BIODIVERSITY INFORMATION RETRIEVAL ACROSS NETWORKED DATA SETS

Sarinder Kaur Kashmir Singh*, Kaharuddin Dimyati, Susan Lim Lee Hong, Amir Feisal Merican

Institute of Biological Sciences, Faculty of Science, University of Malaya
**Department of Electrical Engineering, Faculty of Engineering, University of Malaya*
sarinder@um.edu.my, kahar@um.edu.my, susan@um.edu.my, merican@yahoo.com

ABSTRACT

Globally, biodiversity resources are inevitable digital and stored in wide variety of formats by researchers or stakeholders. In the Malaysian perspective, although awareness of digitizing the biodiversity data has long been stressed, the semantic interoperability of the biodiversity collections is still an issue to be looked into. This is essentially because when data is shared, the copyright crisis occurs hence creating a setback among researchers wanting to promote or share their findings through online presentations. Hence, this has become a hindrance for researchers in this country to share their valuable information and knowledge in this area with their peers locally or even internationally. To solve this, we present an approach to integrate data through wrapping of various datasets stored in relational databases located on networked platforms. The approach, which uses tools such as XML, PHP, ASP and HTML to integrate databases in heterogeneous environment, does not only solve copyright problems by suggesting distributed warehouses and required fields for sharing but also give the data owner the benefit of having their database under their own jurisdiction. The approach presented in this paper is important for scientists as findings in science are useful should be shared among the scientists for a better living.

Key words: Information Retrieval, Biodiversity Databases, Database Integration

1. Introduction

In the past 20 years, scientists have engaged themselves in the informatics discipline of querying multiple remote or local heterogeneous data sources, integrating manually received data and manipulating it with advanced data analyzing and visualizing tools. The access of relevant data, combining data sources and coping with their distribution and heterogeneity is a tremendously difficult task (Lacroix, 2002). However in recent time, information retrieval has become faster through networking related datasets. In this paper, biodiversity datasets are confined to databases containing taxonomic information of flora as well as fauna. Searching integrated multiple biodiversity databases at once have, in many occasions, significantly alleviated the process of information retrieval in the domain. The underlying mechanism which integrates different data sources has taken the load off the researchers in information gathering and mining.

While most of the current systems, which serve the above purpose, are either too specialized or complicated, there is a critical need to adopt an alternative approach which is simple yet dynamic enough for the scientific community. The introduction of the new approach is based on some studies on the current trends on information retrieval from networked datasets.

2. Literature Review

A literature study was done on existing systems for information retrieval in Biology. They are GenoMax (InforMax, 2001), Kleisli (Wong, 2000a), DiscoveryLink (Haas et al., 2001), SPICE (Jones *et al.*, 2000) and DiGIR (Biodiversity Research Center, 2005). The summary is presented in Table 1. Even though GenoMax, Kleisli and DiscoveryLink do not support biodiversity data, these systems were reviewed to evaluate their architecture

Table 1. Existing Biological Database Integration Systems

System Developer	Data type supported	Approach	Strengths	Weaknesses	References
GenoMax by InforMax	Genomic	Data warehouse	1. simple graphical user interface	1.scripting language is not designed for large-scale database style manipulations 2. difficulties in adding new kinds of data sources and analysis tools	Wong (2002) InforMax (2001)
Kleisli by geneticXchange Inc.	Genomic	Mediator based	1. high level query language 2. ability to store, update, and manage complex nested data and a good query optimizer besides being equipped with two application programming interfaces so that it can be accessed in a JDBC-like manner from Perl and Java 3. good and simple user interface	1.programming of queries is complicated	Wong (2000b)
DiscoveryLink IBM	Biomedical	Mediator based	1. high-level query language	1. only supports wrappers written in C++, which is	Wong (2002)

			2. perform further manipulations on the results	not the most suitable programming language for writing wrappers 2. not straightforward to add new data sources or analysis tools into the system 3. very limited in its capability for handling long documents and as a tool for creating and managing data warehouses for biology 4. programming of queries is complicated	
SPICE Centre for Plant Diversity & Systematics, Plant Science Laboratories, University of Reading, England	Biodiversity	Mediator based	1. objects are distributed optimally 2. implementation language independence and platform independence ensure that SPICE can interoperate effectively across all databases of interest	1. interface of SPICE allows for searching through scientific names and common names only 2. imposes a set of requirements on the kinds of data model that the individual databases can have	Jones (2000)
DiGIR Biodiversity Research Center (BRC) Informatics in collaboration with the Museum of Vertebrate Zoology at UC Berkeley and the California Academy of Sciences	Biodiversity	Mediator based	1. friendly and guided interface	1. imposes a set of requirements on the kinds of data model that the individual databases can have 2. does not retrieve images from a database 3. cannot integrate FileMaker Database Management Systems	Biodiversity Research Center (2005)
BioCASe	Any Type	Mediator based			

