BIODIVERSITY DATABASE

AGAN ADUN SHASHI MOHAN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2013

BIODIVERSITY DATABASE

AGAN ADUN SHASHI MOHAN

RESEARCH REPORT SUBMITTED
IN FULFILLMENT OF THE REQUIREMENT
FOR THE DEGREE OF
MASTER OF BIOINFORMATICS

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2013

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: Agan Adun Shashi Mohan (No. K.P/Pasport: 810730-08-6665)

No. Pendaftaran/Matrik: SGJ100002

Nama Ijazah: Sarjana (Bioinformatik)

Tajuk: Pangkalan Data Biodiversiti

Bidang Penyelidikan: Bioinformatik

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerjatersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalamHasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain samaada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon Tarikh

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh

Nama: Dr Sarinder Kaur a/p Kashmir Singh
Jawatan: Pensyarah Kanan

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Agan Adun Shashi Mohan (I.C/Passport No):810730-08-6665

Registration/Matric No: SGJ100002

Name of Degree: MASTER OF BIOINFORMATICS

TITLE: Biodiversity Database

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any Copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name: Dr Sarinder Kaur a/p Kashmir Singh

Designation: Senior Lecturer

ABSTRACT

There is a growing interest in academia to provide biodiversity data to both the scientific community and the public. Biodiversity Informatics Laboratory Portal was developed to provide useful information to both scientists, who can easily navigate and find all parasite related resources in one place, and bioinformaticians, who are provided with interoperable resources containing data which can be mined and integrated. This thesis demonstrates how Web service technology provide the ability to create a dynamic integration of multiple heterogenous databases while maintaining the existing database schemas and not losing control over the administration of the databases.

ABSTRAK

Terdapat minat yang semakin meningkat dalam bidang akademik untuk menyediakan data biodiversiti kepada kedua-dua komuniti saintifik dan orang awam. Portal Makmal Informatik Biodiversiti telah dibangunkan untuk menyediakan maklumat yang berguna kepada saintis, yang dengan mudah boleh menavigasi dan mencari semua sumber parasit yang berkaitan di satu tempat, dan pakar bioinformatik, yang disediakan dengan sumber mengandungi data yang boleh dilombong dan bersepadu. Tesis ini menunjukkan bagaimana teknologi servis 'Web' menyediakan keupayaan untuk mewujudkan integrasi dinamik pelbagai bentuk pangkalan data di samping mengekalkan skema pangkalan data yang sedia ada dan tidak kehilangan kawalan ke atas pentadbiran pangkalan data.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my utmost gratitude to my respected supervisor, Dr. Sarinder Kaur a/p Kashmir Singh for her invaluable guidance throughout the duration of my project.

My heartfelt thanks to my parents for their encouragement and understanding during the ups and downs during the duration of the project.

Last but not least, I would like to express my gratitude for the help and encouragement of my friends.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
DiGIR	Distributed Generic Information Retrieval
DNA	Deoxyribonucleic Acid
ERD	Entity-Relationship Diagram
HTTP	Hypertext Transfer Protocol
JAX-RS	RESTful Web Services with Java
MRB	Mouse Resource Browser
NARIS	Natural History Information System
OS	Operating System
PDB	Protein Data Bank
REST	Representational State Transfer
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
UDDI	Universal Description, Discovery and Integration
W3C	World Wide Web Consortium
WS	Web Service
WDSL	Web Services Description Language
WYSIWYG	What-You-See-Is-What-You-Get
XML	Extensible Markup Language

CHAPTER 1

INTRODUCTION

Today, the Internet has allowed the accessibility of public biological databases by researchers. However, the heterogeneous biological databases give rise to many problems such as being time-consuming and difficult to an untrained user. Researchers may have to peruse and retrieve the data of interest from the query results manually before subjecting it through a second query in a different database. This becomes more complex and time-consuming when involving large amount of data, especially biological data. In this case, there arises a need for multiple database automated data integration.

Many different categories of biological databases exist that are separated according to different criteria. The original AceDB that focuses on *Caenorhabditiselegans*species is one example of database that focuses on a single organism, while other databases may focus on a single type of research data, such as radiation hybridisation of protein-protein interactions. There exist other databases that include databases of pathways, enzymes, transcription factors, individual organism's protein motifs besides the sequence databases, Medline and other bibliographic and PDB structural data databases, which are commonly accessed as important points of reference. (Salter, 1998)

Biological data that are able to be integrated with other related biological data types are often more of interest to researchers. For example, when a DNA sequence that is found to have a similar function when compared to other sequences, an inference can be made about

its functionality, especially when this information is used to develop a new drug for the treatment of a particular disorder. This shows the importance of this integrated biological data, especially for drug companies. Salteralso explained the importance of approaching the biological information through an integrated viewpoint. A researcher may want to know the associated bibliography, structural and functional information of a desired gene, and the development of search engines to will collate and cross-reference primary databases rapidly through derivative indexing (Salter, 1998).

The aim of this research is to develop a web portal that will provide multiple database integration and web services for users.

CHAPTER 2

LITERATURE REVIEW

2.1 Biodiversity Databases

‘‘Biodiversity’ is defined as the overall view of the ecosystem, which includes the interaction and diversity of many species within all the different ecosystems. Biodiversity is usually used to refer biological diversity at three levels such as (1) genetics, (2) species and (3) ecology. High throughput experiments such as whole genome sequencing, functional annotation, expression analysis and others have rapidly accelerated the development in bioinformatics, especially in the study of molecular events. However, unless this genetic information can be correlated with physiology, neurobiology,native habitat, or with genealogical relationships of the species, researchers will not benefit from the true value of these information. Inter-compatible molecular datasets would also contribute in the advancement in the field of biodiversity informatics.

Taxonomists create the nomenclature and classification databases are created by taxonomists, that can also contain all kinds if information about the characteristics, economic importance, conservation and management oforganisms. Technologiesare developed by the field of Bioinformatics for the management of genomic and proteomic data. Database technology,electronic storage media, WWW, and digitalization of data and creation of public databases have revolutionisethe way that biodiversity information is created, maintained, distributed and used. Theapplication of information technologies to

the management, analysis and interpretation of data regarding life particularly the species level organization are also included in the field of Biodiversity Informatics. (Shanmughavel, 2007).

