisation

Social science surveys are analysed to understand the trend in societies and provide statistics for policy making and planning. Many analyses performed in social science often include occupation as a significant variable.

Occupational information is regularly collected by social science researchers and is usually analysed, or supplied to others, in the form of small electronic datasets, which typically detail occupational unit groups (OUG's). However social researchers are often unaware of how to use such data in an efficient and scientifically consistent way. In particular, occupational data is often analysed and released without documentation, which subsequently raises the barrier for other research efforts [1][15]. Publication of occupational datasets is commonly established via web links, furnished with usage instructions (although the data may be represented in other formats and media for instance email, disc and tape archive). There are also no formal semantic annotations used to define the datasets. Doing this can provide substantial benefits for data searches and access. The current trend is to have descriptions in natural language, which can be ambiguous, within materials accompanying the data sources supplied to end-users. Many of these data resources are also not indexed and therefore do not experience good exposure within the community.

Table A1 describes the existing format of occupational information datasets which have thus far been considered within the project. There are many further occupational information resources in use within the social science research community.

The Grid defines a scalable architecture where data and computational resources are virtualised, abstracted, and collaborated on within virtual organisations [2]. It is therefore highly desirable that occupational data utilisation be made possible in a Grid environment to overcome present issues. This paper illustrates how both the suppliers of occupational information datasets, and the social researchers who may wish to access this data can benefit from a Grid infrastructure developed in the GEODE project [3].

1.2 CAMSIS

CAMSIS scales are measures used by social researchers which indicate the average level of

advantage associated with different occupational unit group positions. CAMSIS scale scores are one of a number of alternative measures of occupational position which are widely used in this field. They are calculated on the basis of a statistical analysis of patterns of social interaction exhibited between individuals from different occupational unit groups.

The use of CAMSIS scales by social researchers illustrates a typical practical scenario of the current practices of distribution and utilisation of occupational data described in Section 1.1. CAMSIS scales are downloaded from a web link with usage instructions put up as descriptions in web pages.

The CAMSIS project [9] is coordinated by members of the GEODE research team so is used as the focal point of initial developments with the service.

1.3 Structure of paper

The paper discusses the GEODE project's intention to improve the practice of occupational data distribution, utilisation, and linking occupational information to CAMSIS scale scores. It presents the requirements and approaches of GEODE in Section 2. Section 3 illustrates the design and the architecture of GEODE and how it is influenced by the application requirements and current technical capabilities of Grid middleware. The results of the development work are discussed in Section 4.

2. Purpose of GEODE

2.1 Objectives and requirements

GEODE [3] aims to improve the current utilisation of occupational data by using the Grid. The goal is to create a virtual community where data resources are virtualised, indexed, and are uniformly accessed by users in a secure manner resulting in a gateway where occupational information can be discovered, exchanged and collaborated on. Occupational data analysis services are rendered to the community members. Occupational data researchers who are the members of this virtual organisation can have their occupational data resources abstracted, described and made accessible in a grid environment, thus standardising the practice of publishing quality datasets.

The GEODE project aims to deliver a usable application that is highly accessible for the users, most of whom have limited prior exposure to Grid services, or to formalised standards of data indexing. The choice is naturally a web interface (because of the ubiquity of web access) representing the view of the Grid, as further discussed in Section 2.4.

The project is also investigating the feasibility of extending its application scope to incorporate other forms of social science datasets in addition to occupational data.

2.2 Occupational data community

An occupational data virtual organisation should encompass disparate data resources made accessible to social science researchers belonging to the community. This is achieved using the MDS (Monitoring and Discovery System) [11], provided by the Globus Toolkit, which provides data aggregation and notification services. The organisation should have fine grained control over user access and the security of the data resources. An indexing service is to be deployed to hold registry information on resources and services. Resources register with the organisation through the indexing service, where resource sharing is then made possible. This is elaborated in the following subsection.

Services make themselves known to the community very much in the same manner as through registration with the index service. The difference is in the metadata used to register with the index service.

2.3 Virtualisation of data resources

The GEODE infrastructure leverages data abstraction Grid middleware (OGSA-DAI [4]) to create a framework for dynamically deploying data resources. The OGSA-DAI middleware, in addition to being able to automatically perform registration with the indexing service, contains the provisions to register custom metadata together with the database schemas.

