

BioSO: Bioinformatic Service Ontology for Dynamic Biomedical Web Services Integration

Ramez Elmasri, Jack Fu, Feng Ji, Qing Li
Department of Computer Science and Engineering
University of Texas at Arlington

P.O. Box 19015, Arlington, TX 76019, U.S.A

Email: {elmasri, fusheng, jifeng}@cse.uta.edu, cli@accurohealth.com

Abstract—Web services have recently become a new trend for gathering biomedical information. However, it is not easy to integrate and obtain a concise/complete query result among hundreds of services. In this paper, we propose a multi-level service integration architecture for dynamically integrating web services in the biomedical domain. Our ultimate goal of BioSO system is to create a unified, public, scalable, and interoperable biomedical service platform to benefit scientists in data searching and publishing.

Keywords: biomedical, web service, multi-level integration

I. INTRODUCTION

Biological and medical research creates large amounts of data spread over diverse databases such as GenBank, PDB, etc., which need to be processed, integrated and organized in order to query them efficiently. Figure 1 illustrates integration strategies from three perspectives (instance layer, schema layer, and service layer). Traditionally, computer scientists pursue a schema layer integration on relation/attribute names among various database schemas, while life scientists also focus on the instance layer by seeking a universal agreement on identification of biogenetic entities such as International Protein Index (IPI) [1] and Life Science Identifier (LSID) [2] standards in order to achieve the goal of integration. Some other integration work also has been done as mentioned in [3], [4], [5], [6], [7], and most of this work focuses on integration based on the instance and schema layers. However, there are some inevitable drawbacks in instance and schema layer integrations.

For Instance layer integration, since most of the biological databases have different schema designs and their own identifiers for the same biological entities [8], [9], the identifier of each biological entity needs to be updated and maintained frequently in order to ensure data consistency and integrity. Schema layer integration requires the detailed schema information of each data source and relies on schema matching and reasoning techniques [10]. In addition, each data source may employ different modeling techniques to create their schema such as Entity Relationship (ER) [11], XML, RDFS, etc. This increases the difficulty of integrating work at the schema layer. Due to the nature of life science data (highly distributed, dynamic, complex, incomplete, heterogeneous)[12], we focus on service layer integration instead of directly dealing with schemas and data. In this paper,

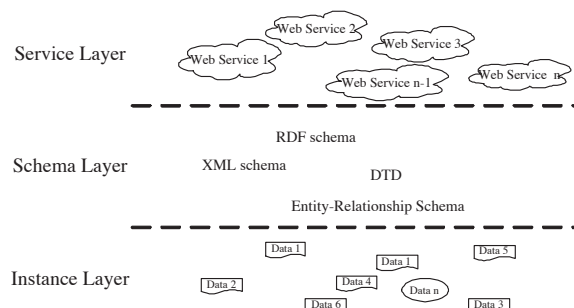


Fig. 1. Three Integration Layers

we propose a multi-level biomedical web service integration architecture, which incorporates multi-level modeling concept [13] with web service integration. Some similar work has been proposed in genome research such as SPDBSW[14] and DAS (Distributed Annotation System) [15] that defines a communication protocol to allow exchange of biological sequence annotations between client and server in order to achieve resource integration. It has been employed in UCSC, UniProt, Ensembl, Flybase, Wormbase, etc. Our multi-level web service integration approach is more flexible, versatile, and platform independent as it can provide loosely coupled integrated across different application domains through some related industrial and academic standards such as XML, SOAP, WSDL, UDDI, etc.

The rest of this paper is organized as follows. In section 2, we briefly introduce some web services in the biomedical domain. In section 3, we propose a system architecture in order to integrate biomedical web services and describe seven main components in our framework. we conclude this paper and discuss directions for future work in Section 4.

II. BIOINFORMATIC WEB SERVICES

In this section, we will give a few typical web service examples in the biomedical domain and describe the capability of each. (Because of space limitation, we just briefly describe two web services here).

