Towards an Ontology-based Mediation Framework for Integrating Biological Data*

Amine Kerzazi, Ismael Navas-Delgado, José F.Aldana-Montes

E.T.S. Ingeniería Informática
Universidad de Málaga, Campus de Teatinos, 29071 Málaga
{kerzazi, ismael,jfam}@lcc.uma.es

Abstract. In the context of Life Sciences, the frame of Systems Biology is emerging. It is supported by all high-throughput methods which generate large amounts of data that cannot be processed simply by the human mind. The integration of data from heterogeneous knowledge sources involves the consolidation of heterogeneous data geared at generating new knowledge that can not be obtained from single data sources. In this paper, we introduce the new improvements in the mediator components, their function and importance for biological data integration.

Key words: Semantic Directory, ontology, Data Integration, mapping.

1 Introduction

Integration of data from heterogeneous knowledge sources represents the consolidation of heterogeneous data geared at generating new knowledge that can not be obtained from single data sources. The field of data integration in the Semantic web has gained popularity in recent years; integrated access to multiple distributed and autonomous data sources is a key challenge for many semantic web applications. In this paper, we introduce an ontology-based mediator framework (Khaos Ontology-based Mediator Framework *http://khaos.uma.es/KOMF/*[1]) which uses a generic infrastructure to register and manage ontologies, their relationships and also information relating to the resources. KOMF has been successfully instantiated in the context of Molecular Biology for integrating dispersed data sources. The most important proposal to solve the data integration problem is the wrapper/mediator architecture. In this architecture, a mediator (an intermediate virtual database with a schema G according to a previous definition of the data integration system) is established between data sources (with a set of schemas S) and applications.

The study of data integration proposals has enabled the design of the novel architecture proposed, KOMF. This architecture can be used to develop different data integration systems, and can even be used to emulate existent systems.

* Supported by P07-TIC-02978 (Junta de Andalucía) and TIN2008-04844 (Spanish Ministry of Education and Science).

This architecture can be used for any kind of database system, including centralized, distributed, or parallel systems. The main purpose of our architecture is to provide a semantically unified interface for querying heterogeneous information sources. The architecture is composed of four major kinds of elements (Figure 1):

- Semantic directories[2] store and manage meta-data concerning a number of domain ontologies, as well as the relationships among the ontologies and the data source/databases.
- Data services include the access to data sources/databases that could be queried through the Web.
- The mediator provides a way of using queries from the applications to produce integrated results.
- Applications should provide end-user-focused interfaces, so users do not need to know that a mediation system is being used.

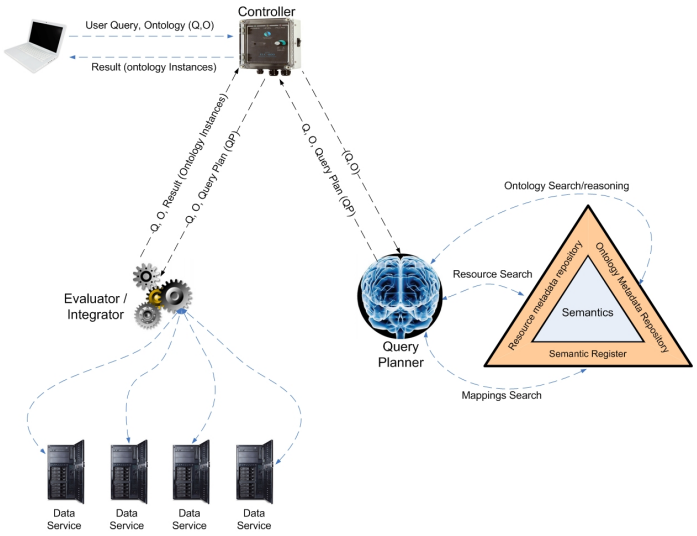

Fig. 1. Figure 1: KOMF architecture.

2 Use cases

We have started working on a pilot system called the Amine System Project (ASP, *http://asp.uma.es/WebMediator*) for the integration of biological information, related to Biochemistry, Molecular Biology and Physiopathology, of a group of compounds known as biogenic amines. Two general objectives can be distinguished in this project:

1. Development of new and more efficient tools for the integration of information stored in databases, with the aim of detecting new emergent properties of this system.
2. Generation of in silico predictive models at different levels of complexity. It is being carried out by a multi-disciplinary group consisting of biochemists, molecular biologists and computer scientists.

The main activity carried out in this pilot has been the development of a prototype for solving a specific problem. The result is AMMO-Prot, the ASP Model Finder. The problem to be solved by this tool is the following:

A common and useful strategy to determine the 3D structure of a protein, which cannot be obtained by crystallization, is to apply comparative modelling techniques. They start working with the primary sequence of the target protein to finally predict its 3D structure by comparing the target polypeptide to those of solved homologous proteins.

Another tool developed taking advantage of KOMF has been the System Biology Metabolic Modelling Assistant (*http://www.sbmm.uma.es*)[2], which is a tool developed to search, visualize, manipulate and annotate identity data and assist in annotating the kinetic data. The inputs to search pathways are the pathway's name or Kegg code, a set of enzymes or a correctly annotated model (homemade or not). The application queries KOMF via conjunctive queries and obtains a set of RDF instances. Results are rebuilt to a format that is understandable for the application and the user can ask for information on enzymes, compounds and reactions. Users can also edit the pathway freely or using an assistant, adding well formed kinetic rules. At the end of the process the user can export the pathway to SBML format, enriched automatically without any previous configuration.

3 Discussion and Conclusions

This paper has introduced KOMF, an ontology-based mediation framework, which uses ontologies as mediated schema. Since the mediated schema is an ontology, queries are created over the ontology that constitutes the mediated schema and results are ontology instances. The use of ontologies enables reasoning to be included at different levels, making it possible to infer new knowledge. This framework has been validated in molecular biology and systems biology. In this context we have defined a domain ontology and a set of data sources has been registered and successfully integrated. KOMF it is used for integrating data from different biological information related to Biochemistry, Molecular Biology and Physiopathology of a group of compounds known as biogenic amines. It is also used by The System Biology Metabolic Modeling Assistant to search, visualize, manipulate and annotate identity data and assist in annotating the kinetic data. As future work, our intention is to infer information that is not stored anywhere (but is a logical consequence of the stored one) by using reasoning. Furthermore, we are defining how to integrate data transformation tools in order to enable the

transformation of integrated data for solving more complex tasks (like protein structure prediction, protein alignment, etc.).

References

1. Othmane Chniber; Amine Kerzazi; Ismael Navas-Delgado y José F. Aldana Montes. KOMF: the Khaos ontology-based mediation framework. En Luciano Milanesi and Paolo Romano (eds.). Bioinformatics Methods for Biomedical Complex System Applications. 19-21 May 2008, Villa Monastero, Varenna, Italy. págs. 57-60. NETTAB, 2008.
2. Navas-Delgado I; Kerzazi A; Chniber O y Aldana-Montes J. SD-CORE: a semantic middleware applied to molecular biology. In Proceedings of On the Move to meaningful Internet Systems: OTM Workshops; 9-14 November 2008; Monterrey. 2008:976-985.
3. Reyes-Palomares A, Montanez R, Real-Chicharro A, Chniber O, Kerzazi A, Navas-Delgado I, Medina MA, Aldana-Montes JF, Sanchez-Jimenez F: Systems biology metabolic modeling assistant: an ontology-based tool for the integration of metabolic data in kinetic modeling. Bioinformatics 2009, 25:834-835.