The Michigan Data Documentation Initiative (DDI) [5] defines a set of XML schemas for annotating social science datasets, thereby promoting the semantic description of the data. The occupational data in the GEODE community is also annotated with social science (custom) metadata (DDI) to give it semantic definitions that are used for yielding more accurate searches than using keywords. The semantic metadata are registered in the community index service when the related data resource is added to the GEODE gateway.

2.4 Usability and accessibility

Many social science researchers are unfamiliar with advanced computer applications. Therefore it is desirable to develop GEODE as an application that can be used with minimal learning and configuration. Though a custom application has been considered, a web portal is much more appealing to the users.

GEODE has developed a web portal as the user interface by which occupational data researchers interact with the grid infrastructure. Through the portal, users can administer their data resources, search the data index, and make requests for statistical services. The portal is accessible via the Internet using standard web browsers. Application users are not bound to specific machines and software in order to perform tasks. This will greatly increase usability in the social science community. The portal was developed with the GridSphere Portal Framework [6], an open-source and widely used tool for portal development.

2.5 Services

The Grid-specific services are built with Globus Toolkit 4 [7]. This WSRF [8] implementation was recently accepted as an OASIS standard. At present, GEODE has developed specific services developed to make queries to the index service, and to link occupational data to CAMSIS scale scores [9]. As the scope of GEODE evolves, services can be readily implemented and deployed on the Grid and accessed via the portal drawing on the service-oriented characteristic of Grid services.

2.6 Security

Security is a major concern, especially when sharing data in the virtual community. Globus Toolkit uses GSI [10] to establish security in a Grid environment. GSI offers authentication, authorization, credential delegation, and single sign-on that GEODE leverages to administer resources, trust and portal service operations. OGSA-DAI makes it possible for resource security to be configured when deployed.

Users delegate their credentials (proxy certificates) to the GEODE services to allow operations to be carried out on their behalf and accounted for.

2.7 Framework Extensibility

Finally, it is highly desirable to make the infrastructure capable of incorporating social data about aspects other than occupation. This benefits social science researchers in other fields, whilst maintaining the same framework whose scope can be extended to provide data and services for more users. Therefore a generic design of the GEODE infrastructure is required to adapt to non-occupational social data.

3. Architecture

3.1 Overview

The high level architectural design of GEODE is depicted in Figure 1. The dotted box represents the occupational data virtual community that comprises the index service, various data services (G1, G2, O4) and application services (in this case the CAMSIS linking service). Users interact with the grid indirectly through the GEODE portal as their web interface. The individual components and functions are elaborated in detail in this section and in section 3.2.

3.1.1 Index service

The GEODE Index Service is considered to be the main core of the virtual community, as services and resources are first discovered prior to performing the actual operations. It is deployed on the default index service supplied by GT4. This index aggregates registration information from both data and application services, where the respective service metadata is propagated by each registering service. This design is not subject to a single index, though currently one instance is deemed sufficient.

3.1.2 Data services

Based on OGSA-DAI configurable data services and appropriate drivers, the GEODE data services provide the data abstraction on occupational data to overcome issues of heterogeneity of data sets including relational tables and text files (comma separated). Data resources are deployed dynamically and register with the index service where discovery could be made. O1, O2 and O3 are examples of the data resources abstracted by the respective data services. GEODE maintains two data services G1 and G2 (as shown in Figure 1). G1 virtualises data that are curated at Stirling locally, and G2 is a collection of resources harnessed from a wider international community of social scientists. The resources of G1 and G2 feature the occupational information described in Table A1.

OGSA-DAI provides a framework for automatic derivation of data access metadata, with provisions made for custom metadata also.

