Data Integration for Virtual Enterprise in Cyberworlds

Cheng Leong Ang, Robert Gay, and Olga Sourina
Nanyang Technological University, Singapore
eosourina@ntu.edu.sg

Abstract

Virtual Enterprise enables companies to collaborate through sharing of resources, and therefore allows them to enjoy the benefits of virtual integration. The essence of Virtual Enterprise implementation on the Web lies in data integration. This paper describes key research issues in implementation of data integration in virtual enterprise using semantic web service. The architectural framework of the proposed data integration infrastructure adopts a mediated ontology approach to data integration in which each data source is described by its own ontology and translations between different ontologies are by means of mediation. We also propose to use the concept of an active data warehousing system for virtual enterprise data aggregation. The paper briefly describes the preliminary study for the project proposal.

1. The Semantic Web Service Infrastructure

Virtual integration can be looked at as Cyberworlds creation on the Web. Virtual enterprise enables companies to collaborate through sharing of resources, and therefore allows them to enjoy the benefits of integration. In addition, companies can also take full advantage of cheaper and higher-quality goods and services as they become available, which is not possible when internal sources are the only sources of supply. Finally, virtual integration also reduces their risk exposure to changing technologies or changing buyer preferences and enhances their ability to assemble diverse kinds of expertise speedily and efficiently. However, none of these can happen without efficient data integration. For many enterprises, especially the smaller ones, the proposed Semantic Web Service infrastructure for automatic data integration offers the only means to realize the full benefits of virtual integration and to enable them to collaborate and compete effectively in the dynamic global market.

The Semantic Web Service has the potential of becoming the most powerful technology for the virtual integration of enterprises. There are two equally important and complementary aspects of virtual integration, namely process integration and data integration. While much research has done into the use of the Semantic Web Service for automating process integration, the same is not true of its use for automating data integration. Automatic data integration on a global scale is especially important for virtual enterprises operating in a dynamic global market. The large number of data sources they need to access, the evolving nature of the data sources, and the changing business environments all render manual data integration infeasible. By offering data sources as web services, the so-called data services, the data sources can be made accessible globally. However, today’s Web is meant only for human interpretation and use. Web data lacks machine-understandable semantics, so it is very difficult or impossible to achieve automatic data integration with the Web Service. The Semantic Web Service could overcome the limitations.

What is needed therefore is a Semantic Web Service infrastructure that can support automatic data integration. As proof of concept, we propose to develop a prototype of the Semantic Web Service infrastructure for automatic data integration on a global scale. The work will involve detailed design and implementation of the prototype, and solving many new challenging research issues for the prototype implementation. We have developed an architectural framework for the proposed infrastructure and have conducted a detailed literature survey to identify what have been done and what need to be done, as well as potential research issues for the development of the infrastructure.

2. Architectural Framework

The architectural framework of the proposed data integration infrastructure as shown in Figure 1 adopts a mediated ontology approach to data integration in which each data source is described by its own ontology and translations between different ontologies are by means of mediation. The framework is based on the fact that most, if not all, industries still do not have a common shared ontology in use to date. Due to the difficulty of building a consensus about what terminologies and structures should be used, it is doubtful if such ontology will ever exist. In fact, it is more likely that there will be many ontologies to be mediated among. It is also likely that the number of such ontologies will increase with increasing business in the industry and many of them will also change over time. In such a situation, it is almost impossible to build a
complete global ontology for any industry. So the user may have to start with an incomplete ontology. What he can do is to update and improve on it, making it more complete while it is being used. The proposed framework has taken that into consideration. It not only can support incremental building of ontologies, but has also the ability to decide if human intervention is required in order to avoid erroneous results, and hence any possible disastrous consequences, due to the use of incomplete ontologies.

2) A systematic methodology for ontology building.
The proposed infrastructure is basically an ontology-based data integration system. For such a system, it is important to support the user with a genuine methodology and software tools for building and using ontologies.

