International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:1, 2011

Biological Data Integration using SOA

Noura Meshaan Al-Otaibi and Amin Yousef Noaman

***Abstract*—**Nowadays scientific data is inevitably digital and stored in a wide variety of formats in heterogeneous systems. Scientists need to access an integrated view of remote or local heterogeneous data sources with advanced data accessing, analyzing, and visualization tools. This research suggests the use of Service Oriented Architecture (SOA) to integrate biological data from different data sources. This work shows SOA will solve the problems that facing integration process and if the biologist scientists can access the biological data in easier way. There are several methods to implement SOA but web service is the most popular method. The Microsoft .Net Framework used to implement proposed architecture.

Keywords—Bioinformatics, Biological data, Data Integration, SOA and Web Services.

I. INTRODUCTION

RECENT years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced, and protein and gene interaction data are accumulating [1]. These biological data is available in a wide variety of formats, annotated, and stored in flat files and relational or object-oriented data bases. The value of any kind of data is greatly enhanced when it exists in a form that allows it to be integrated with other data. An important aspect of bioinformatics consists in building a scientific digital library, integrated view of all data of interest widely distributed and constantly updated in heterogeneous remote public data sources or local private ones. Access to heterogeneous biological data sources is mandatory to scientists. A single query may involve flat files such as GenBank or SwissProt, web resources, or the references data source PubMed [2, 3].

Integration of biological data is just one phase of the entire molecular biology research and genomic hypothesis discovery process [6].

Several works showed that the integration of heterogeneous bio-molecular data sources can significantly improve the performances of data mining and computational methods for the inference of biological knowledge from the available data. Also, Integration of biological data allows uniform access of federation of several data sources [4]. Despite the importance, the following challenges make data integration one of the longest standing problems facing the database research community: how to solve the system heterogeneity; how to build the global model; how to solve the semantic heterogeneity; and how to deal with queries automatically, etc [8].

Contact us at nmalotaibi@kau.edu.sa , anoaman@kau.edu.sa

SOA is a novel architecture aimed to build collaborative computing systems. SOA is essentially a distributed architecture, with systems that span computing platforms, data sources, and technologies [8, 5]. SOA provides a standard method to integrate both data sources and software applications by regarding them as interoperable services. Thus, client applications will combine these services to implement their intended tasks [4]. The implementation of SOA using Web services technologies is the current state of the art in systems integration. This research will elaborate on implementing integration of biological data by using SOA. Also the research will discuss the following questions: Does using SOA solve the above challenges that facing integration process? Does the biologist scientists can access the biological data in easier way than using other integration approach?

II. BACKGROUND

A. Data Integration

Convergent advances in biochemistry techniques, biotechnologies, and information technology and computer science provided the basis for the development of bioinformatics and made available huge and growing amounts of biological data. Nowadays public database infrastructure spans a very large collection of heterogeneous biological data, opening new opportunities for molecular biology, bio-medical and bioinformatics research, but raising also new problems for their integration and computational processing. Indeed the integration of multiple data types is one of the main topics in bioinformatics and functional genomics [4]. The process of heterogeneous database integration may be defined as "the creation of a single, uniform query interface to data that are collected and stored in multiple, heterogeneous databases." [7]. The example of databases that hold the biological data: Swiss rot and PIR focus on protein sequences, while Protein Data Bank (PDB) stores protein structures, Embank stores DNA sequences, BIND specialize in protein–protein interactions [4]. The main goal of integration is to provide mechanisms that can unify a number of (computer) systems. Three important aspects of system integration are distribution, autonomy and heterogeneity.

• *Distribution:* Often the source of database is distributed. The user need not know the location and other details of each available resource. Such details are usually handled automatically by the integrated system [12].

• *Autonomy:* It is very often the case that integrated resources belong to different organizations or research groups. Each data sources are working autonomous without any control by another integrated system [12].

• *Heterogeneity:* In an open and diverse environment it is very common that some or all of the data sources are different from each other. The differences are either semantic or technical. Technical heterogeneity difference occurs because of different hardware platforms, operating systems, database management systems (query languages, data models) and programming languages. The semantic heterogeneity is conceptual differences that occur in the data models/schemas of the data sources i.e., the organization of data and the relationships between such data. For examples there are synonyms when attributes of two schemas have different names but refer to the same concept [12].