A. NCBI Web Service

NCBI Entrez databases play a key role in the searching of portals for retrieving many biological/biomedical resources.

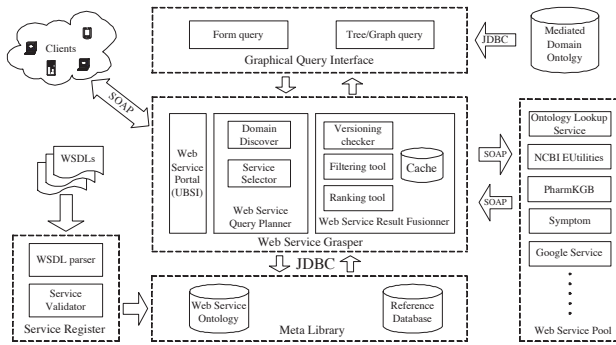


Fig. 2. BioSO System Architecture

NCBI provides web services that allow users/developers to access Entrez utilities (such as ESearch, EInfo, ESummary, etc.) via SOAP[16].

B. PharmGKB Web Service

PharmGKB database is developed by Stanford University for storing genomic, molecular, cellular phenotype, clinical, pharmacokinetic and pharmacogenomic information.

III. WEB SERVICE INTEGRATION IN BIOMEDICAL DOMAIN

A. System Architecture

Figure 2 demonstrates the architecture of our ongoing web service integrating system (BioSO). There are seven main components in this system:

- 1) GQI (Graphical Query Interface) allows the user to generate a query by filling a query form or choosing concepts/relationships through a graph.
- 2) MDO (Mediated Domain Ontology) is a data source providing concept/relationship information for the GQI module.
- 3) WSQP (Web Service Query Planner) is responsible for generating web service query plans based on the domain and level/classification properties of the given query.
- 4) WSRF (Web Service Result Fusionner) module filters results gathered from different web services, compares versions of the results for the same biological entity, and checks consistency/integrity among them.
- 5) WSO (Web Service ontology) is the key component in our framework, which not only keeps the WSDL description of each web service, but also records classification (level/domain) and service relationship information. It provides essential functions/data needed by other components with the purpose of completing the query process successfully.
- 6) RD (Reference database) is a cross-reference depository, which keeps identifier mapping information among diverse databases such as Gene ontology, GeneBank, PDB, KEGG, PharmGKB, etc. By referring to this component, our system can establish solid connections for the same biological entity among services.

TABLE I
DOMAIN/LEVEL INFORMATION

Domain	Level
disease	tissue, organ
drug	undefined
biological	gene

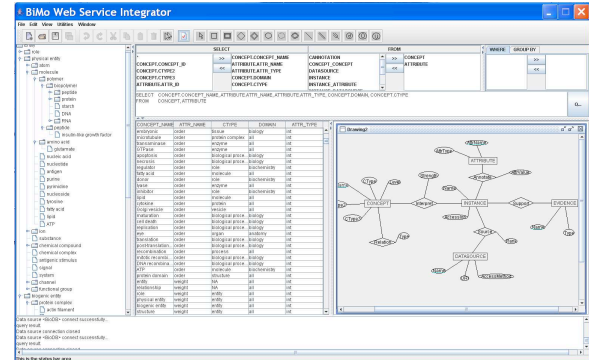


Fig. 3. BioSo System prototype

- 7) WSP (Web Service Portal) is designed for achieving interoperability, reusability, and scalability among services across a varied base of underlying and changing technologies by providing an UBSI (Unified Biomedical Service Interface). A client side can remotely invoke a service through WSP module, and is not required to deal with complex/inconsistent interfaces of each web service. WSP will look up the WSO module and dynamically invoke corresponding services for a client.

A more complete description of these components is given in the full paper.

EXAMPLE. Find the genes and symptoms related to Colorectal Neoplasms, and drugs that can be used for treatment.