The mediated ontology approach adopted imposes three specific requirements on the methodology: 1) the methodology must be able to support the systematic integration through mediation of a large number of existing ontologies from different domains, 2) it must be able to support incremental building of ontologies and handle ontology evolution, and 3) it must be based not just on theoretical analysis, but also on empirical analysis in order to be able to integrate local data sources in the real-world. Of course, to be comprehensive, the methodology must also be able to handle all types of sophisticated mappings (i.e. non 1-1 mappings) between ontologies, and ontology constraints. The development of a systematic methodology will be a major step in the work on using ontologies for automatic data integration.

For the development of the methodology, we propose to explore the use of an integrated Idef0/Idef1x methodology developed in [3] for building data models. The integrated Idef0/Idef1x methodology uses a top-down approach for domain analysis, and a bottom-up approach for building a global data model. It allows the global data model to be built incrementally and is able to handle changes in local data models.

3) Automatic generation of ontologies. Although the use of ontologies to integrate data sources has many advantages in terms of information sharing among applications, the creation of ontologies is not a trivial matter. It is a difficult, tedious, time-consuming and therefore an expensive process. Creating ontologies manually is clearly not feasible on the Web scale. The need to automate ontology generation for data integration

3. Key Research Project Issues

Although fully automatic data integration may not be possible in the dynamic environment considered, we should be able to achieve a high degree of automation requiring only little human intervention through ontology learning. The work here involves detailed design and prototype implementation of the proposed Semantic Web Service infrastructure for automatic data integration, as well as solving new challenging research issues. The proposed research into the data integration infrastructure faces many challenges because of the heterogeneity of the data sources and because of the mediated ontology approach adopted. We defined the following research issues in our project:

1) An agent-based approach to detailed design and implementation. The proposed infrastructure is meant for large-scale data integration in the Internet, where new data sources are being created in an ever-increasing rate, and where existing data sources and business requirements for Web services data are changing constantly. Given this scenario, it is extremely difficult to integrate data using traditional approaches based on centralized and predefined control mechanisms.

To meet this need, we propose to develop an agent-based infrastructure for data integration [1]. This requires, among other things, methods and technology for the automatic control, maintenance and evolution of the adaptive mobile agent systems [2]. Our vision is a data integration infrastructure that is dynamic, autonomous, self-organizing and proactive, i.e. an intelligent data integration infrastructure.

Figure 1. Overall architectural framework of the proposed Semantic Web Service infrastructure for automatic data integration

![Figure 1. Overall architectural framework of the proposed Semantic Web Service infrastructure for automatic data integration](image-url)
has previously been considered in the literature [4-5]. To meet the need, we propose to explore the use of ontology learning [4], specifically the use of the multitude of automation techniques from the many different disciplines for ontology extraction, ontology mapping and merging, and ontology evaluation. Examples of such techniques include computational linguistics, machine learning, information retrieval, graph matching, databases etc. Despite extensive research, there are still many open research problems in these ontology areas. There is to date not a comprehensive domain-independent method for learning ontology mappings, i.e. one that can handle all types of mapping conflicts and ontology constraints [6]. Our aim is to develop a prototype of an ontology-learning system for the semi-automatic, if not fully automatic, generation of ontologies, and to implement it as an integral part of the proposed ontology-based data integration system.

4) **Ontology extraction from computer applications.** For the reasons given above, there has been a significant amount of research into techniques whereby ontologies can be extracted automatically from data sources, whether structured, semi-structured or unstructured. However, we feel that research efforts here should also be expanded to cover ontology extraction from computer applications, and be directed towards achieving automatic generation of query ontologies for data sources directly from the applications to achieve automatic process-data integration. In view of that, we propose to explore the possibility of mining application programs for ontological information. What we would like to see here is a software system that can automatically analyze an application program, extract its ontology, and formulate a query for data sources, whether structured or semi-structured, by reformulating the very data input statements of the application program in terms of its ontology.

Here applications can be conveniently grouped into three categories: those that are expressed in some dedicated knowledge representation languages like Prolog, those in O-O languages like C++, and those in conventional procedural code such as C. First two categories can offer a more immediate and more fertile ground for ontological extraction than the last. Some preliminary work on ontology extraction from Prolog programs was done, investigation of ontology extraction from O-O programs has yet to be carried out.