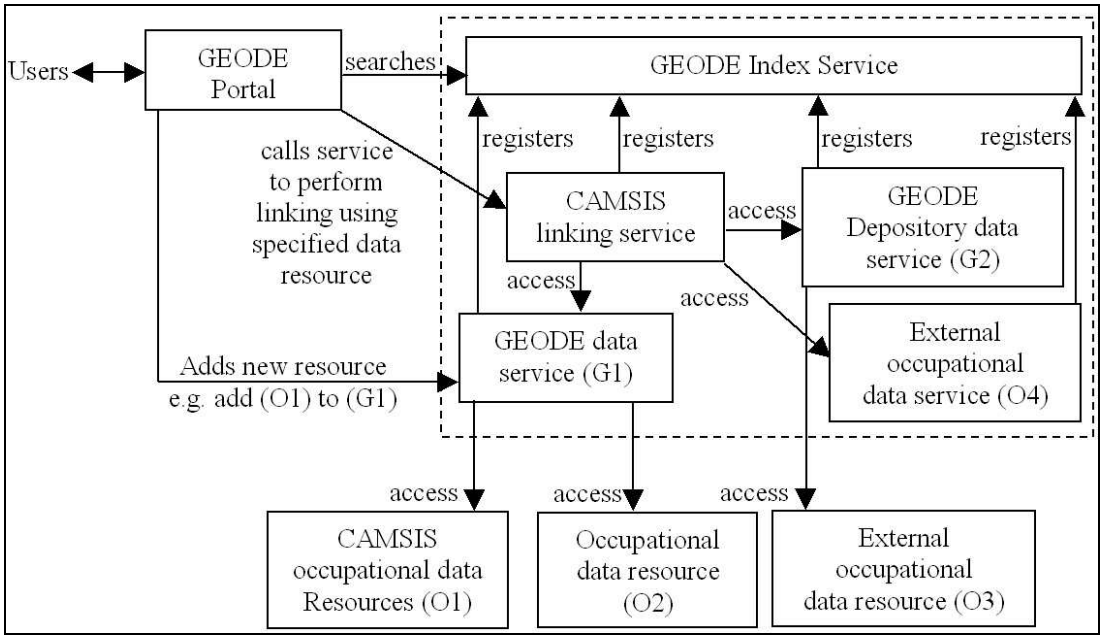

Fig. 1. GEODE Architecture

The DDI schema is used to annotate the data resource with social science metadata which upon registration is aggregated within the index service alongside with the data access metadata.

The GEODE architecture is scalable. An arbitrary number of data services can be deployed (intuitively on different machines/sites) to handle the vast amount of virtualised datasets, both internal and external to GEODE. Users may choose to provide data using their own OGSA-DAI data services (depicted as O4 in Figure 1), considered as external services that register with the index to be made available to all. There can be an arbitrary number of external data services like O4 registering to the GEODE grid, but for illustration purposes only one (O4) is shown.

Users do not interact with the data services directly, but rather do so via the application services which in turn are invoked from the portal.

3.1.3 Application services

The application services are grid services developed using GT4 to provide specific data analysis functions on the data resources. The application services access the data resources and render specified results to the user via the portal (user interface).

Authorization will be verified prior to accessing the services. This is where the GSI is involved to set up a security context with the users and to perform operations on their behalf (via credential delegation).

At the initial stage, only the CAMSIS score linking service will be implemented. As the Grid services are based in Service-Oriented Architecture (SOA), services can be developed and deployed into the Grid with minimal effort.

3.2 Portal

The GEODE portal provides the web interface to users by which operations and functions are invoked by proxy on the Grid services. GridSphere is used to develop the components that make up the entire portal, namely the presentation view, presentation logic, and the application logic. The view is implemented with JSP and the logic with portlets that controls the presentation flow.

The emphasis is on the application logic, developed as a portlet service, which interfaces with the Grid environment. The portlet service invokes the operations of the Grid services and returns results to the presentation logic. GEODE follows the Model-View-Controller design pattern which the occupational data Grid (model), presentation (view) and portlet service (controller) represent.

3.3 Extensibility

The GEODE architecture is designed to be as generic as possible. One of the most promising benefits is to be able to apply or extend the structure towards other social science statistical

data resources with similar requirements. In a generic context, a data Grid with registration to index services along with implemented services can fulfill the requirements of data sharing and collaboration to a considerable and substantial level. In addition to data abstraction and location transparency, this architecture allows control of services whereby the data provider may have the flexibility to provide data services as well as using the services set up within GEODE.