B. Service Oriented Architecture

The SOA was proposed initially as an emerging paradigm for business process integration inside or across organization boundaries [5]. SOA is an approach to defining integration architectures based on the concept of a service. A basic tenet of SOA is that the use of explicit service interfaces and interoperable, location-transparent communication protocols means that services are loosely coupled with each other. Services are software modules that are accessed by name via an interface, typically in a request-reply mode. Services can be invoked independently by either external or internal service requesters to process simple functions. Service consumers are software that embeds a service interface proxy (the client representation of the interface) [13].

III. Literature Review

As Internet accessible biomedical databases proliferate there is an increased need for tools capable of integrating information available from a variety of sources. Clinicians and researchers could benefit from a more consolidated and unified view of the available biomedical data. Systems biology researchers need to integrate disparate genetic information from multiple public sources to merge with their own experimental data [9]. A wide variety of technologies, techniques and systems have been explored and exploited over the past 15 years [10].

In following subsection we review some approaches (mediator and data warehousing) used for integrated biological data. Then the proposed approach (SOA) will be introduced to solve the problem of integration.

A. Mediator approach

Mediator-based integration concentrates on query translation. A mediator in the information integration context is a system that is responsible for reformulating at runtime a query given by a user on a single mediated schema into a query on the local schema of the underlying data sources. Typically, each individual source will also require the definition of a “wrapper” component, which will be used to export a view of the local data in a useful format for mediation [4, 6]. This approach required mapping to capture the relationship between the source descriptions and the mediator and thus allow queries on the mediator to be translated to queries on the data sources. Specifying this correspondence is a crucial step in creating a mediator, as it will influence both how difficult the query reformulation is and how easily new sources can be added to or removed from the integration system.

The two main approaches for establishing the mapping between each source schema and the global schema are global-as-view (GAV) and local-as-view (LAV). In the GAV approach the mediator relation nothing but a query over the data sources. The GAV approach greatly facilitates query reformulation, however handling the addition or removal of a source in a GAV mediator is much more difficult as it requires a modification of the mediator schema to take into account the changes. In a LAV-based mediator every source relation is defined over the relations and the schema of the mediator. It is therefore up to the individual sources to provide a description of their schema in terms of the global schema, making it very simple to add or remove sources but also complicating the query reformulation and processing role of the mediator [6].

This approaches is satisfied the integration of biological data. The scientist can access the different sources by sending the query and receiving result. Also, each source preserve own data autonomies. Also, this approaches is flexible, it allow adding or removing any sources. But there are problems in this approach the mapping between local and global schemas need to manually specify. Also this approach have complex schema to satisfied mapping between local and global view. Another drawback of this approach, sources must be available during query executions [6].

B. Data warehouse

Data warehouse is bringing all data from multiple sources into a local warehouse and executing all queries on the data contained in the warehouse rather than in the actual sources. The first step in data warehousing is to develop a unified data model that can accommodate all the information that is contained in the various source databases. The next step is to develop a series of software programs that will fetch the data from the source databases, transform them to match the unified data model and then load them into the warehouse. The warehouse can answering any question that can handle by source database also have integrated knowledge not in individual source database. Systems that rely on the data warehouse architecture are usually restricted to consider a few source databases, but can achieve a higher degree of integration of the data sources [1, 6]. The data warehouse achieves the high degree of integration of biological data sources. Also it reduces the response time to answer the queries because requests send to single place. But the major drawbacks of data warehouse are update issue. The all data insert and update in source data base must be re-imported in warehouse. Also, another drawback is difficult to add new source without change the warehouse schema [6].

C. Service Oriented Architecture

In this section we review the important work that used the SOA approach for integration different data sources. These works are Service-Oriented Data Integration Architecture and Dynamic Data Integration using web services.

