Koon Leai Larry Tan, Paul S. Lambert, Vernon Gayle and Kenneth J. Turner
Enabling Quantitative Data Analysis on Cyberinfrastructures and Grids.
In Gary Olson and Rob Procter, editors, Proc. 3rd International Conference on e-Social Science,
pages III.20-III.31, Ann Arbor, Michigan, USA, October 2007.

Enabling Quantitative Data Analysis on Cyberinfrastructures and Grids

Koon Leai Larry Tan[1], Paul Lambert[2], Vernon Gayle[2], Ken Turner[1]

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

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

Email address of corresponding author: klt@cs.stir.ac.uk

Abstract. The social, behavioral and economic sciences (SBEs) do not currently benefit from a unified workflow environment for the quantitative analysis of social survey data. Some unified models integrating data storage, data management and data analysis do exist, for example the NESSTAR, IPUMS and LIS projects. However all of these services are focused on a limited number of data resources and functionalities. The Cyberinfrastructure could be exploited to develop and support a more generic workflow environment. In this paper, we build upon earlier work in providing a specialist data access service to social scientists (the GEODE project), to outline a proposed framework for a generic quantitative social science infrastructural service based on open standards.

Introduction

In this paper we discuss some of the current practices in providing, obtaining, accessing and using quantitative datasets in the social, behavioral and economic sciences. We evaluate approaches to accessing and analyzing quantitative data in the social sciences, and propose a generic Grid framework/middleware for supporting quantitative social science research. The specific benefits of this framework for social scientists are discussed and illustrated with research examples. We use the GEODE project as a case study where some of the ideas have been implemented, and also describe how it could be further developed.

Current Practice

Publishing and obtaining data and resources

In characterizing current activities, it is useful to distinguish three groups of data resources. Firstly, analytical data is the subject of the research. It is typically 'micro-data' on the subjects of analysis – such as individual level responses from questionnaire surveys. Analytical data is typically shared between small numbers of users in controlled conditions. For instance, secondary survey researchers may access their analytical data by downloading existing survey datasets from dedicated provision services, such as the UK Data Archive [UKDA] or the IPUMS project [IPUMS]. In some examples, analytical data is accessed remotely, by running queries on secondary data stored at an external site, such as in the example of the Luxembourg Income Study [LIS].

Secondly, aggregate social science data resources comprise more generic information that may be linked with analytical data. Aggregate data is often shared widely, for instance being freely available online. The GEODE project focused on one example of linking aggregate data (occupational information resources) with analytical data (survey micro-data) [GEODE].

A third type of data comprises processing scripts, such as software instructions and commands, which may be applied by researchers in a generic way. These can include information on the commands necessary to perform a certain analytical task, or the commands needed to achieve a transformation in the nature of another data file. Several support services in the UK publish processing scripts online, such as the instructions furnished by the UK Economic and Social Data Service on working with major UK social surveys [ESDS Government].

Data resources are often shared amongst social science researchers, but current practices in publishing and accessing data have some limitations. Data resources and processing scripts are published in various environments, from small privately-owned web-sites to large-scale public repositories. These implementations (and in turn the appropriate discovery process) usually vary according to different local approaches, an inconsistency that can limit the potential use of resources. Therefore, whilst many social research scenarios involve linking together analytical data, aggregate data and processing scripts, social scientists often lack proficiency in undertaking such linkages. A contribution to SBE research resources could therefore be made by facilitating the linkage between these three types of data resource.

Provision of analytical data

Analytical data is often accessed under clearly defined conditions concerning the production and distribution of the data. Census datasets provide one typical scenario. A large number of datasets derived from official census data are available as public resources, residing in well-known web-sites, such as the UK Census database [Census.ac.uk] and the IPUMS project [IPUMS]. These datasets are typically described using online codebooks with information on variable semantics, alongside published details on the relevant project (search functions may also be available, local to a specific data provision service, for discovering and obtaining relevant data). However metadata is mostly provided as text written in natural language which requires human interpretation. Whilst different datasets may be compatible in terms of variable values, semantics and format, a significant limitation is that there is no standard practice in providing descriptions on such datasets.