In principle the GEODE Grid can be extended and used for non-occupational social science data as it is designed generically. For example different DDI metadata can be customised for alternative data resources in instances when social scientists have similar requirements for both the storage and distribution of data. Possible areas of application may include the management of geographical and educational data resources, although the scope of this project does not include such implementations.

3.4 Issues

This section discusses the influences and experiences on the technical implementation.

3.4.1 Resource administration

Though GT4 features resource security, OGSA-DAI has yet to utilise this capability (version 2.1). Therefore a temporary measure of data resource administration is clearly required to manage the resource security. Data resources can be deployed by authorised users onto configurable data services. It is natural that the owners or a list of authorised users can manipulate the state, performing tasks such as undeploying resources.

3.4.2 Operations on resources

OGSA-DAI implements activities in a way that they are all invoked via one single operation. Therefore it is not possible to have different security configurations that GT4 supports for individual operations. Although not critical presently, it may become a growing consideration that will impact the project practically. E.g. an activity to modify of resource metadata may have a requirement of authorisation using with an access control list. This requires activity-level security configuration which OGSA-DAI supports shortly after the submission of this paper.

3.4.3 Credential delegation

There are a few ways to delegating credentials to services with GT4. One way is that a client perform the delegation to the services directly. Alternatively the client can store credentials in a depository, where services are then informed of details to retrieve the credentials in order to have the delegated rights. The latter is more favourable as it does not confine proxy credentials to specific locations to perform delegation.

There are currently 2 implementations of credential depository, namely MyProxy [16] and GT4's DelegationService [17]. MyProxy is only available as a software installation in Unix/Linux flavours. GEODE is implemented under Windows and is preferable keep a single environment for maintainability. DelegationService is a Grid service that provides similar functions. Hence this service can be easily deployed into WSRF containers. GEODE aims to use DelegationService as the proxy credential depository. However the current limitation in using DelegationService is that credentials can only be delegated to services that run within the same service container. In cases where services are deployed in multiple containers, the limitation could be resolved by having a DelegationService deployed in each container. Ideally the DelegationService can be used to perform delegations to services in disparate containers.

4. Results

This section reports the results of the prototype development, the current status and progress of GEODE.

4.1 GEODE Prototype

A basic Grid architecture has been developed and deployed, comprising the indexing service and the G1 data service (shown in Figure 1). The indexing service is a mandatory component to establish the virtual community. G1, being an OGSA-DAI configurable data service behaves similarly to G2 and O4. Therefore the latter two services are not required to be deployed for the prototype development. There are no restrictions to how many G1 services deployed. Relational databases and comma-separated value files (local disc and HTTP access) can be deployed as data resources in GEODE.

Custom OGSA-DAI activities were developed for deploying data resources and modifying metadata. Resources, when deployed, automatically register with the