After studying the literature, it can be said that GenoMax, Kleisli and DiscoveryLink are far too complicated. Not only the programming of queries is difficult but adding new data sources are also intricate. While SPICE and DIGIR are built specially for biodiversity data, these systems impose a set of requirements on the kinds of data model that the individual databases must have. For instance, DiGIR requires five fields such as *Date Last Modified, Institution Code, Collection Code, Catalogue Number and Scientific Name*. Though

GenoMax, Kleisli and DiGIR have simple and friendly user interface, their underlying architecture is not easily understood.

3. A Possible Solution

The literature done is this paper suggested that a simple and dynamic system is needed to serve the purpose of querying multiple databases using a single search engine. This paper focuses on a new solution for retrieving information from biodiversity datasets. The conceptual architecture is presented in Figure 1 and the search mechanism is described in the following subsection.

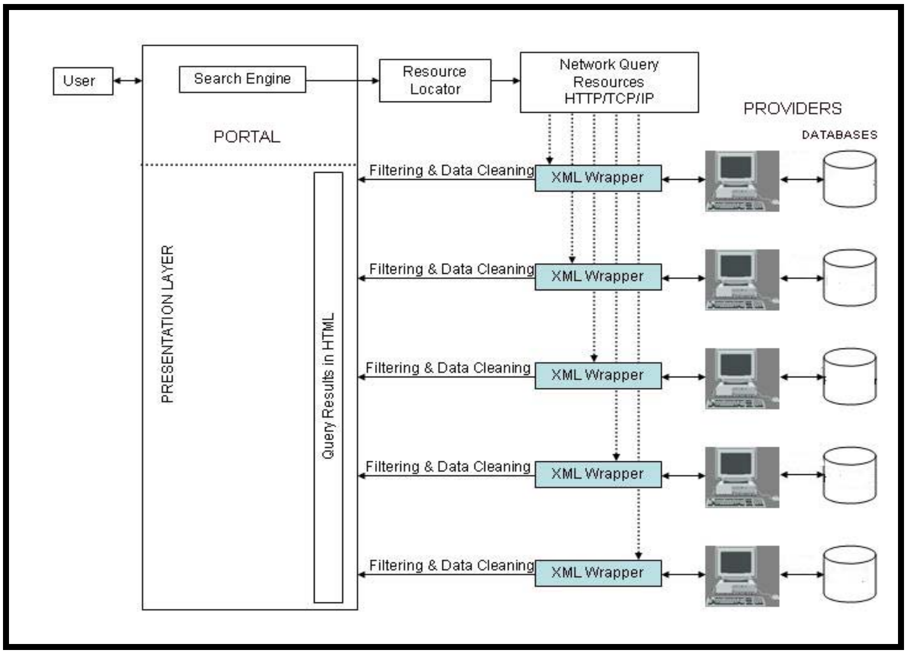

Figure 1. Architecture for CABIR

3.1 Search Mechanism

The presentation layer basically communicates with the user. Therefore, it was designed vigilantly to meet the users' expectations of a friendly interface. It has two main components which are the search engine and query results. Users key in the query via the search engine using simple terms. The search engine thus uses a smart query agent to manipulate the query to make the search more extensive and robust. This generates a larger set of results which will then be filtered based on users view. The presentation layer basically sends query and receives results. Based on the users' selection, the resource locator will then look up for the resources which are web biological data sources connected using the proposed solution. At this point users can select more than one database and the system will execute the processing simultaneously to these repositories. Once the resource locator has identified the web data sources, query will be sent via the TCP/IP protocol using the uniform resource locator (URL) to retrieve all the necessary data. XML wrappers contains provider information document, xml schema and xml documents containing query results. The provider information document consists of the database connectivity details, namespace and query statements. This document will be installed at the client side where the database resides. The xml schema will map the query results into a well formed data structure which applies the Darwin Core V2 global standard (Biodiversity Research Center, 2005). Once the results are produced in the XML document, the data goes through a clean-up phase to meet the users requirements. At this point, surplus data and empty fields are filtered out. Thus, data is sent to the presentation layer ,which is then converted into HTML for viewing. The conversion is done using XSLT (Extensible Stylesheet Language Transformations). XSLT is a transformation language for converting XML instances.