Many distributors of analytical data have an agreed model of authentication to facilitate user access among participating members. For instance, the UK Data Archive [UKDA] and Census database [Census.ac.uk] use the Athens system premised on institutional authorizations [JISC - Athens]. Organizations implementing security models differently from one another could not easily provide such seamless cross-boundary access without putting effort into security integration.

Provision of aggregate data and processing scripts

Social scientists may also compile aggregate data and processing scripts, and share them with fellow researchers. Figure 1 shows an overview of one commonly used approach. It indicates how outcomes of one researcher's project are themselves discovered and used by another researcher as a data resource. This model is similar to the use of occupational data published at the websites of the CAMSIS [CAMSIS] and PISA [Ganzeboom] projects.

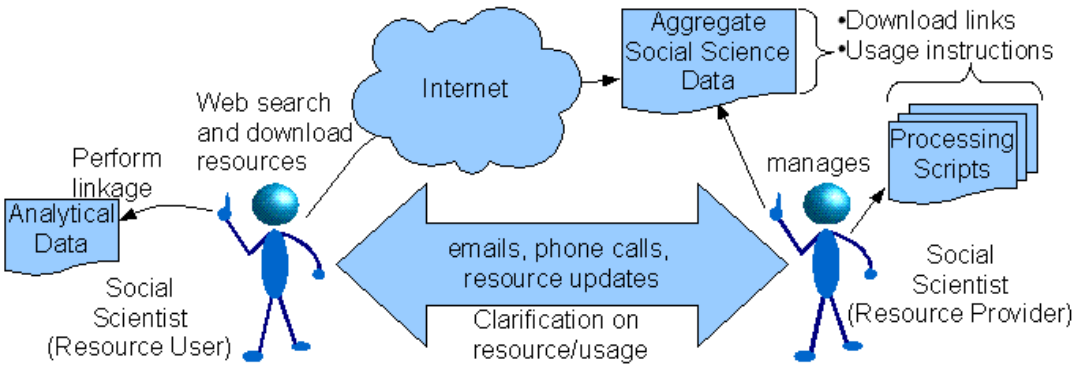

Figure 1. Informal approach of sharing resources

Unifying data provisions

A feature of contemporary quantitative data analysis in the SBEs is that the organisations and individuals involved in publishing and distributing analytical data, on the one hand, and those who publish and distribute aggregate data and processing scripts, on the other hand, are in large part separate. This may reflect the specialist nature of the tasks involved in either aspect of data production and provision. However one impact is that there has been little integration between the formats for data provision involving either resource.

Standardised metadata structures offer one possibility for integrating the distribution of analytical data with aggregate data and processing scripts. Lambert et al. (2007) describe how this can be done with the example of data resources associated with the analysis of occupations [IJDC 2007], in this case using the Data Documentation Initiative [DDI], version 2.1 metadata structure. The DDI specifies a comprehensive set of XML schemas for annotating social science datasets at various levels (from document to variable). This comprises a standardised method of metadata annotation to facilitate better semantic interoperability amongst datasets. The benefit of using standardised metadata is two-fold. First, it can allow for machine interpretable processing and semantic resource searches. Second, the metadata is maintained separately from the data itself, so it is possible to annotate data of different formats in a similar fashion.

Many archives have already moved in this direction. Once resources are in this fashion, users may be able to perform searches, run analyses, and obtain comprehensive metadata with regards to targeted datasets. In one example, the NESSTAR service provides software and an architecture for annotating analytical datasets using the DDI XML schema [NESSTAR]. Figure 2 illustrates this recent approach.

The dissemination of social science datasets has benefited from a standard, structured metadata notation and data dissemination such as in the NESSTAR implementation. However this approach does not provide data abstraction, from the specific formats of resources, that is accessible to services located elsewhere. Better interoperability, access and exploitation could be achieved by combining data virtualization with the standardised exchange of metadata.