2.2 Web Portal

A web portal is a web site that combines information from multiple sources and displays in a uniform way. Information from each of the sources is usually displayed in a dedicated area of the web portal, which is called a portlet. These portlets can be configured by users according to their needs and preference.

Web portals can offer many services besides its search engine, such as providing news feed, e-mail, databases, stock prices and entertainment. Portal also provides enterprises a method of providing consistent look and feel that is integrated with access control and procedures in running multiple applications and accessing multiple databases, which would have been different entities in the absence of these methods.

2.3 Web Services

World Wide Web Consortium (W3C) has defined a web service as software system that will provide support in the interoperability of the interaction for machine-to-machine over a network. Its interface is developed in a machine processable format, particularly in WDSL. “Other systems interact using SOAP-messages with the Web service, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related

standards" (Haas et al., 2004). Simple Object Access Protocol (SOAP) is a lightweight protocol functions as a decentralized, distributed environment for distributing information. This protocol is developed based in XML and consists of three parts, which are a framework defined by an envelope that describesthe content of the message and the method to process it, instances of application-defined data types expressed bya set of encoding rules, and remote procedure calls and responses represented by a convention. The combination of a variety of other protocols can potentially utilise SOAP, except in the case of the only bindings defined for the usage of SOAP with HTTP and HTTP Extension Framework in this document (Box et al., 2000).

Web services are available in two implementations, RESTful and SOAP, both of which are widely accepted in both academia and industry. Representational State Transfer (REST) services are very lightweight services that allow for simple implementation. These services are confined to only the standard HTTP operations of GET, PUT, POST and DELETE. REST services work by binding methods within the service to the HTTP operations. This simplicity is accomplished by standards like JAX-RS, which handles the binding of JAVA code to the standard HTTP operations. For example, to implement a Java class as a REST service, using JAX-RS, the developer need only annotate the source code to indicate which methods are called on a given operation. If a user accesses the service via the URL with a GET request, the method that was annotated for GET is invoked. RESTful services require only a Web server to be functional.

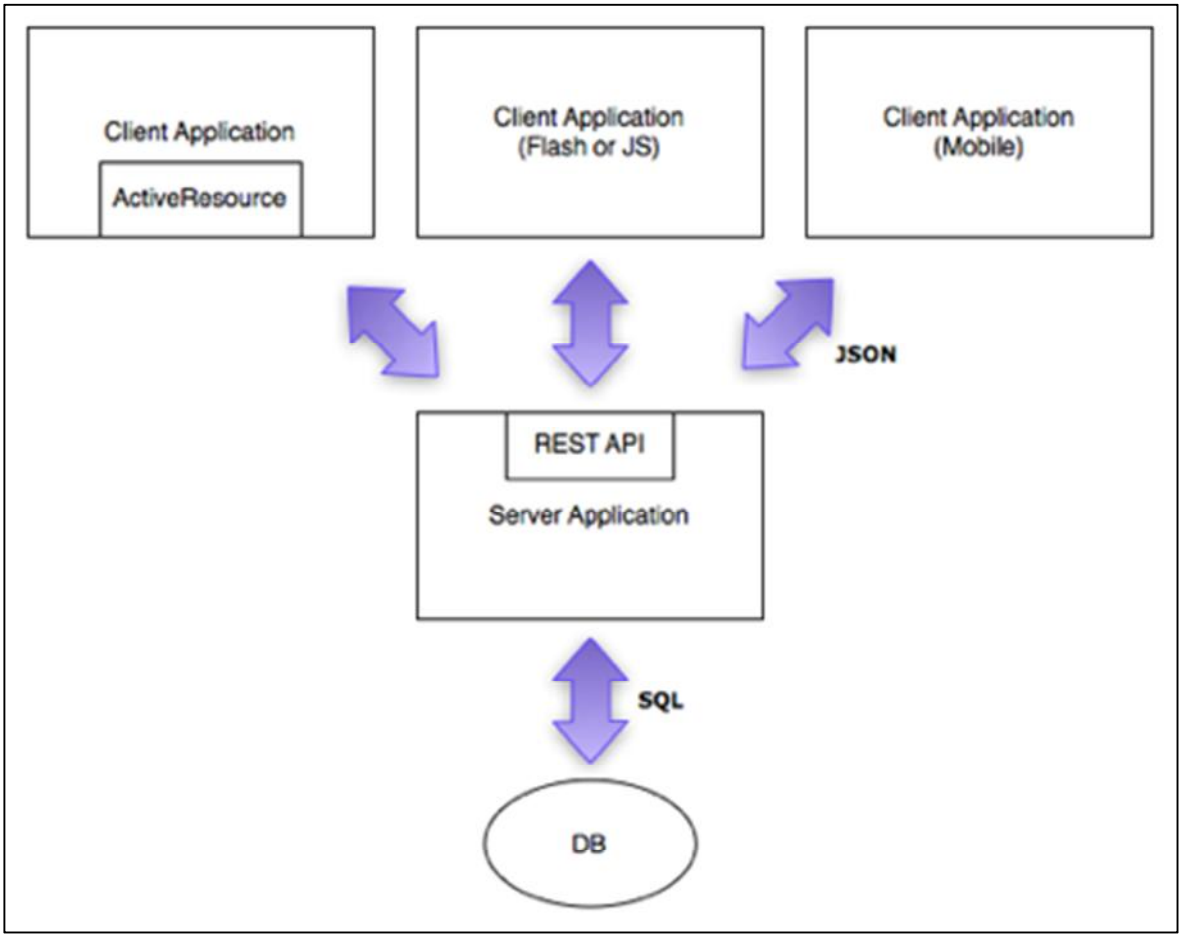


Figure 2.1 REST web services architecture

SOAP services are also known as “big services”. This architecture got this name from that fact that it requires more back-end infrastructure to support it. SOAP services utilize a stack of standard protocols, known as WS-* to describe the interface used to interact with the service, security, and constraints amongst other features.