3.2 Prototype

The process of building a database integration system initially necessitates integration of databases with the Web. This process required a Web server, application program and connectivity. In this research, Apache and Internet Information Service (IIS) were employed as Web servers for the databases to be Web accessible. ASP and PHP were chosen as scripting languages to integrate database with the Web. ASP was used for Windows-based providers whereas PHP for UNIX-based providers. As for the database connectivity, ODBC, DataDirect32-BIT SequeLink 5.4 and Oledb were the driver managers used (see Figure 2). Two methods were used for database connection in CABIR. They are DSN (Data Source Name) connection and DSN-less connection, depending on type of DBMS used. Structured Query Language (SQL) was used to retrieve and manipulate the data from the relational databases. It was also applied to relate the tables using the *JOIN* command.

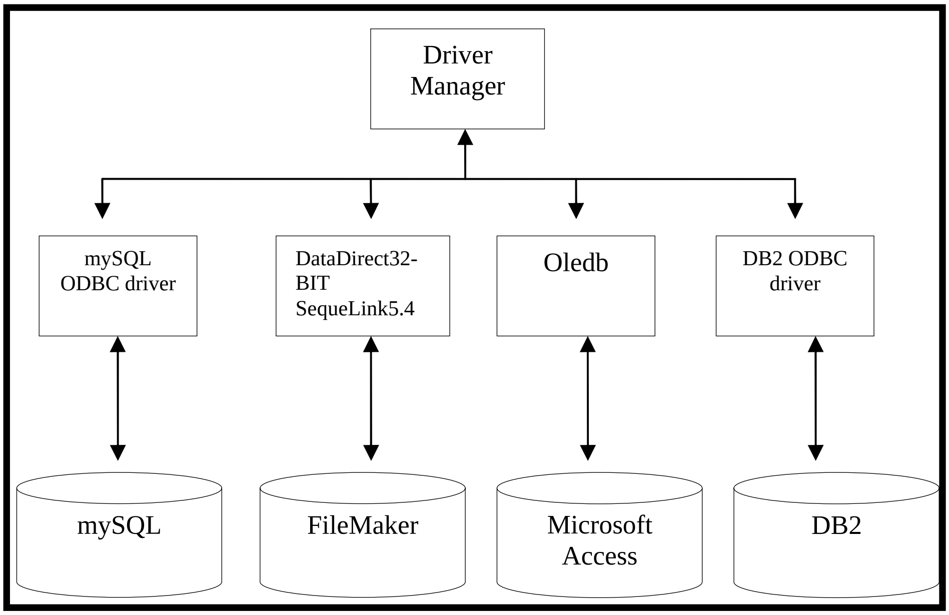

Figure 2. Database connections in proposed solution

Each database in the proposed system required a provider which connects the database to the Web service. The provider also contained the SQL strings to perform the desired query to the database according to the request from the wrappers. The data extracted from the database was returned as XML schema and XML document. Figure 3 shows the architectural view of the database integration process for one database. The application sends a query to the Web server, through the Internet. The query is then forwarded to the specific provider. The provider returns results in XML format which is then converted into readable HTML format. Figure 3 illustrated the process for a single database. The Web server contains a connection with the back-end database. The document that contains the ASP/PHP script and database connection is called a Provider. In this solution, heterogeneous databases are used and they are queried simultaneously.

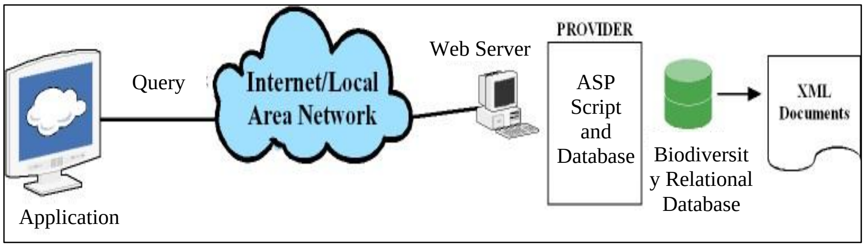

Figure 3. Search process for a single database.

4. Discussion and Conclusion

The proposed solution was tested on biodiversity databases and it is expected to also work for data in other biological domains and for a variety of different data sets outside the biological domain. This is the components or modules in the proposed solution are independent of each other, especially the data format.