User submits above query through GQI module, and then WSQP module partitions the query into sub-queries based on domain and level information as shown in table I. WSQP module will look up WSO for choosing related web services. The last job for WSQP module is to dynamically create signatures and invoke corresponding remote services. Once WSRF module gathers results from each service, it will pass into versioning, filtering, and ranking tools before presenting the final integrated result.

Figure 3 shows the prototype of our system, and we are currently developing the web service ontology.

IV. FUTURE WORK AND CONCLUSION

In this paper we address the problem of Integration in the biomedical domain from different perspectives (instance, schema, and service layers). We proposed a multi-level service integration architecture for integrating web services in biomedical domain. Our architecture can dynamically search and invoke remote web services. In addition, we will define a UBSI (Unified Biomedical Service Interface) based on existing service interfaces, flexibility, and performance concerns. Also,

we need to define ranking and filtering strategies for WSRF module. Our architecture can be used with other mediator systems for cross referencing the diverse bioinformatics sources and be utilized in other applications such as system biology modeling and health care system. Due to the space limitations, we cannot include the detailed functions and implementation issues in this extended abstract.

REFERENCES

- [1] P. J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler, "The international protein index: An integrated database for proteomics experiments," *Proteomics*, vol. 4, no. 7, pp. 1985–1988, 2004.
- [2] OMG, "Life sciences identifiers specification," 2005. [Online]. Available: <http://www.omg.org/docs/dtc/04-05-01.pdf>
- [3] T. Hernandez and S. Kambhampati, "Integration of biological sources: current systems and challenges ahead," *SIGMOD Rec.*, vol. 33, no. 3, pp. 51–60, September 2004. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1031583>
- [4] V. M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, N. Ivanova, and N. C. Kyrpides, "The integrated microbial genomes (img) system: a case study in biological data management," in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 2005, pp. 1067–1078.
- [5] V. Jakoniene and P. Lambrix, "Ontology-based integration for bioinformatics," in *Proceedings of VLDB Workshop on Ontologies-based techniques for DataBases and Information Systems, Trondheim, Norway (2nd–3rd September 2005)*, 2005.
- [6] P. Mork, R. Shaker, and P. Tarczy-Hornoch, "The multiple roles of ontologies in the biomediator data integration system." in *DILS*, 2005, pp. 96–104.
- [7] S. TriSZI, K. Rother, H. Muller, T. Steinke, I. Koch, R. Preissner, C. Frommel, and U. Leser, "Columba: an integrated database of proteins, structures, and annotations," *BMC Bioinformatics*, vol. 6, no. 1, p. 81, 2005. [Online]. Available: <http://www.biomedcentral.com/1471-2105/6/81>
- [8] S. Drăghici, S. Sellamuthu, and P. Khatri, "Babel's tower revisited: a universal resource for cross-referencing across annotation databases," *Bioinformatics*, vol. 22, no. 23, pp. 2934–2939, 2006.
- [9] A. C. R. Martin, "Mapping pdb chains to uniprotkb entries," *Bioinformatics*, vol. 21, no. 23, pp. 4297–4301, 2005.
- [10] A. Y. Halevy, "Answering queries using views: A survey," *The VLDB Journal*, vol. 10, no. 4, pp. 270–294, 2001.
- [11] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006.
- [12] S. Potter and J. S. Aitken, "A semantic service environment: A case study in bioinformatics." in *ESWC*, 2005, pp. 694–709.
- [13] R. Elmasri, J. Fu, and F. Ji, "Multi-level conceptual modeling for biomedical data and ontologies integration." in *CBMS*, 2007, pp. 589–594.
- [14] T.-S. Jung and W.-S. Cho, "Spdbsw: A service prototype of spdbs on the web." in *BNCOD*. Springer-Verlag Berlin Heidelberg, Inc., 2007, pp. 49–57.
- [15] BioDAS, "<http://www.biodas.org/>," 2007.
- [16] W3C, "Simple object access protocol (soap)," 2007. [Online]. Available: <http://www.w3.org/TR/soap/>