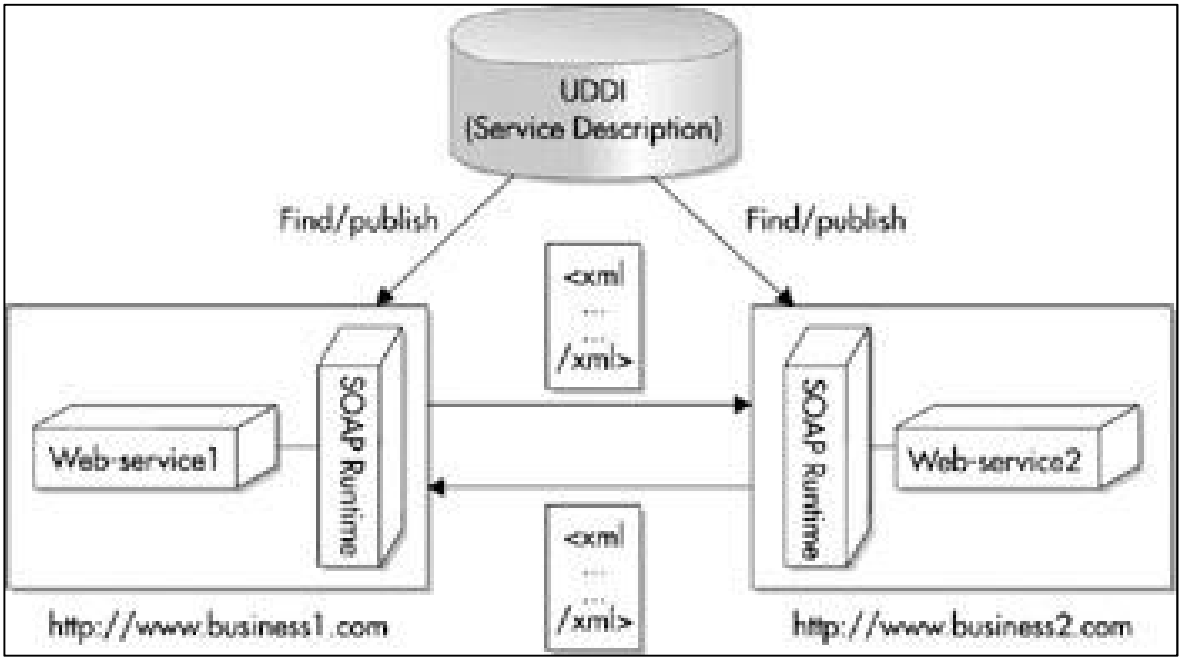


Figure 2.2 SOAP architecture

The Web Service Description Language (WSDL) is a well-defined XML specification that describes the service in sufficient detail that the reader of the file could invoke it (Curbera et al., 2002). WSDL is the cornerstone and the container or attachment point for most of the WS-* protocols. SOAP is the transport protocol for Web Services (Curbera et al., 2002). A SOAP message contains header information for the routing message and the payload, which is either the input or the output of the service as described in the WSDL file.

Universal Description, Discovery and Integration (UDDI) is a specification for a registry that allows for Web services to be published. From here, others can discover and download the WSDL for the published services. Information from the WSDL file is used to invoke the service. (Curbera*et al*., 2002).

2.4 Web Services in Biodiversity Informatics

The Korean Natural History Research Information System (NARIS) is a central depository for biodiversity data that collects and integrates the data from multiple institutes and natural museums in Korea. Additional biological data such as molecular level diversity and genome sequences are also collected from these integrated resources. Twelve institutes and museums in Korea have already been integrated within this depository, which utilises Distributed Generic Information Retrieval (DiGIR) as the protocol and Darwin Core2.0 as the metadata standard for data exchange. NARIS has also been implemented with functions such as statistical analysis and data quality control, particularly in the integration of biological data from the National Center for Biotechnology Information (NCBI) with NARIS, which includes molecular and genetic information. (Lim *et al*., 2007)

Mouse Resource Browser (MRB)is a database for the search and acquirement of mouse resource information from more than 200 mouse resources that are further divided into 33 different categories. MRB utilizes the CASIMIR DDF framework to hold the core and technical information for each of the resources (Zouberakis*et al*, 2010).