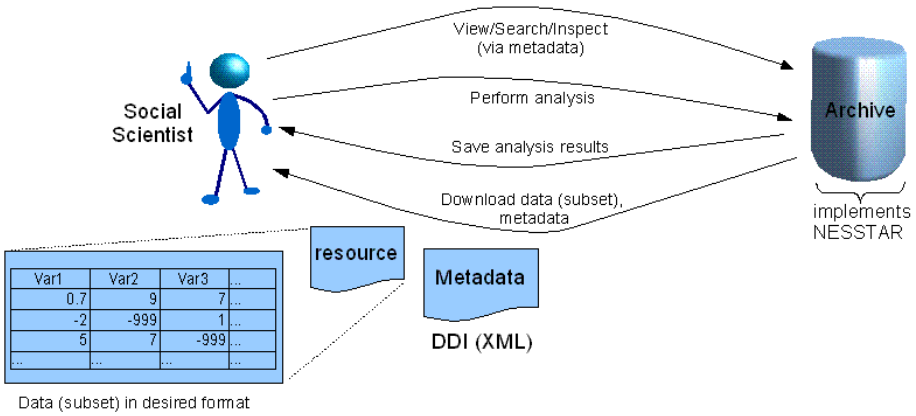

Figure 2. Recent and current practice of resource providers

Services

Resource providers may also support online analysis. With NESSTAR, users can perform simple statistical functions including data sub-setting, cross tabulation, basic regression, and graphical visualization of results on remote analytical data. Similarly the LIS project provides for the functionality of running numerous statistical analyses on remote datasets, achieved through an email service [LIS]. These examples allow for central control over the integrity of the analytical data.

Often these implementations are closed group, in that they support only the data they are associated with. In addition the methods of usage and access are proprietary in that the data owners develop their own interfaces for clients. Whilst these provide abstraction and transparency of access to datasets, there is no standard way of accessing the functions available. In addition, such services concentrate overwhelmingly upon analytical tasks involving whole datasets or simple subsets of them, and have only very limited provision for the more extended tasks in manipulating analytical data (such as recoding variables and selecting data sub-sets) which are central to the analysis of SBE resources. The latter tasks are usefully labelled 'data management' activities in working with analytical data, to be contrasted with 'statistical analysis' activities. Important tasks within data management activities involve linking analytical data with aggregate data and processing scripts.

It could be of benefit to social science researchers if there was a suite of statistical analysis and data management functions that could be deployed and executed on diverse datasets. User-defined workflows could be developed using this suite as the foundation. A further advantage would be that functions could be performed on datasets located in disparate locations. This would imply data virtualization to achieve location and format transparency.

Requirements

The productivity of research can be increased by improving interoperability between data resources and services for statistical analysis and data management. This could involve generic framework services being developed and shared which could act on virtualized datasets. The increase in productivity would be based upon improving collaboration and exploitation of existing data through integrated data resource services, embracing both statistical analysis and data management across analytical data, aggregate data and processing scripts. On the contrary, this is not equivalent to the so-called "number crunching" performance improvements associated with other cyber-infrastructure provisions. Such

performance is less relevant since data resources in the SBEs are not usually large and tend not to be beyond the storage of an average machine.

A further relevant characteristic of quantitative data analysis in the SBEs is that users frequently wish to access, and process tasks on, numerous related datasets. For instance, researchers frequently re-run a number of closely related statistical models on the same datasets, and/or repeat the same analytical operations on datasets which have slight variations between them, such as through variables with small differences in their coding, or datasets with different volumes of missing data. This requirement for multiple replications of similar tasks on similar datasets also motivates a coordinated structure for data access and analysis.

We propose a framework for supporting social science quantitative data activities which would not be specific to any discipline or subdiscipline. The framework must be able to support the activities of data management and analysis, covering data discovery, sharing standard and user-specified analytical functions, and service discovery.

Such a framework would require data virtualization alongside an agreed metadata structure. A standard security mechanism should be considered for supporting seamless cross-boundary data access through authentication, authorization policies, single sign-on, and accounting. Such conditions would support more seamless data exchange due to the integration of access transparency and compatible semantics.

