SGII: Towards Semantic Grid-Based Enterprise Information Integration*

Jingtao Zhou, Shusheng Zhang, Han Zhao, and Mingwei Wang

National Specialty Laboratory of CAD/CAM, Northwestern Polytechnical University, Xi'an, China, 710072
zhoujt@mail.nwpu.edu.cn, {zssnet, Wangmv}@nwpu.edu.cn, zhaohanboy@hotmail.com

Abstract. To fully leverage the information from different data sources and applications, an enterprise needs a generic, interoperable and flexible infrastructure to integrate and coordinate information across back-end data sources on semantic level. Through undertaking research at the intersection of the Semantic Web and Grid, the Semantic Grid expects to establish a semantic interconnection environment to effectively organize, share, cluster, fuse, and manage globally distributed versatile resources. In this context, we introduce SGII, an EII (Enterprise Information Integration) infrastructure based on Semantic Grid vision to achieve adaptive and intelligence information sharing. A survey of existent solutions is made to provide evidence of the benefits from Semantic Grid in the context of integration and interoperation of enterprise information. A primary architecture for SGII is introduced based on the analysis of realizing the vision of an infrastructure for semantic information integration on grid.

1 Introduction

In spite of extensive R&D and successful pilots, current enterprise information infrastructure is poorly suited for dealing with the continuing, rapid explosion in enterprise data. One underlying problem has remained unsolved: data resides in thousands of incompatible formats and cannot be systematically managed, integrated, unified or cleansed [1]. On the other hand, because semantic inconsistency has become an even great problem on the explicit information or knowledge sharing among users or applications, access level data integration has never given enterprise a competitive edge. Only by semantic, or knowledge level integration can enterprise gain insight into its workflow and business process.

Meanwhile, grid is expected to be the solution to the "islands of computation" problem [2], but it could also be the solution to the "islands of information" problem when it is extended to Semantic Grid, which wants to create an internetcentered interconnection environment on the grid to effectively organize, share, cluster, fuse, and manage globally distributed versatile resources based on the interconnection semantics [3]. In the context of enterprise information integration and sharing, Semantic

* Funding for this work is supported by Chinese 863/CIMS and Chinese National Research Foundation for Doctoral Program of Higher Education through The National Key Laboratory of CAD/CAM in NWPU Contract No. (2002AA414210) and (20030699032) respectively.

H. Zhuge and G.C. Fox (Eds.): GCC 2005, LNCS 3795, pp. 560 – 565, 2005.

Grid could provide a competently basic infrastructure for ensuring the information interoperation and collaboration both on system and semantic level.

2 Survey

2.1 Enterprise Information Integration

Enterprises have long recognized the value of data integration. Efforts can be roughly classified into two categories: application centric integration (ACI) and data centric integration (DCI).

ACI, such as point to point integration and enterprise application integration (EAI), integrates relative data by linking applications through custom-coding or integration broker that acts as a hub to route messages between connected applications. Tight-coupling of technology and systems, N^2 problem at data layer, multiple vendors for multiple systems, lack of common protocols make these solutions difficultly suitable to an open, dynamic environment of businesses and operations in or across enterprises.

DCI can be implemented by creating either a centralized repository for data access and analysis, such as data warehouses, or a data integrating layer over a set of distinct and autonomous data sources, such as federated information systems. Although some researches on DCI follow a semantic integration approach, they are hard to implement interoperability across corporate boundaries or between them and other information systems because of the using of proprietary metadata, representation.

2.2 Data Integration on Grid

In the context of information sharing, Grid technologies distinguish current technologies in enterprise as discussed above by providing not a generic approach but also an open and standard-based infrastructure.

The current efforts of the Data Grid community concentrate on providing a global, uniform access methodology for all database resources. Meanwhile, information grid projects shift the emphasis on information integration and mediation. Moreover, the emerging of Semantic Grid is beginning to take this further, from information to semantic or knowledge. Some projects, such as COG [4] and Dart-Grid [5] explore this trend in the context of information integration.

The COG project aims to integrate disparate data sources on semantic level by using a central Information Model (i.e. ontology). However, although COG means "Corporate Ontology Grid", it does not seem to intend to use general Grid technologies. In essence, it is a solution following an ontology-based information integration approach [6]. Compared to COG, Dart-Grid is an OGSA-based Database Grid originally motivated by the application of web-based data sharing and database integration for Traditional Chinese Medicine. In particular, data sources integrated by Dart-Grid are mainly databases, other data sources such as documents, and data sources that stream data in real or pseudo-real time from applications are not supported by current Dart-Grid. Furthermore, details of some crucial issues concerned by enterprise, such as security, authorization, transaction, etc., are not addressed.