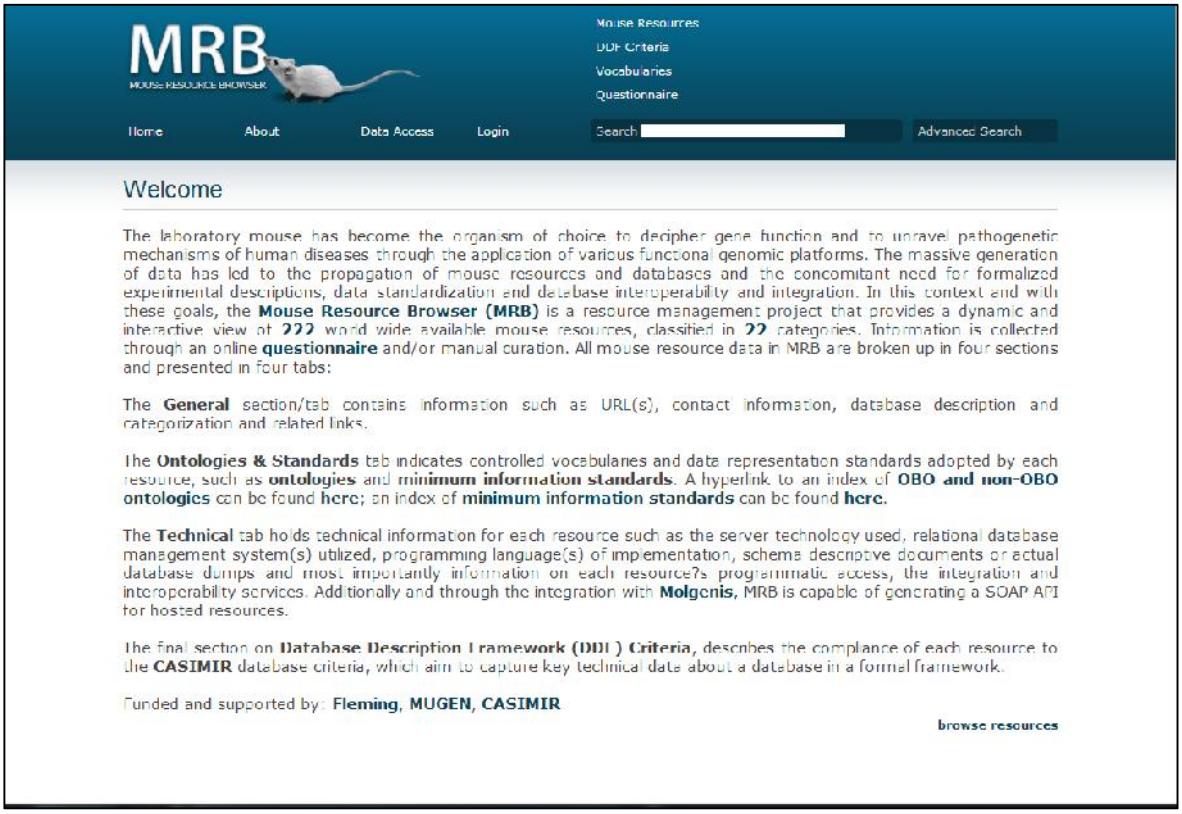


Figure 2.3 Mouse Resource Browser

CHAPTER 3

METHODOLOGY

3.1 System Development Process

The system was developed utilising the Waterfall model because the model allows the separation of tasks and management control. A schedule with particular deadlines can then be determined for each task to be achieved within each allocated time. This allows the development of the final product to be submitted to the user on time. Figure 3.2 shown below shows the steps involved in system development using the Waterfall model.

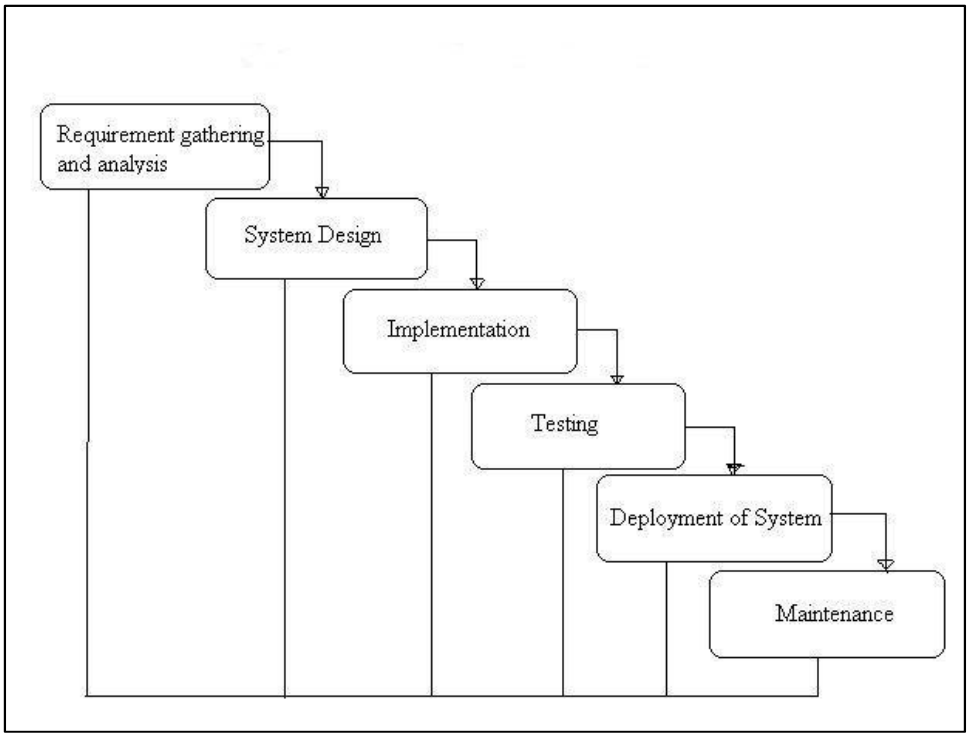


Figure 3.1 Waterfall model

3.1.1 Requirement Gathering and Analysis

All possible requirements the development of the system is gathered within this phase. Information such as project feasibility, project aims and needs are gathered through studies of previous works and user instructions, and the possibility of further incorporation of future requirements for the system are also studied. Finally, a Requirement Specification document was created to serve as a guideline for later phases of the model.

3.1.2 System Design

The Requirement Specification document from the previous phase was studied in this phase to plan the overall design of the system. The document also acts as a guideline for determining the hardware and software requirements for the project. The overall system architecture was also defined using information gathered from the previous phase. The system design phase provides specifications for the next phase of the model.

3.1.3 Implementation and Testing

The implementation step involves the development of the functionalities and application logics for the system. The coding in the development phase serves as the primary input for the design elements described in the design document. The optimization of the codes should also be conducted for the purpose of resource saving.

Source code generated during this phase is then subjected through a testing phase for the purpose of debugging. This phase can be manually or automatically conducted using available testing tools.

3.1.4 Deployment and Maintenance

The system developed should be provided with additional maintenance support to handle future bugs that may arise after the deployment of the system. This step can be more time-consuming than the duration taken during the system design and development phase. Any further modifications to handle errors that may occur due to unexpected input values into the system would be done within the maintenance phase.

3.2 Software Tools

The system software will facilitate the system administrator to handle the system resources efficiently, providing support to the installation and operation of the application software and providing support to the development and operation of additional software. The software tools involved in this project are displayed in Table 3.1.

Table 3.1 Software Tools

ITEM	SOFTWARE
Operating System	Genuine Windows® 7 Home Premium
Database Software	MySQL Server 5.5.27
Database Interface	PHPMyAdmin 3.5.2.2
Application Development Software	Adobe Dreamweaver CS 6
Application Server	PHP Version 5.4.7
Web Platform	Apache Version 2.4.3
Web Browser	Microsoft Internet Explorer, Mozilla Firefox, Google Chrome

3.2.1 Windows 7

Windows 7 is a series of Operating System (OS) developed by Microsoft for personal computers and mobile devices, such as tablets and smartphones. Windows 7 is the successor of Windows Vista, which includes support for building SOAP-based web services in native code using a new network API, application installation times shortening through new features, simplified development of installation packages, and a new Extended Linguistic Services API that improves the globalisation support.

3.2.2 MySQL Server

MySQL Server is a part of the MySQL package that runs as a server providing multi-user access to a number of databases. MySQL Server executes the instructions that are written by users. Instructions can be sent to MySQL Server in several ways, but PHP language is primary way used on most web sites. Users communicate with the relational database using SQL. PHP is used to create a connection to MySQL server and to send the SQL statements to the server so that the statements can be executed.