Potential users should be able to discover variously owned and located data resources in a standardised manner, preferably from a single point. Search functionality with regard to metadata will result in better semantic matches. Therefore it is necessary to have aggregation of the metadata and data resource discovery services. This can be achieved with metadata registries and specific service implementation for facilitating semantic search.

Statistical functions should be made available as services via a standard means of access that can act on datasets of diverse formats, at disparate locations, and with different security measures. These statistical functions could be arbitrarily deployed and configured, and be accessed by clients and peer services. Semantic descriptions of service capabilities are also essential to enable meaningful searching of services. A unified semantic property model for describing social science service capabilities is required; this might be developed using ontologies and taxonomies. Registry services are also required to facilitate service discovery.

We propose a grid approach since the anatomy of the Grid [Foster] meets most of the requirements supporting the proposed services, and since grid development toolkits exist which could be suitable (e.g. Globus Toolkit, OMII-UK). The proposed infrastructure would not necessarily require innovative methods, but contribute through enabling social science activities on the Grid.

Quantitative Data Virtualization

Data abstraction

Quantitative social science datasets are found in various formats according to the corresponding versions of statistical packages used by researchers (SPSS and Stata are

currently two of the most widely used packages[1]). OGSA-DAI can be used to develop resource abstractions for these datasets [OGSA-DAI]. This middleware features a framework for linking data resources and metadata which facilitates data access and data manipulation activities upon registered data resources. Currently there are no implementations of OGSA-DAI data resources that provide abstraction alongside facilities to perform specific statistical analyses. Implementation of such facilities in combination with data management functionality would be required. Figure 3 illustrates the layer of data abstraction with metadata, and relevant activities, in terms of the OGSA-DAI framework.

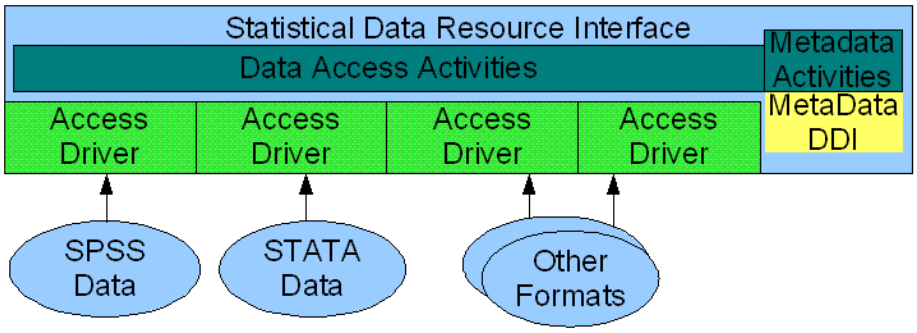

Figure 3. Social science data virtualization using the OGSA-DAI framework

Metadata and discovery

OGSA-DAI is able to support metadata covering the structure of data resources and statistical analysis activities via an implementation design pattern. Similarly it also supports storage of user-defined metadata for each data resource. The design pattern for exposing metadata in data resources uses XML and therefore fits well with the DDI structure. Metadata management can be developed as generic OGSA-DAI activities for social science data resources, as shown in Figure 4. Abstraction to metadata management is possible with high-level (e.g. visual) interfaces to OGSA-DAI activities, while absolute control is available by using the activities directly.

The data resources and their relevant metadata can be registered with registry and discovery services. A grid development toolkit like GT4 provides Monitoring and Discovery Services [MDS4] which have indexing service with aggregation and trigger capabilities. OGSA-DAI supports automatic registration to index services for data resources. Given this arrangement, it is possible to query and discover registered datasets. However the search functionality may be inadequate as these index services expose querying at the raw level (XPath). An abstract form of service discovery, encapsulating the structure of DDI, would be better able to facilitate searching at the semantic level, where there is no requirement for detailed user knowledge of DDI. Figure 4 depicts a possible arrangement for data discovery.