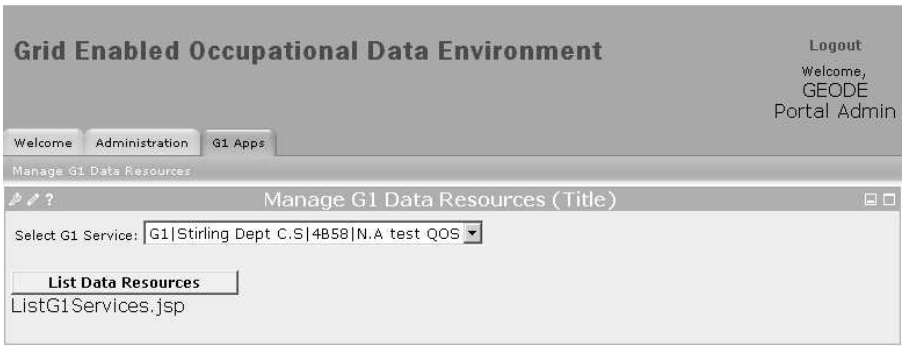

Fig. 2. List of G1 services registered in indexing service

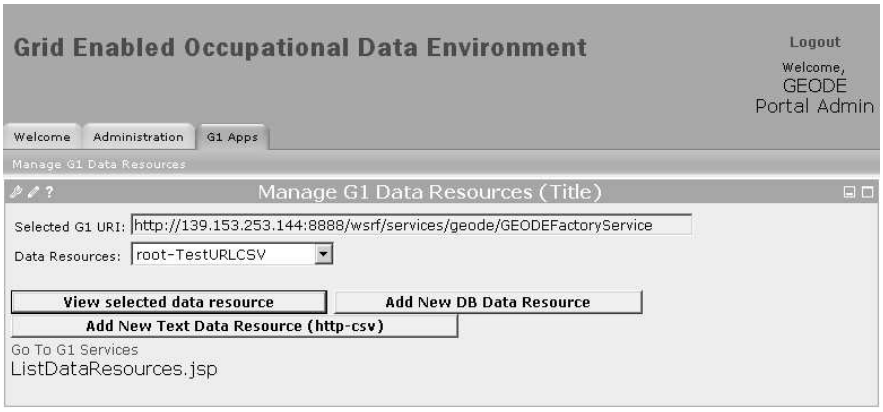

Fig. 3. List of data resources in selected G1 service

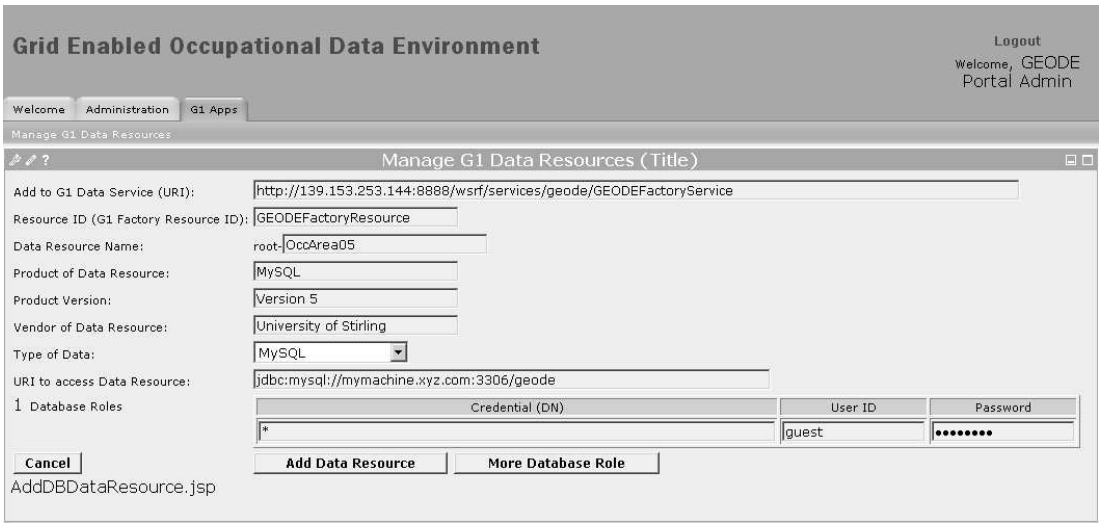

Fig. 4. Add new data resource to selected G1 service

indexing service with the initial specified metadata. The metadata when altered is reflected into the indexing service. Presently the metadata that can be altered is the list of comma-separated value files that are accessible via HTTP. When changed updates to the schema of the data resource that represents it occurs.

The GEODE portal was developed with a user interface that interacts with the indexing service and G1. The views (portlets) accept the user's input, which the portlet services use to communicate with the indexing service and G1. Figures 2 to 5 shows screenshots of the GEODE prototype portal. Figure 2 illustrates the portal querying the indexing service for all G1 services and displays them as a list. The portal