3.2.3 Adobe Dreamweaver CS 6

Adobe Dreamweaver Creative Suite 6 (CS6) is a What-You-See-Is-What-You-Get (WYSIWYG)web development application that is available for both Mac and Windows operating systems. Dreamweaver can hide the HyperText Markup Language (HTML) code details of pages from the user, making it possible for non-coders to create web pages and sites. Dreamweaver has a "live" view that enables the user to preview the web page in to allow for further troubleshooting or debugging. It has tools for site management, such as the replacement of lines of text or codes using site specified parameters, and a feature for creating multiple, similiarly structured web pages using templates that available in the software.

Dreamweaver supports the usage of third-party programs called "Extensions" that provides added functionality. Extensions can be written in HTML and Javascript by the web developer, or can be downloaded and installed from online sources. Extensions are made both commercially and freely available by a large extension developer community for Dreamweaver for most web development tasks.

3.2.4 Apache

The Apache HTTP Server, commonly referred as Apache, is an open source Web server that has played a key role in the initial growth of the World Wide Web. Its source code is freely available support the usage of third-party add-ons that extends the core functionality, which includes server-side programming support and authentication support. Many web applications are designed to support the environment and features provided by Apache. Apache is primarily used to support online static and dynamic Web pages.

3.2.5 Web Browser

A web browser is a software application that enables a user to display and interact with media and information usually located on a Web page of a website on the World Wide Web or on a local area network. These media and information includes text, images, videos, music, games and other formats. The web browser allows the user to access this information quickly and easily by utilising hyperlinks that are linked to the Web page containing the information. Web browsers format the HTML information to display the Web page appearance that may differ based on the browser being used.

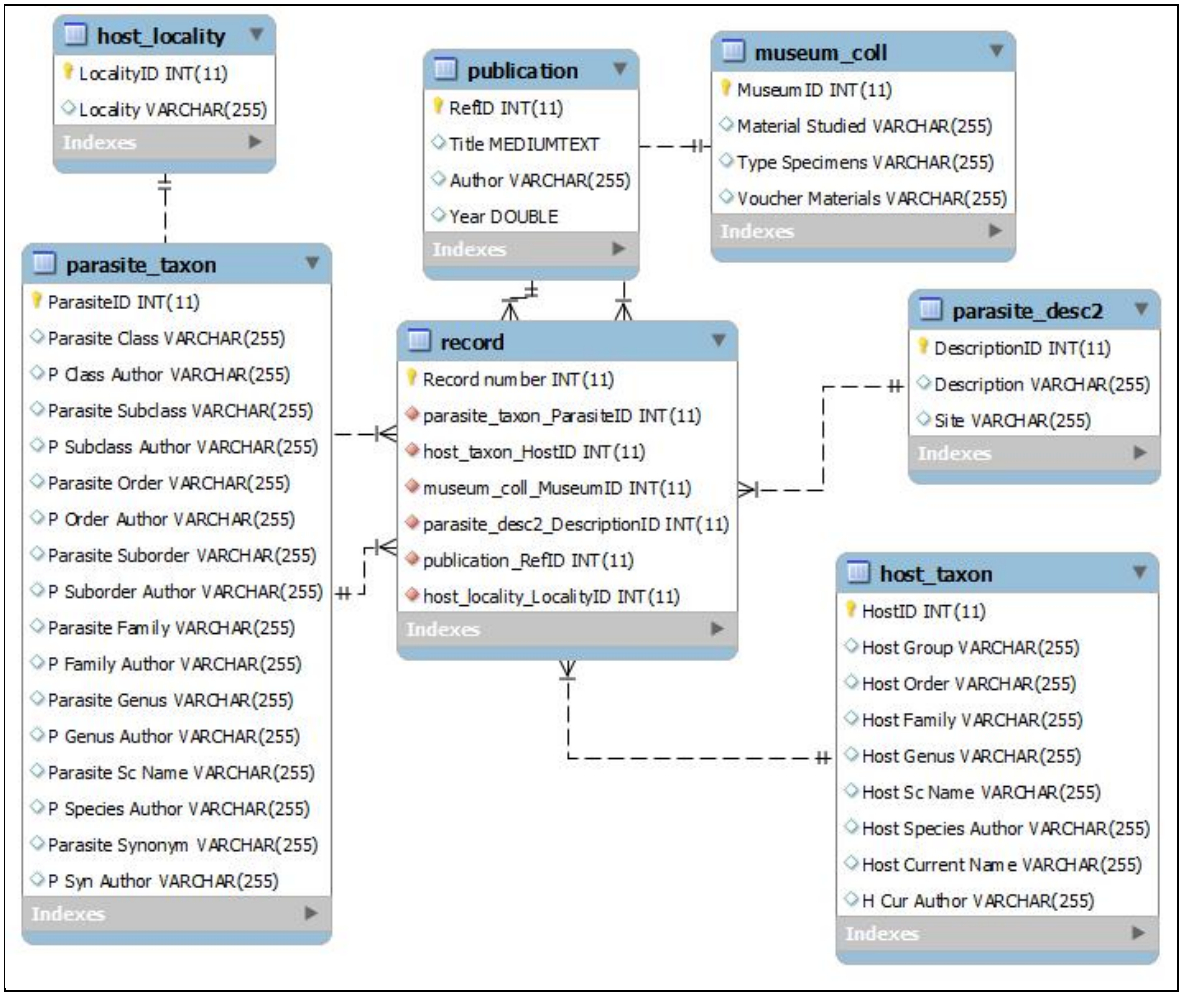
The most popular Web Browsers are Internet Explorer, Mozilla Firefox, Google Chrome, Safari and Opera, which are freely available for personal computers. All the web browsers can run on Mac and Windows OS. Web browsers are the most commonly used type of HyperText Transfer Protocol (HTTP) user agent. Browsers can also be used to access information provided by Web servers in private networks or content in file systems.

CHAPTER 4

RESULTS

4.1 Entity Relationship Diagram

Entity relationship diagram (ERD) for the system is the model that displays the entities of the system and the relationship lines between each entity. The entities represent the tables within the database, and the relationship lines represent all the keys that relate one entity to another entity. The lines indicate the relationship between each entity. The entities can have a one-to-many relationship such as between the parasite_desc2 and record entities or a one-to-one relationship such as between the host_locality and parasite_taxon entities. Figure 4.1 shows the ERD for one of the system's databases.