1. Service-Oriented Data Integration Architecture

The emerging of SOA and the application of ontology provide a new starting point for solving the problem of Data integration. By making use of services and ontology technology, the system heterogeneity is solved by services; the global model is constituted by two levels of ontologism to solve the semantic heterogeneity, to provide the semantic reasoning, to facilitate the query, and to make automatic query answering. A service is provided for every database to solve the system heterogeneity just like providing the hardware driver before a hardware is plugged into the main board. Every component is a service; the whole system is a service also. A data source can be included by different SOA systems just like one SOA system can contain different sources, thus new systems are easy to be developed and the underlying resources can be utilized fully.

As SOA it takes several advantages: concise structure, adaptive to dynamic changes, adaptable to additional functions, scalable and easy to be integrated into other systems [8].

Authors discussed the problem of integration in general by using SOA approach. This method achieves the higher degree of integration and solves the problem in previous approach unless the sources of data must be available during the query execution.

2. Dynamic Data Integration using web services

The authors was address the problem of large scale data integration, where the data sources are unknown at design time, data are from autonomous organizations, and may evolve. Experiments are described involving a demonstrator system in the field of health services data integration within the UK. Current web services technology has been used extensively and largely successfully in these distributed prototype systems. The work shows that web services provide a good infrastructure layer, but integration demands a higher level "broker" architectural layer.

A service-oriented data integration architecture (SODIA) has been proposed to provide a dynamically unified view of data on demand from various autonomous, heterogeneous and distributed data sources. Service providers publish their data sources as data access services (a kind of Web service, which allows for the provision of a data intensive service), which may then be discovered, bound at the time they are needed and disengaged after use. Hence the changes such as organization structures, backend data sources, data structures and semantics could be managed dynamically and potentially reduce the maintenance cost. To achieve this, those data access services should be semantically described and discovered [11].

The authors discussed dynamic data integration using web services. This approach solves the problem of integration in domain of health care. The major benefit of this approach is to get integrated view of data from autonomous, heterogeneous, distributed and continually changing data sources.

IV. Methodology

In this study SOA approach is propose to solve problem of integration biological data. The goal of system enables the scientists to acquire some knowledge from large amounts of data rather than manually access these sources. Proposed system is web based application because most of biological data sources are available online. It has been built using web services on Microsoft .Net environment.

The web service is use to provide a uniform regime for "plumbing together" data resources that present themselves as services with programmatic interfaces. Web Services (WS) technology is the most popular implementation of SOA, which provides flexibility through decoupling the service provider from the service consumer. Web service define standardized interface (Web Services Description Language, or WSDL), a Web services define a standardized communication protocol (Simple Object Access Protocol, or SOAP), a standardized repository for registering and discovering Web services (Universal Description, Discovery, and Integration, known as UDDI), and standardized message encoding using XML.

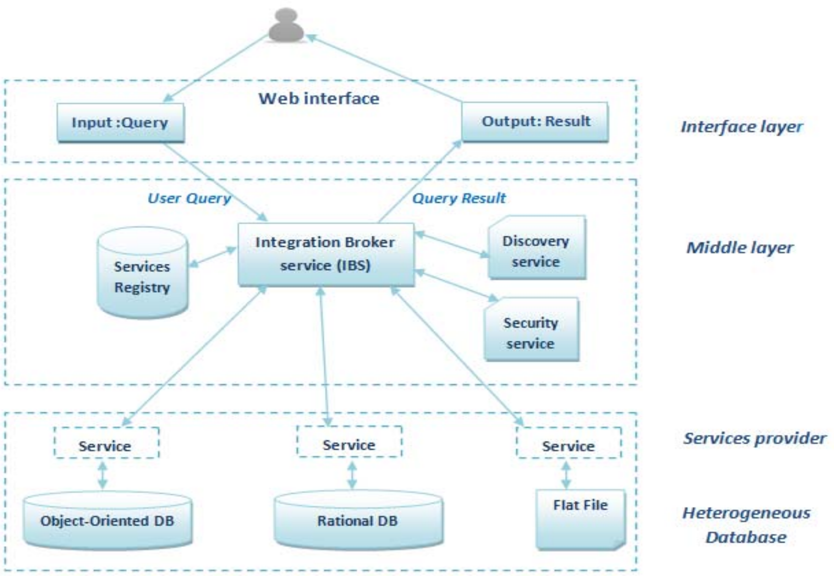

Fig.1 Proposed architecture for integration biological data with SOA

These standards enable a Web service to reside anywhere and be accessed from everywhere, making Web services well suited to the role of providing aggregated functionality within a composite application and being the standard for inter application interfaces [5].