[1] There are alternative views on the direction that service provisions for quantitative data in SBE should go. One perspective is that provisions should exploit freeware for statistical analysis (such as the advanced analytical package 'R'). However we argue that is it unrealistic to restrict services to minority freeware, when leading proprietary software is widely used (and widely available to academic researchers). Therefore our orientation is toward services which are compatible with, and may complement, existing proprietary packages

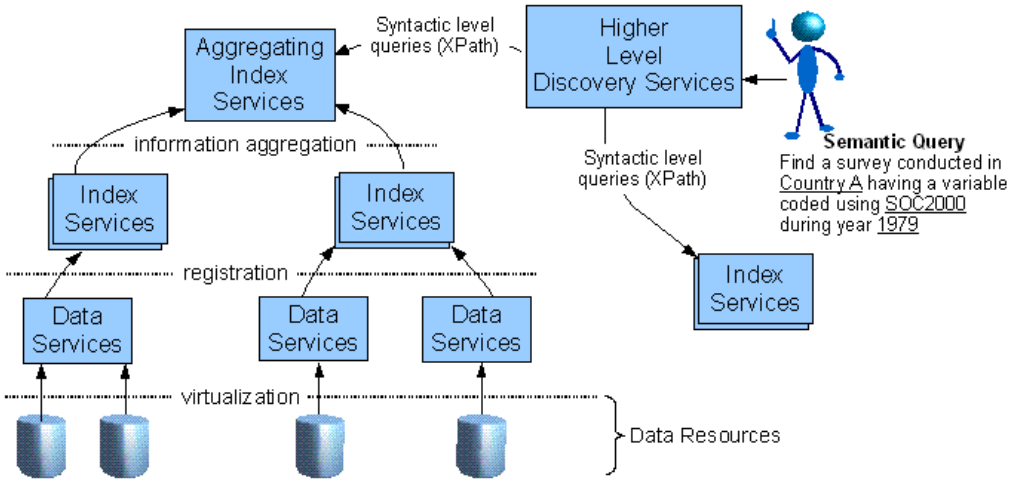

Figure 4. Data resource registration (e.g. metadata) and discovery via semantic search service

Nevertheless, using a standardised metadata scheme such as the DDI may not be sufficient on its own to fully support discovery. For example, a dataset may have a variable that is known to be of a certain classification (with a range of values). Though DDI supports the means to identify such classifications in variables, the responsibility for correct input lies with the entity that does the annotation.

Services and Discovery

Statistical services and proliferation

Services that render statistical functionality should be developed and deployed. Through these services a user should be able to perform statistical analysis on (a set of) selected data resources. It is then possible to connect an arbitrary number of data resources, and in turn an arbitrary number of statistical services. It is necessary that services be able to interpret the metadata (DDI) of virtualized data resources, interact with data services, and support peer services.

The Service Oriented Architecture of the Grid allows new services to be created from existing ones. New services may be created by social science researchers, deployed, and made accessible to others. Researchers can provide their analyses to fellow researchers, who can in turn use them in their own analyses. This may result in further creation of new resources, constantly expanding the capabilities and sustainability of the proposed Grid services.

GT4 implements the WSRF (Web Service Resource Framework) specifications. Consequently, Web Service orchestration/choreography specification standards such as the WS-BPEL 2.0 [BPEL] should be considered for supporting this capability. BPEL can specify behaviour of executable processes comprised of peer services. The deployment of a specified executable process behaviour results in a new service. There are many BPEL implementations across commercial vendors and open source and scientific workflows [OMII-BPEL]. These implementations can be used, but not exclusively, for service composition.

Service metadata and discovery

Users should be able to discover services according to their capability, along with other possible criteria specifications. Services deployed for sharing should be well-annotated for

appropriate discovery. When a new service is created involving the reuse of peer services, adoption from relevant existing service metadata should be made along with the new service metadata for totally new capabilities. This new service should be made discoverable via registration to indexing services.

Though there is middleware supporting service/resource discovery, there has not been a unified way of describing metadata for services specifically for SBEs. There has been work on using ontologies and models (e.g. RDF, UML, OWL-S) to describe, locate and match properties of services. Other approaches such as taxonomy, process choreography (e.g. BPEL abstract process), and possible combinations of techniques should also be considered. A unified model for describing the capabilities of statistical services and the behaviour of their operations as are relevant for social scientists is necessary. This would be a basis for compatible service discovery, whereupon discovery services should be implemented which include the functionality to reason about and match services against user specifications. Figure 5 shows high level view of the proposed service discovery and service composition, making reference to resources from the CAMSIS project.