.Figure 4.1 Entity Relationship Diagram for sample Parasite-Host Database

4.2 Search Architecture

The system utilizes the federated search function which is an information technology which *'automatically searches across multiple distributed resources'* (Aguello, 2011). A query request will be distributed to the search engines participating in the federation and will return the result to the federated search engine. The federated search engine will then integrate the results from the participating search engines into a single list for display to the user.

This system is based on the utilization of Sphinx, which is a full-text search engine. Sphinx is integrated with the SQL databases, which is then indexed by the search engine. Search engine will then access the Sphinx search function using native search API.

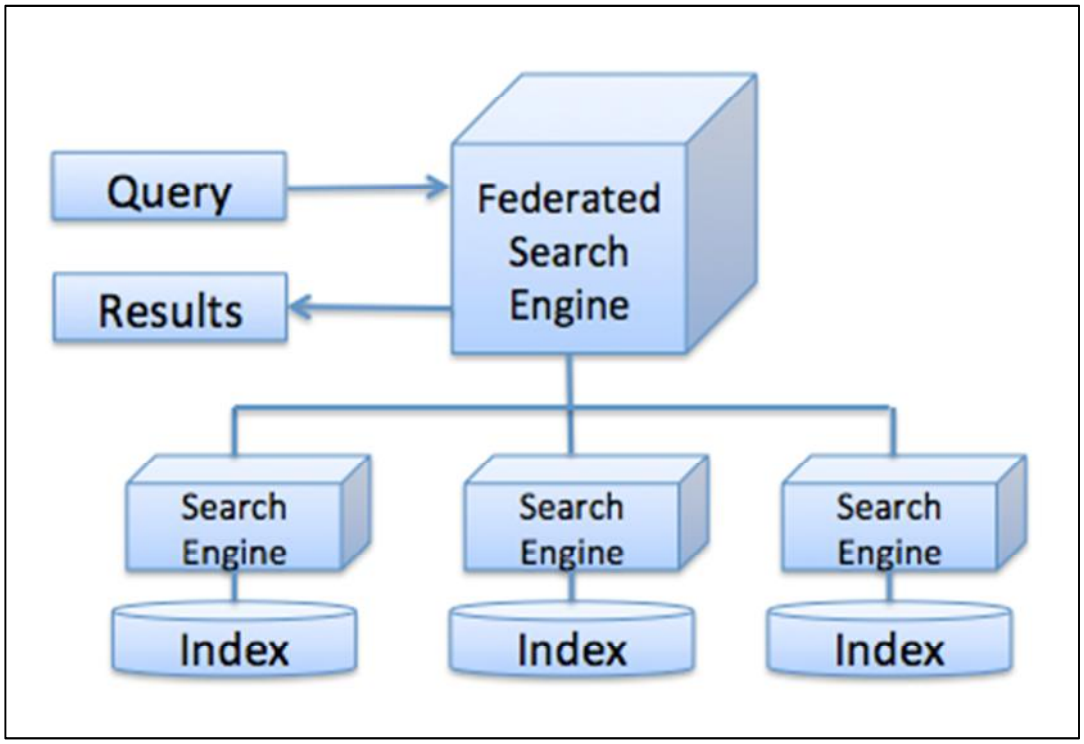


Figure 4.2 Search Architecture using Federated Search

4.3 Home Page

The home page will be displayed when users login into the web portal using any browser such as Internet Explorer or Mozilla Firefox. In this page, users can view the news as last updated. The portal will display news feed on the right of the site that the site administrator may provide to update the users on current events.

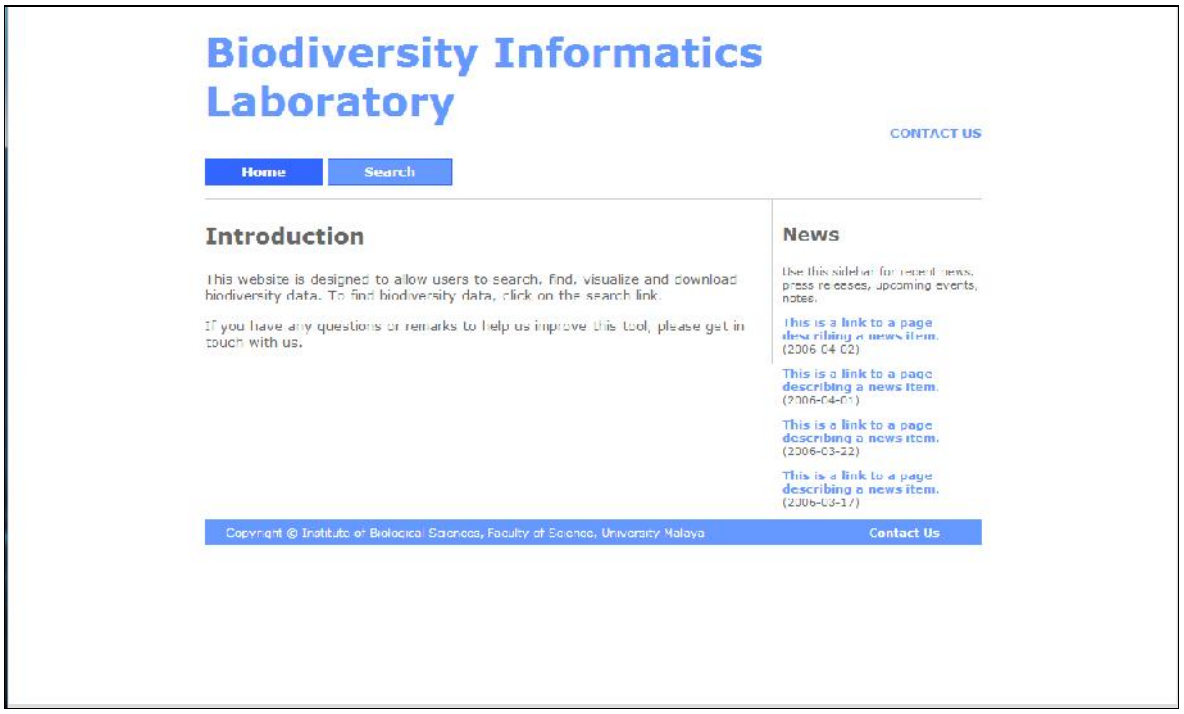


Figure 4.3 Home Page

4.4 Search Page

The parasite search page will displays the user's search result for particular parasite based on criteria that have been provided by the user. The criteria include search based on the parasite name, host's name or locality. Figure 4.3 displays the search form for the search function.