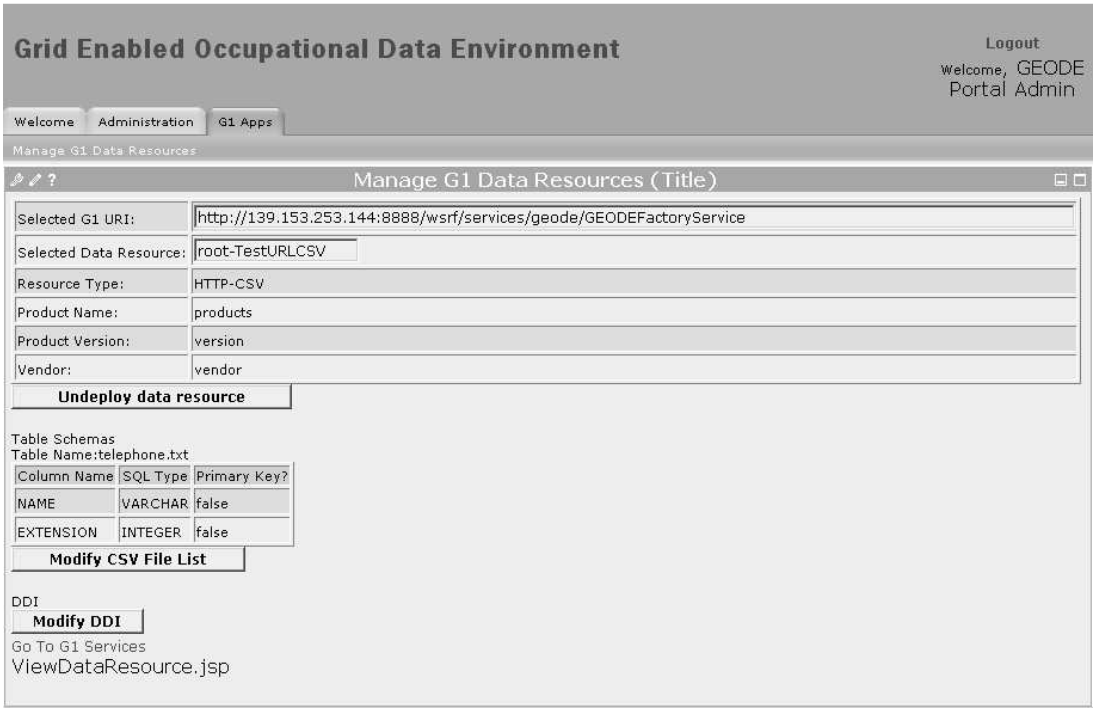

Fig. 5. View data resource in selected G1 service

is also able to list data resources deployed selected G1 services as seen in Figure 3. It is now possible to deploy (see Figure 4) and undeploy (undeploy button in Figure 5) data resources, and to modify the metadata via the portal. In addition, the portal is able to query the indexing service for data resources registered with G1 and to display the metadata of selected data resources. Checks are put in place to guard against the issues listed in Section 3.4.

4.2 Project progress

DDI has been incorporated manually and tested successfully in registering and retrieving from the indexing service. GEODE is currently developing the portal interface and data service activities to manage semantic definitions interactively. The requirements for linking data resources to CAMSIS scores will be examined in detail, and implementation will commence thereafter.

GSI and credential delegation will be assessed and implemented once the functional requirements of GEODE are finalised. To simplify the complexities of client GSI set up, creating proxy certificates and delegating credentials, the web start of the CoG (Commodity Grid) kit [12][13] will be investigated. Java Web Start [14] allows applications to be deployed and launched with a single click from a Web browser, thus omitting complicated and specific installations for GEODE users. This allows researchers to utilise GEODE on other machines instead of being constrained to their own machines.

Prospective users have been identified and would engage in the assessment of using the GEODE prototype when it is ready. This will gather valuable feedback to help in the development of a useful GEODE portal.

5. Conclusion

5.1 Benefits of GEODE

The prototype has illustrated that the design and implementation provides a viable and scalable framework for GEODE and the community members. GEODE and its design in principle can be extended and applied for non-occupational social science data.

GEODE encourages good occupational data utilisation via the portal for occupational information exchange and services, and having the datasets semantically annotated. Occupational data researchers have a common channel for dataset publication that improves data quality definitions, usability and better publicity. The complexity of accessing CAMSIS stratification scale scores will be greatly reduced as a result of using the CAMSIS data linkage service. Likewise other

occupational information linkage services can be applied to meet other user requirements.

5.2 Future work

The possibilities of extending GEODE are great. For example, researchers can create collaborations as resources which authorise members of the community can use, operate on, and share. Datasets in XML format can be virtualised readily with OGSA-DAI as the middleware, although this is not required in the scope of GEODE. Analytical and often time-consuming statistical services can be developed and deployed to the Grid. There is a substantial user-community in the social sciences who would benefit from utilising the GEODE services. Additionally, GEODE can be extended to incorporate non-occupational data and services. Ideally GEODE can be established as the portal for a wide range of social science data.