In SOA based environment, database sources are wrapper into web services, which provide common interface for resources. Web services allow users to access remote resources in the same way in spite of the difference in internal details. Hence, any change in data source, data structures and semantics could be managed dynamically and potentially reduce the maintenance cost.

There are a lot of technologies for implemented SOA such as EJB or EMS. The problem of these technologies, none of them are able to implement self-describing entities as is possible using Web Services. By using the Web Service Definition Language WSDL it is possible that a client can be completely "agnostic" about a service, gaining "understanding" during run-time of all the semantics exposed in a service without "knowing" "a priori" protocols, binding, data types, policies, Service Level Agreement SLA [14].

A. System Architecture

The proposed architecture consists of three layers. The bottom layer of architecture consists of different database sources and set of services provider. These sources are different in hardware platforms, operating systems, database management systems, access protocols, transport formats and programming languages. The services providers implement services and describe them using service description languages such as WSDL. Service providers publish the service description file into a service registry, such as UDDI. This layer acts as interfaces of accessing local databases. The middle layer contains the Integration Broker Service (IBS) that is a composed service, which integrates different data access services (service provider) and functional services such as the Discovery Service (DS) in order to provide the end-user with an integrated uniform view of the data. The discovery service is used to discover and bind to service implementations at run time. The top layer is interface layer. This layer represents service requester. Client send query through this layer and receive result also through this layer. Fig.1 represents the architecture of proposed system [8, 11, 12].

B. Implementation

The most popular implementation of SOA is web services. Web services have several advantages mentioned in the beginning of methodology section. Two of the most popular technologies for implementing SOA through web services are Microsoft's .NET and Sun Microsystems's Java Platform, Enterprise Edition technologies. In our proposed architecture we use Microsoft's .NET to implement SOA approach to integrate biological data. .NET is a proprietary solution for code running and for development, designed for use with windows operating systems and servers. .NET platform can be used to provide a variety of applications, from desktop and mobile systems to distributed web solutions and web services. An important .NET component for SOA is ASP.NET environment, used to provide the level of web technology in SOA (and subsequent additions and the Web Services Enhancements (WSE) extension. Building services-oriented solutions with .NET usually involve the creation of service providers as ASP.NET Web Services and client services using auto-generated proxy classes. Service developer code is transformed in Common Language Runtime and then executed. [15, 16].

A. Scenario of proposed system

This section discussed the scenario of the proposed system. This scenario begins from sending query until the result return. The scenario as follow:

• User send Query about specific biological data (such that DNA, Protein and Gene) through web interface.

• Query is processed by Integration Broker Service (IBS). IBS contact with service discovery that used to discover and bind to service implementations at run time.

• The IBS is distributed Query to services that met user requirements (services that return by service discovery) and use this service to access appropriate database sources.

• The result of query is return to IBS that forward result to requestor.

The above scenario was represent negotiation between users and system Based on SOA architecture.

B. Procedure

In this section we discuss the activities necessary for accomplish our study. The activities as follow:

• *Data collection and analysis:* Because biological data are spans across different online resources, we need domain expert and integrated expert to collect and capture resource representation and descriptions.

• *Implementation:* Include construct a service for each data sources in our system, design interface of system and building web application.

• *Testing:* will do in gene center and require from researcher to send query for specific gene. Then measure response time for result and average time necessary to access and send query. Also, another measurement is measure response of system to change in content of databases.

V. BENEFIT OF PROPOSED SYSTEM

This section discusses the expected benefits from proposed system. Those benefits as follow:

• The system provides the unified access for different biological data sources.

• SOAs can provide dynamic service discovery and binding, which means that service integration can occur on demand.

• SOA provides a standard method to integrate both data sources and software applications by regarding them as interoperable services.

• The loose coupled feature of SOA facilitates the distribution of computational intensive processes across multiple nodes.