Furthermore, semantic interoperation based on single ontology such as in COG and Dart-Grid may fall short where one information source has a different view on a domain, or information sources are on changing [6]. Accordingly, in this way, it is difficult to achieve semantic interoperability across department and corporate boundaries, realize inter-domain information sharing.

Based on the review of the survey, we observe that common infrastructure and generic approach in the context of semantic information integration for enterprise business and daily operation have to date not been addressed in semantic gird community. This is a key contribution of our project.

3 Vision

SGII, described in this paper, is a generic, interoperable, pervasive, open and flexible Semantic-Grid-based infrastructure for EII in a generic way, while creating a semantic interconnection environment for enterprise information sources. The vision of SGII focuses on four aspects:

1) For infrastructure, SGII wants to provide an open platform that enables distributed information sources to be shared, managed, coordinated, and controlled both on system and semantic level.
2) For enterprise data sources, SGII aims at making them machine understandable and interoperable by enriching them using semantics and wrapping them using services.
3) For applications in enterprise, SGII expects to provide understandable and operable information and services for individual and common application domains.
4) For users in enterprise, SGII can provide a global, real time, and 360-degree view of enterprise information on semantic level.

4 Elementary Architecture

To enable the vision of SGII, we propose the elementary architecture of SGII based on the Open Grid Service Architecture [7], which makes SGII a generic infrastructure for semantic integration of enterprise information. Every function in SGII is independently realized as a grid service on top of the Open Grid Service Infrastructure and translated into semantic grid service through semantic enrichment. In essence, SGII is semantic grid service oriented architecture.

As shown in figure 1, the architecture consists of four spaces or layers: *Data Sources Space*, *Support Service Space*, *Mediation Service Space* and *Application Support Service Space.* Except Data Sources Space, the services of low-space provide more common functionality than the services of upper-space, and help reducing complexity of upper-space services.

Data Sources Space is a collection of data sources which can be classified into two categorizes: *Grid-Enabled Information Repositories* and data sources that *Stream Data from Applications*. Note that we regard stream data from applications as data

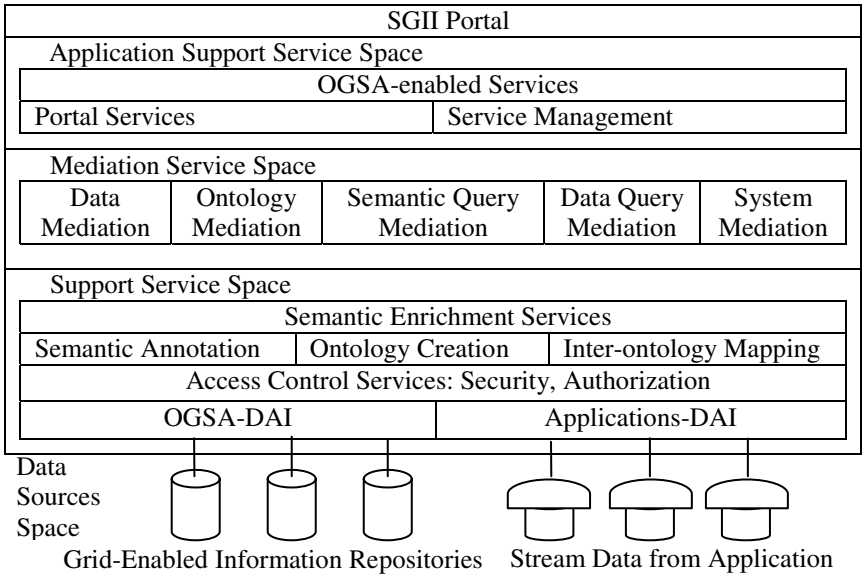

Fig. 1. Elementary Architecture of SGII

sources because there are growing needs to integrate results from applications. In particular, we may also integrate data through application when the direct access of data source is forbidden. This is the common occasion in an enterprise. The Grid-Enabled Information Repositories generally refer to the storage of all information relevant to enterprise businesses and daily operations, including data and metadata. Moreover, the information produced by SGII (including management information, operational information, and metadata for data sources and services, etc.) is also stored into grid-enabled information repositories.

Support Service Space constructs common and basic services for SGII by providing *Access Control Services* and *Semantic Enrichment Services*.

The Access Control Services are responsible for the guarantee of flexible information access control with the support of *OGSA-DAI* [8] and *Applications-DAI*. Based on OGSA, the Data Access and Integration (OGSA-DAI) infrastructure can model heterogeneous data sources as grid services and integrate them into SGII architecture. Analogy with OGSA-DAI, Applications-DAI (Data Access and Integration of application data) is responsible for wrapping data sources that stream data from applications as OGSA-compatible grid services and integrate them into SGII architecture. Furthermore, OGSA-DAI and Applications-DAI provide basic data access capability to SGII.