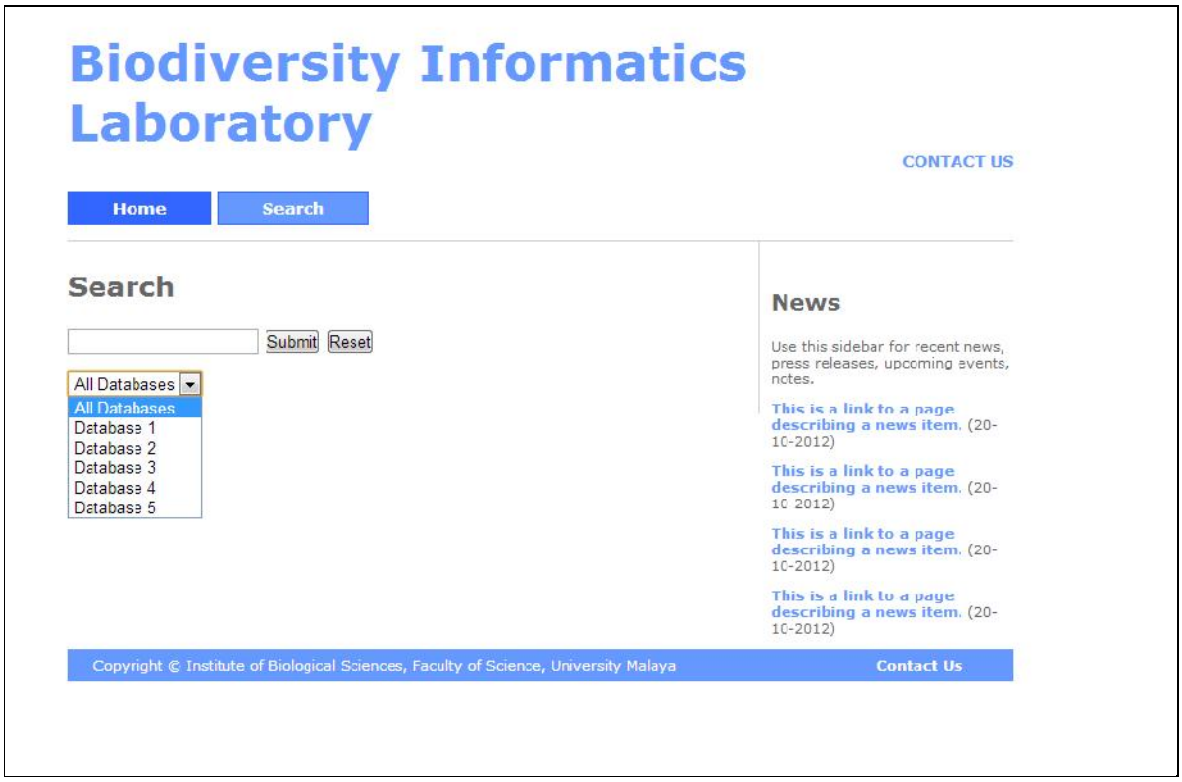


Figure 4.4 Search Form

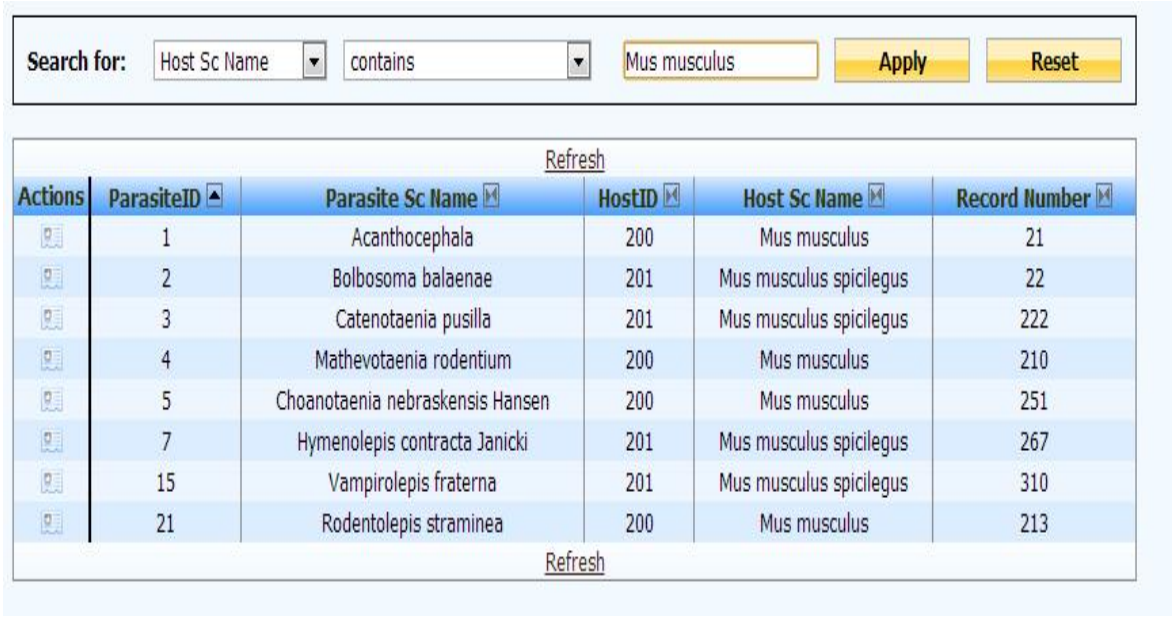


Figure 4.5 Search Result Page

4.5 Contact Form

Figure 4.5 displays the contact form for user feedback and enquiries.