Table 2. The proposed solution criteria against the standard database integration systems

	Aim of Integration	**Data Model**	**Source Model**	**User Model**	**Level of Transparency**	**Overall Integration Approach**
GenoMax	Data Mining	Structured static data	Mostly complementary	Expertise in software functionality, data mining tools, life science informatics analysis approaches, collaboration twork, and other aspects of the user interface	Sources specified by head database	Warehouse based
Kleisli	Query-Oriented	Semi-structured, object-oriented	Mostly complementary	Expertise in query language	Sources specified by user	Mediator-based
DiscoveryLink	Query-oriented middleware	Structured, object-relational	Mostly complementary, some overlap	Expertise in query language	Sources selected by system	Mediator-based
SPICE	Query-oriented middleware	Structured, object-relational	Mostly complementary	Expertise in query language	Sources specified by user	Mediator-based
DiGIR	Query-Oriented	Structured, object-relational	Mostly complementary	Expertise in query language	Sources specified by user	Mediator-based
Proposed Solution	Query-Oriented	Structured, object-relational	Mostly Complimentary Can also accommodate other kind of data with a change of data model	Novice	Sources specified by user	Mediator-based

Besides that, the proposed solution also has generic characteristics of existing database integration systems (see Table 2). These systems were used as models to build the proposed solution, especially DiGIR which was implemented during the preliminary study (Sarinder *et. al*, 2007). While having the generic characteristics, the proposed solution is made simple with powerful underlying facilities. Thus, it is suitable for scientific community.

The proposed solution is a promising tool that will undoubtedly impact positively on the scientific community, especially the novice users. It is a simple tool that meets the requirements of querying heterogeneous and remote biodiversity databases. Currently, this solution is being adopted at University of Malaya for querying various biodiversity databases. Being a web based technology, it is not restricted to just local repositories. It has an open data format which allows addition of new data sources.

5. Conclusion and Future Work

Despite the strong points that are mentioned in this paper, the proposed solution can continue to develop with new features and updates. The following attributes are suggested for future research; (i) testing the solution with other biological and non-biological data, (ii) wizards to add providers, (iii) link with other similar systems and (iv) more search fields to search on. With these, the proposed solution is hoped to attract wider spectrum of users.

6. Acknowledgements

We would like to thank Prof Haji Mohamed and Prof Halijah Ibrahim from University Malaya and Dr Chua Tock Hing from Monash University for providing us with data used in this study. We also would like to thank Prof Mashkuri Yaacob, Dr Lai Weng Kin and Mr Jingyuan Luke of MIMOS Sdn Bhd, Malaysia for allowing us to use their facilities to test the system.

7. References

Biodiversity Research Center (2005). *Distributed Generic Information Retrieval* (online). Natural History Museum. Available from:
http://www.specifysoftware.org/Informatics/informaticsdigir/ (Accessed 10 January 2005).

Haas, L.M., Schwarz, P.M, Kodali, P., Kotlar, E., Rice, J.E. and Swope, W.C. (2001). DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Sys J.* **40.**489–511.

InforMax, 2001. *GenoMax* (online). United States. Available from:
http://www2.informaxinc.com/solutions/genomax/index.html (Accessed 15 December 2005).

Jones, A.C., Xu, X., Pittas, N., Gray, W.A., Fiddian, N.J., White, R.J., Robinson, J.S., Bisby, F.A. and Brandt, S.M. (2000). SPICE: A Flexible Architecture for Integrating Autonomous Databases to Comprise a Distributed Catalogue of Life, To appear in *Proc. 11th Int. Conference and Workshop on Database and Expert Systems Applications (DEXA 2000)*, Springer-Verlag (Lecture Notes in Computer Science).

Lacroix, Z. “Biological Data Integration”. IEEE Transactions on Information Technology in BioMedicine. 2002, VOL. 6

Merican, A.F, Othman, R.Y., Sarinder, K., Ismail, N., Kok, C.Y., Khoo, P.C., Yong, C.F., Moak, S.F.L. (2002). Development of Malaysian Indigenous Microbial Online Database System, *Asia Pacific J. Molecular Biology and BioTechnology* (in press).

Roderic, D.M. (2005). A Taxonomic search engine: Federating taxonomic databases using Web services. *BMC Bioinformatics.* **6**:48

Siegel, J. (1998). OMG overview: CORBA and the OMA in enterprise computing. *Communications of the ACM.* **41910**.37-43.

Wong, L. (2000a). Kleisli, a functional query system. *J. Funct. Prog.* **10**.19–56.

Wong, L. (2000b). Kleisli, its exchange format, supporting tools, and an application in protein interaction extraction. *In*: *Proceedings of .IEEE Intl. Symp. Bio-Informatics and Biomedical Engineering*. 21–28.

Wong, L. (2002). Technologies for integrating biological data. *Briefings in Bioinformatics,* 3, 389-404.