Appendix

Table A1: Selected Occupational Information Datasets used in GEODE

1. CAMSIS indexes, www.camsis.stir.ac.uk/versions.html	
Format:	Index matrix, SPSS and plain text data files
Units:	Variety of national OUG, plus gender, employment status
Output:	CAMSIS scale scores
2. CAMSIS occupational information value labels, www.camsis.stir.ac.uk/occunits/distribution.html	
Format:	One-to-one translation, SPSS and plain text files
Units:	Variety of national OUGs
Output:	Text labels to numerical OUG codes
3. ISEI tools, home.fsw.vu.nl/~ganzeboom/pisa/	
Format:	One-to-one translation, SPSS and plain text files
Units:	ISOC-88 and -68 international OUG schemes
Output:	ISEI and SIOPS occupational scale scores
4. E-SEC matrices, www.iser.essex.ac.uk/esec	
Format:	Index matrix, MS-Excel and SPSS syntax
Units:	ISCO-88 OUG, and employment status (es) data
Output:	E-SEC class position for OUG-es combination
5. Hakim gender segregation codes (Hakim, C. 1998 *Social Change and Innovation*, Oxford University Press, pp266-90).	
Format:	One-to-one translation, paper printout
Units:	ISCO-88 and UK SOC90 OUG schemes
Output:	Gender segregation information for OUG codes

References

[1] P.S. Lambert, "Handling Occupational Information", *Building Research Capacity*, pp 4:9-12, Nov. 2002.

[2] I. Foster, C. Kesselman, S. Tuecke, The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International J. Supercomputer Applications, 15(3), 2001.

[3] P.S. Lambert, V. Gayle, K. Prandy, R.O. Sinnott, K.L.L. Tan, K.J. Turner, GEODE Grid Enabled Occupational Data Environment, http://www.geode.stir.ac.uk, Oct. 2005.

[4] OGSA-DAI, Open Grid Service Architecture, Data Access and Integration, http://www.ogsadai.org.uk, Feb 2006.

[5] DDI, Data Documentation Initiative, http://www.icpsr.umich.edu/DDI/, Feb. 2006.

[6] Gridsphere Portal Framework, http://www.gridsphere.org, Dec. 2005

[7] I. Foster, Globus Toolkit Version 4: Software for Service-Oriented Systems, IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2005.

[8] WSRF specification, Web Services Resource Framework (WSRF) v1.2 Specification.

[9] P.S. Lambert and K. Prandy, CAMSIS: Cambridge Social Interaction and Stratification scales, http://www.camsis.stir.ac.uk/, Aug. 2005.

[10] The Globus Alliance, Grid Security Infrastructure, http://www.globus.org/toolkit/docs/4.0/security/key-index.html, Apr. 2006.

[11] Jennifer M. Schopf, Monitoring and Discovery in a Web Services Framework: Functionality and Performance of the Globus Toolkit's MDS4, http://www-unix.mcs.anl.gov/~schopf/Pubs/mds-sc05.pdf, Apr. 2006.

[12] Java CoG Kit – Webstart Applications, http://www.cogkit.org/release/4_1_2/webstart/, Feb. 2006.

[13] Java CoG Kit, http://www.cogkit.org/, Feb. 2006.

[14] Java Web Start Overview White Paper, http://java.sun.com/developer/technicalArticles/WebServices/JWS_2/JWS_White_Paper.pdf, May 2005

[15] P.S. Lambert, K.L.L. Tan, V. Gayle, K. Prandy, R.O. Sinnott, Developing a Grid Enabled Occupational Data Environment, to appear Second International Conference on e-Social Science, Manchester, UK, June 2006.

[16] The Globus Alliance, GT4.0: Credential Management: MyProxy, http://www.globus.org/toolkit/docs/4.0/security/myproxy/, Apr. 2006.

[17] The Globus Alliance, GT4.0: Security: Delegation Service, http://www.globus.org/toolkit/docs/4.0/security/delegation/, Apr. 2006.