• Provide the security feature by security service that is used to authenticate the user and control the data items a given user can access.

VI. CONCLUSION

This paper proposed the architecture of how to use SOA architecture to integrate different biological data sources. Service Oriented Integration is a more application-agnostic approach for integration. SOA promotes assembly, orchestration, and choreography based on service, made possible by a Service-based integration.

REFERENCES

[1] L. D. Stein., INTEGRATING BIOLOGICAL DATABASES.: Nature Publishing Group, May 2003, Vol. 4.
[2] B. Smith, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration: Nature Publishing Group, November 2007, nature Biotechnology, Vol. 25. 11.
[3] Zoé Lacroix, Biological Data Integration: Wrapping Data and Tools. JUNE 2002, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, Vol. 6. 2.
[4] M. Mesiti, et al, XML-based approaches for the integration of heterogeneous bio-molecular data. Varenna, Italy: BioMed Central Ltd, October 15, 2009, BMC Bioinformatics, Vol. 10.
[5] Dr. G.K. Behara, Service Integration. October 2008.
[6] T. Hernandez, and S. Kambhampat, , Integration of Biological Sources: Current Systems and Challenges Ahead. September 2004.
[7] W. Sujansky, Heterogeneous Database Integration in Biomedicine. : Elsevier Science (USA), January 22, 2002, Journal of Biomedical Informatics, Vol. 34.
[8] G. Hao, S. Ma, J. Lv, Y. Sui,.A Service-Oriented Data Integration Architecture and the Integrating Tree: IEEE, 2006. Proceedings of the Fifth International Conference on Grid and Cooperative Computing.
[9] P.Mork, A. Halevy, and P.Tarczy-Homoch , A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases. Mork. Washington: AMIA, Inc., 2001
[10] C. Goble, and R. Stevens, State of the nation in data integration for bioinformatics. Manchester: Elsevier Inc, February 5, 2008, Journal of Biomedical Informatics.
[11] Fujun Zhu Turner, et al, Dynamic Data Integration using Web Services. Durham Univ, UK: IEEE, 2004. Proceedings. IEEE International Conference on web service. 0-7695-2167-3
[12] K.A. Karasavvas, R. Baldock,and A. Burger. Bioinformatics integration and agent technology. [ed.] Elsevier Inc. s.l. : Science direct, May 8, 2004, Journal of Biomedical Informatics.
[13] Keen, Martin, et al, Patterns: Implementing an SOA Using an Enterprise Service Bus. : IBM Corp., July 2004.
[14] Raymond Kurland, Service-oriented Architecture (SOA) defined: TechniCom Group., October 2007.
[15] Ilie Tamaş, Radu Bucea-Manea, Rocsana Ţoniş, Applications and Service Integration of Decision Support Systems using .NET Platform, 2008
[16] B.V. Kumar, Prakash Narayan, Tony Ng, Implementation SOA using Java EE: Evolution of Service Oriented Architecture.: 1st ed, Prentice Hall, Dec 23, 2009

Ms.Noura Al-Otaibi: was born in Saudi Arabia in October 1985.currently study in master program in King AbdulAziz University,Saudi Arabia. In 2008 have a B. SC. in computer science with first honor degree from Faculty of Computing and Technology, King AbdulAziz University, Jeddah, Saudi Arabia. Miss.Al-Otaibi currently is a Teacher Assistant at King AbdulAziz University, College of Science. She interested on Bioinformatics, E-Systems and Brain computer Interface. e-mail: nmalotaibi@kau.edu.sa

Dr.Amin Noaman: received his Ph.D. in computer science from University of Manitoba, Canada, in 1999 in the area of distributing data warehousing. He received his M. Sc. in computer science from McGill University, Canada, in 1995 in the area of software engineering. He received his B. SC. in computer science from King Abdulaziz University, Saudi Arabia, in 1988. Dr. Norman is currently an assistant professor in the computer science department faculty of computing and information technology at the King Abdulaziz University, where his current research focuses on data warehousing, bioinformatics, distributed database systems, decision support systems and mobile database.
anoaman@kau.edu.sa