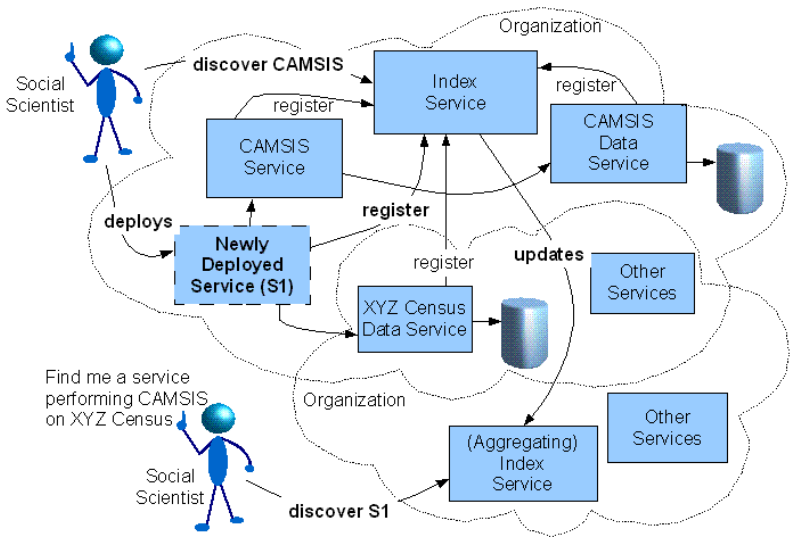

Figure 5. Service Discovery and Proliferation

Security

Security is a primary concern which is particularly acute when data is sensitive. The current standard approach is that of an authentication model, which also enables seamless access to other data hosts who participate in the security federation (e.g. Athens). Privately-owned micro datasets are usually not subject to any explicit security measures. However resources might be exchanged in closed-group communication, hence using an implicit form of security. In the grid environment well-known security frameworks like Shibboleth can be used to implement security [Shibboleth].

Shibboleth defines a way for an organization and a digital resource provider (Service Provider) to exchange information securely. The organization (Identity Provider) is responsible for user authentication and providing user attributes to the Service Provider, who decides the authorization outcomes based on the information received. Shibboleth uses open standards such as SAML (Security Assertion Markup Language) for asserting security

information. To achieve single sign-on among organizations (including service providers) using Shibboleth, they must belong to a federation which governs membership and trust. Therefore a federation for social science quantitative data and analysis community should be created, involving data archives, interested organizations and individuals. A governing body could potentially be responsible for managing the federation. Statistical service providers and registry providers could also enable their services with Shibboleth.

Data providers (Service Providers) will have to configure their data services to use Shibboleth. Continuation between Shibboleth and previous approaches is plausible since there has been work on migrating Athens to Shibboleth. Authorization policies should also be implemented upon the data resources. Accounting of access should also be monitored, e.g. by logging. There may be more specific application-related issues such as preventing identification of individuals, anonymizing records etc. If solutions are developed they should be implemented in the data services, but they must also take into consideration the interactions from/with statistical services that may negate the purpose. For example a data service may have implemented an anonymizing prevention measure. However, its access from statistical services may result in obtaining data which compromises the measures in place.

Case Study – GEODE

The main objective of the GEODE project was to support the use of occupational data in social science analyses, by facilitating access to existing aggregate data and processing script resources, and their linkage with analytical datasets. This was done by an online service known as the 'occupational information portal'.

The implementation was inclined towards interoperability between datasets. It support a certain extent of data virtualization, facilitates data discovery (syntactically) using metadata, and implements a specific grid service application for linking datasets with aggregated data such as the CAMSIS resources [CAMSIS]. Figure 6 shows a high-level overview of GEODE.