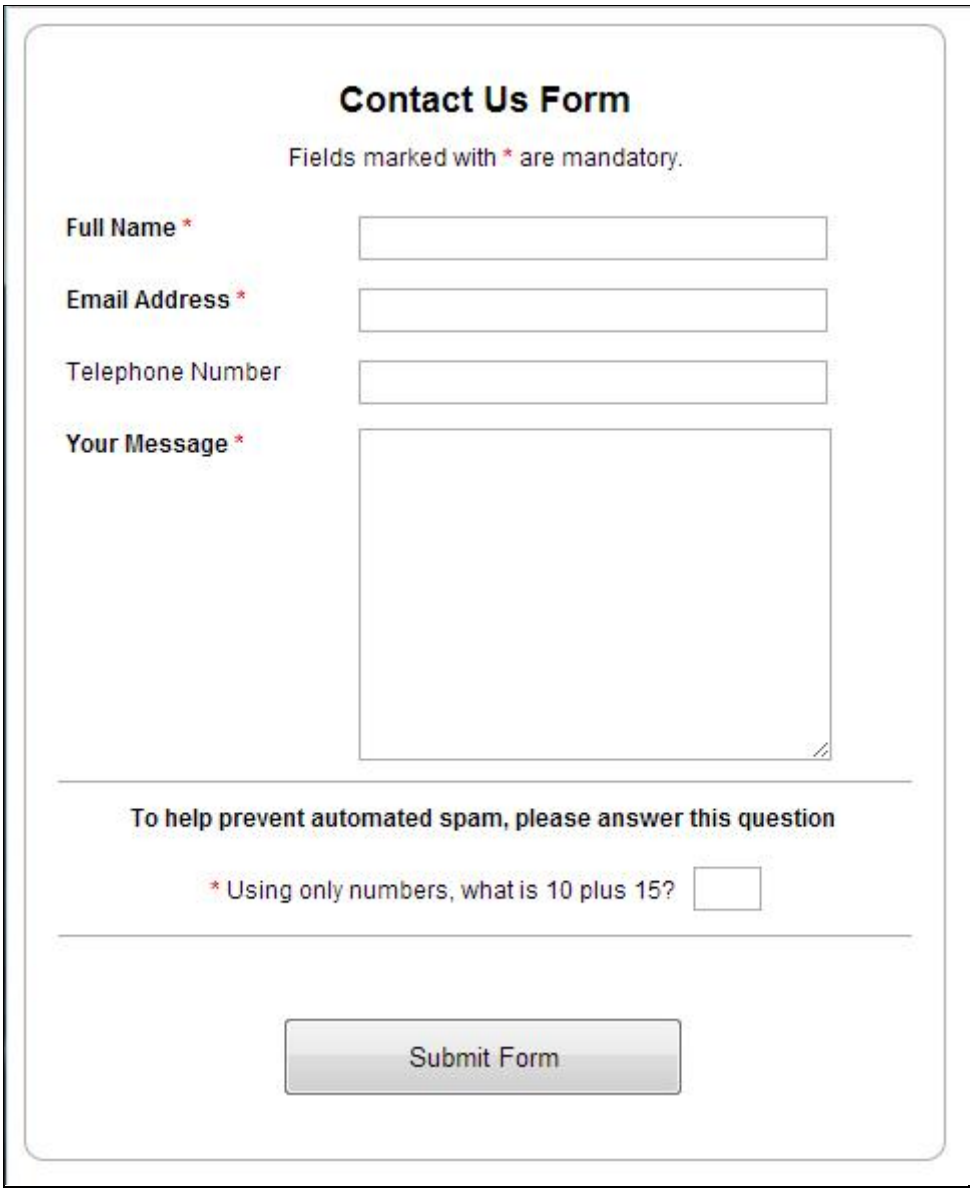


Figure 4.6 Contact Form

Chapter 5

CONCLUSIONS

As a conclusion, the objectives of the research had been achieved. The main aim of this study is to develop a web portal with multiple database search functionality and integrated web services.

There are a few recommendations to be proposed in improving the Biodiversity Informatics Laboratory Web Portal. This system can be brought into future enhancement which can provide better services to users. The system can allow in the future the contributions by non-registered users that have been validated by researchers to add to the database. A more attractive interface can assist the researcher as well as the end-user on accessing the system. The system can have some enhancement on the search function. The search function can be modified by users to suit the kind of results that the user wants.

BIBLIOGRAPHY

Aguello, J. (2011). *Federated Search for Heterogenous Environments*. (Doctoral Dissertation). Retrieved from Language Technology Institute, Carnegie Mellon University. URL: http://www.lti.cs.cmu.edu/research/dissertations.shtml

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., &Weerawarana, S. (2002). Unraveling the Web Services Web: An Introduction to SOAP, WSDL, and UDDI.IEEE Internet Computing, 6(2), 86-93.doi: 10.1109/4236.991449

Haas, H. and Brown, A. (2004). W3C Working Group Note 11 February 2004. URL: http://www.w3.org/TR/ws-gloss/. Accessed 20 October 2012.

Lim, J., Bhak, J., Oh, H. M., Kim, C. B., Park, Y. H., &Paek, W. K. (2008).*An integrated Korean biodiversity and genetic information retrieval system*.[Research Support, Non-U.S. Gov't]. BMC Bioinformatics, 9 Suppl 12, S24. doi: 10.1186/1471-2105-9-S12-S24

Seng P. *Web services programming tips and tricks* [Online]. URL:http://www.ibm.com/developerworks/websphere/library/techarticles/0511_phung-lu/0511_phung-lu.html.Accessed 12 October 2012.

Shanmughavel, P. (2007). *An overview on biodiversity information in databases*.Bioinformation, 1(9), 367-369.

Sheth, A. P., & Larson, J. A. (1990).*Federated database systems for managing distributed, heterogeneous, and autonomous databases*.ACM Comput.Surv., 22(3), 183-236. doi: 10.1145/96602.96604

W3Schools. SOAP introduction [online].

URL: http://www.w3schools.com/SOAP/soap_intro.asp. Accessed 12 October 2012.

Zouberakis, M., Chandras, C., Swertz, M., Smedley, D., Gruenberger, M., Bard, J., . . .Aidinis, V. (2010). *Mouse Resource Browser--a database of mouse databases.*[Research Support, Non-U.S. Gov't].Database (Oxford), 2010, baq010.doi: 10.1093/database/baq010