Semantic Enrichment Services deal with relative topics of semantic enriching, which include *Semantic Annotation*, *Ontology Creation*, and *Inter-ontology Mapping* services. Semantic Annotation service is used to enrich data and service content based on pre-defined ontologies by using automatic or semi-automatic semantic annotation tools. Ontology Creation service wraps a semi-automatic approach as OGSA-compatible grid service to create ontologies on the basis of a process of

semantic discovering, explanation and user modeling. These ontologies constitute the conceptual backbone for information-level and semantic-level interoperability, which is the basis of semantic enrichment. To achieve pervasive interoperability among (extra-) enterprise information sources, SGII requires the support of multi-ontology. Accordingly, Inter-ontology Mapping service makes this to be reality by creating interconnection semantics based on semantic mapping between ontologies.

Mediation Service Space provides services to mediate conflicting processes, semantics, and information without custom written programs for the purposes of information sharing by operationalizing system, ontology and data sources. It includes five categories of services: *Data Mediation*, *Ontology Mediation*, *Semantic Query Mediation*, *Data Query Mediation* and *System Mediation*. Data Mediation is an approach to simplify the work with multiple, federated, and usually geographically distributed data sources. Built on top of OGSA-DAI, it is capable of mediating any grid enabled information repositories by mediation data models or ontologies, of which data source schemas are defined as the views. Ontology Mediation, together with Inter-ontology Mapping service in Support Service Space, is used to facilitate creating and maintaining consistency semantic interconnections between ontologies. It uses semantic relationships as semantic mediation model rather than a global ontology. Semantic Query Mediation performs semantic query following a mediation centric approach based on ontology mediation mechanism. This service decomposes semantic queries into sub-semantic-queries on corresponding ontologies. Finally, each sub-semantic-query is translated into sub-data-queries on corresponding data sources according to the semantic relations between ontologies and data sources. System Mediation Service is used to monitor, tune, and manage system transactions and resolve conflicts between transactions, such as synchronal happening of cross-data query and update on the same data sources. Furthermore, it is also in charge of dynamic clustering process of data and services for specific information integration and sharing task.

Application Support Service Space provides services called *Portal Services* for front end enterprise applications (e.g., e-commerce, B-to-B, CRM, ERP, EIP, etc.). In order to ensure that relative services in SGII can get in enterprise business and operation environment, SGII uses Portal Services on top of OGSA as delegations or broker services, which enables enterprise applications to use these services in a standard and easy way as long as they observe compatible specifications of OGSA. *Service Management* service in this space is responsible for the management of each portal service.

5 Conclusions

The application based on Semantic Grid concept for provision of effective information resource sharing in and across industrial enterprises is promising. However, it needs more industrial case studies to be carried out in order to evaluate finally the utility of information and semantic interoperability based on SGII. Many relevant aspects of a complete SGII integration environment are not addressed in this paper so far. We will show more details in our further work.

References

1. Schreiber Z.: Semantic Information Management (SIM): Solving the Enterprise Data Problem by Managing Data based On Its Business Meaning. Unicorn Executive White Paper (2003)
2. Hannus, M., Penttilä H., Silén P.: Islands of Automation. http://cic.vtt.fi/hannus/islands.html (1987 - updated by Hannus, M. 2002)
3. Hai Z.G.: Semantic Grid: Scientific Issues, Infrastructure, and Methodology. Communications of the ACM, Vol. 48, No. 4 (2005): 117–119
4. Bruijn J.: Semantic Information Integration in the COG project. Technical report by IFI (2004)
5. Zhaohui W., Huajun C., Shuiguang D., Yuxing M.: DartGrid: RDF-Mediated Database Integration and Process Coordination Using Grid as the Platform. Lecture Notes in Computer Science, Vol. 3399. Springer-Verlag, Berlin Heidelberg New York (2005) 351–363
6. Wache, H., Voegele, T., Visser U., Stuckenschmidt. H., Schuster, G., Neumann, H., Huebner, S.: Ontology-based integration of information - a survey of existing approaches. In: Proceedings of IJCAI workshop on Ontologies and Information Sharing (2001):108–117
7. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the grid: An open grid services architecture for distributed systems integration. In: Open Grid Service Infrastructure WG, Global Grid Forum (2002)
8. Atkinson, M., et al.: Data access, integration, and management in The Grid. Blueprint for a New Computing Infrastructure. San Mateo, CA: Morgan Kaufmann (2004)