Aggregate datasets and processing scripts were virtualized as OGSA-DAI data resources, with each dataset annotated with a subset schema of DDI, and customised activities developed. This was usually achieved by converting these data to SQL and CSV (comma separated value) format. Social scientists can use the GEODE portal to access these resources. Each deployed resource registers its DDI metadata with an Index Service, making it visible to future searches. Those resources which involve aggregate data can have a 'mapping' logic configured to them, using information in relation to the DDI metadata. The Matching Service uses this record to link the resource through appropriate values on analytical data held externally by social science users. Thus the problems shown in Figure 1 are avoided, and the resources are therefore managed with higher data integrity.

Experience from this project is that careful metadata annotation practice is required regardless of the standard metadata model. A generic grid framework for quantitative social science statistical data is plausible, though subject to further investigation and tool development. A standard for modelling service metadata would facilitate service discovery. Conversely, there is a requirement for developing discovery for services. The SOA nature and proper service discovery foundation would serve as the basis for service proliferation. Data discovery can be higher level (semantically based) by abstracting the current syntactic aspect.

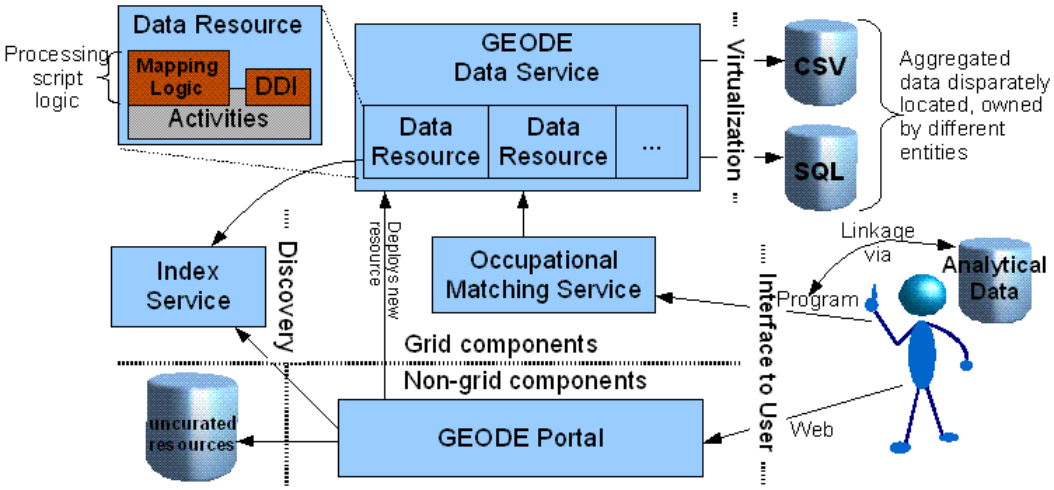

Figure 6. Overview of GEODE

Conclusion

The Workshop on Cyberinfrastructure for the Social and Behavioral Sciences concluded that Cyberinfrastructures can usefully facilitate collaborations and experiments at a very large scale, intensity and at high complexity [SBE-CISE-FINAL]. Additionally the National Science Foundation reports a list of summary recommendations for enabling and advancing Cyberinfrastructure for SBEs**.** In this paper we have demonstrated how the analysis and management of quantitative social science data through the Grid can be aligned with the NSF requirements, and that the complexities associated with proliferation of data resources are suited to a Cyberinfrastructure framework.

A generic grid framework for quantitative social science data is therefore plausible and should be subjected to further investigation and tool development. A standard for modelling service metadata would facilitate quality service discovery. Conversely there is a requirement for developing discovery services. An SOA framework having a proper service discovery foundation will support service proliferation. Data discovery can be of higher level (semantic) by abstracting the current syntactical aspect from users.

There are various ways to meet the objective of supporting a social science community for quantitative analysis. We propose a grid approach focusing on data virtualization, discovery, services, and higher resource exploitation. We highlight requirements to support constant development and sustainability within the social science community. Our recommendations lean towards building a generic middleware based on open standards as the foundation. Possible implementations have been illustrated (e.g. data virtualization), suggestions toward quality use have been given (e.g. metadata annotation practice), and requirements that need further investigations have been mentioned (e.g. service discovery). It is hoped that our suggestions and recommendations be useful for SBE infrastructural projects such as the UK's e-Infrastructure in the Social Sciences project [e-Infrastructure].