5) **Automatic data extraction, transformation, and aggregation.** In the proposed framework, data are extracted from the relevant data sources in response to a user/application query. Once they are extracted, they are mapped into a common target schema, and converted to the required data type and format before they are aggregated. The process concerns three main areas of research, namely automatic data extraction, automatic schema matching, and automatic data conversion. There is a very significant amount of research work in the first area, especially in the extraction of data from structured data sources like relational or O-O databases, and from semi-structured data sources like Web pages or HTML documents/tables. The second research area concerns basically problems of ontology-schema matching and schema-schema matching. As schemas are ontologies with restricted relationship types, the problems can therefore be viewed as subsets of the problem of mapping between ontologies already considered above. In our project, we propose to use an active data warehousing system based for data extraction, transformation, and aggregation. The conceptual architecture is shown in Figure 2.

![Figure 2. A conceptual architecture for an active data warehousing system](image)

6) **Automatic generation of semantic descriptions of data sources.** For the Semantic Web Service to flourish there must be a sufficiently large number of publications of service semantic descriptions in the common service registry or “UDDI”. As far as Semantic Web Service-based data integration is concerned, this can only be realized if the process of generating the semantic descriptions is automated. The process involves basically the extraction of ontologies from data sources. It is a tedious and time-consuming process. The process can however be automated through ontology learning. Here we propose to explore the use of machine learning and other AI techniques for learning the extraction of ontologies. Our aim is to develop a prototype ontology learning system for the automatic generation of comprehensive semantic descriptions of data sources and their access methods. The ontology learning system developed here can be adapted to the requirements of overall ontology learning system we described above.
7) Automatic query processing. When a query is generated, its ontology is matched against the ontologies of data sources published in the service registry to locate the relevant data sources. The number of data sources that need to be accessed is then contained to a minimum by eliminating those that are “redundant.” To speed up query processing, the query is then re-formulated in terms of the selected data sources. Data are then extracted from the data sources by invoking the respective data access and extraction services. The data are then aggregated to produce the query result. To automate and optimized query processing, we propose to explore the use of query containment and query rewriting for answering queries. Query containment is the problem of checking whether, for a given database, the result of one query is a subset of the result of another query; whereas query rewriting is the problem of decomposing a query into sub-queries that access the data sources and produce the same result as the query. The problems of query containment and query rewriting are relatively simple in the case of data integration based on the federated ontology approach, but not so in our case of data integration which is based on the mediated ontology approach involving answering queries using views [7]. We are also aware that XQuery is to become a standard language for querying XML data. In view of these, we would like to propose the development of a query containment algorithm and a query-rewriting algorithm for XQuery to be used in data integration involving answering queries using views.

8) Automatic generation of semantic descriptions of geometric data sources. Besides documents, files, Web pages etc, there are other data sources such as graphics images, 3-dimensional objects, video and audio files etc. Extraction of ontologies from these data sources is even more complicated. The most difficult problem is the ontology extraction from geometric data that are shared through the Web and can come from different applications in manufacturing like 2D and 3D CAD models, or bio-medical engineering like CT, MRI, X-rays data and/or molecular structures, or even from entertainment industry. Though there are tools for the visualization and interaction with geometric data in the web environment like, for example, VRML for 3D objects, there is none for extraction of ontologies from these data. In view of this, we therefore propose to develop methods and tools to support automatic extraction of ontologies from geometric data.

4. Conclusion

Academically, it is hoped that the proposed research will contribute significantly to the advancement of knowledge in automatic data integration, automatic query processing, and automatic data extraction, as well as in the related areas of ontology modeling, and automatic ontology generation. Industrially, it is hoped that the research will not just contribute towards laying the foundation for possible future implementations, but will actually lead to some sort of actual implementations and will therefore help to realize our vision of a hyper-efficient local industry where enterprises, big and small, can collaborate and compete effectively in global market through resource sharing made possible by automatic data integration.

Besides virtual integration of enterprises, the proposed data integration infrastructure can also be used to support e-Manufacturing, e-Commerce, e-Supply Chain, Business Process Management Systems, Collaborative Product Development etc.

6. References