Acknowledgements

The GEODE project was funded by the UK ESRC as Small Grant RES-149-25-1015.

References

[BPEL] - Business Process Execution Language for Web Services version 1.1, http://www.ibm.com/developerworks/library/specification/ws-bpel/ , August 2007.

[CAMSIS] – CAMSIS web site. http://www.camsis.stir.ac.uk/. July 2007.

[Census.ac.uk] - Census.ac.uk: Moving you closer to the data, http://www.census.ac.uk/ . August 2007.

[DDI] - Blank, Grant, and Karsten B. Rasmussen. "The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." *Social Science Computer Review* 22(3): 307-318, 2004.

[e-Infrastructure] - Procter, R., Batty, M., Birkin, M., Crouchley, R., Dutton, W., Edwards, P., Fraser, M., Halfpenny, P. and Rodden, T. 2006. 'The National Centre for e-Social Science' in Cox, S.J. (ed.) Proceedings of the UK eScience All Hands Meeting (AHM '06) Edinburgh: National Centre for eScience.

[ESDS Government] – Economic and Social Data Service: Derived variables, http://www.esds.ac.uk/government/dv/ethnicity/lfs/index.asp, August 2007.

[Foster] – Anatomy of the Grid: Enabling Scalable Virtual Organizations, http://www.globus.org/alliance/publications/papers/anatomy.pdf, 2001.

[Ganzeboom] – HARRY GANZEBOOM'S Tools for deriving status measures from ISKO-88 and ISCO-68, http://home.fsw.vu.nl/~ganzeboom/pisa/. August 2007.

[GEODE] – K.L.L. Tan, V. Gayle, P.S. Lambert, R.O. Sinnott, K.J. Turner. GEODE - Sharing Occupational Data Through The Grid. Proceedings of the UK e-Science All Hands Meeting 2006, September 2006.

[IJDC 2007] - Lambert, P.S., Tan, K.L.L., Turner, K.J., Gayle, V., Prandy, K. and Sinnott, R.O. 'Data Curation Standards and Social Science Occupational Information Resources'. International Journal of Digital Curation 2: 73-91, 2007.

[IPUMS] – Integrated Public Use Microdata Series. http://www.ipums.org/, July 2007.

[JISC - Athens] – JISC Monitoring Unit: Monitored services. http://www.mu.jisc.ac.uk/servicedata/ , August 20007

[LIS] – Luxembourg Income Study. http://www.lisproject.org/, July 2007.

[MDS4] – Globus Alliance. Information Services (MDS): Key Concepts, http://www.globus.org/toolkit/docs/ 4.0/info/key-index.html. August 2007.

[NESSTAR] – Nesstar Limited. http://www.nesstar.com/, 2005.

[OGSA-DAI] – K. Karasavvas, M. Antonioletti, M.P. Atkinson, N.P. Chue Hong, T. Sugden, A.C. Hume, M. Jackson, A. Krause, C. Palansuriya. Introduction to OGSA-DAI Services. Lecture Notes in Computer Science, Volume 3458, Pages 1-12, May 2005.

[OMII-BPEL] – Open Middleware Infrastrucutre Institute: Modelling, monitoring, executing scientific workflows with BPEL (OMII-BPEL), http://sse.cs.ucl.ac.uk/projects/omiibpel/, August 2007.

[SBE-CISE-FINAL] – Workshop on cyberinfrastructure for the Social and Behaviorial Sciences: Final Report, http://director.sdsc.edu/pubs/SBE/reports/SBE-CISE-FINAL.pdf, May 12 2005.

[Shibboleth] – Shibboleth introduction, http://www.athensams.net/federations/shibboleth_intro, August 2007.

[UKDA] – Introduction to UK Data Archive. http://www.data-archive.ac.uk/, August 2007.