

**THE USE OF ON-LINE ANALYTICAL PROCESSING
APPROACH IN DEVELOPING HOST-PARASITE DATABASE
SYSTEM**

NUR IMTIAZAH BT. SHUHAIMI

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

2013

**THE USE OF ON-LINE ANALYTICAL PROCESSING
APPROACH IN DEVELOPING HOST-PARASITE
DATABASE SYSTEM**

NUR IMTIAZAH BT. SHUHAIMI

**DISSERTATION SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE**

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

2013

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: NUR IMTIAZAH BT. SHUHAIMI (I.C/Passport No: 860819-56-6644)

Registration/Matric No: SGR 090145

Name of Degree: MASTER OF SCIENCE (EXCEPT MATHEMATICS & SCIENCE PHILOSOPHY)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

THE USE OF ON-LINE ANALYTICAL PROCESSING APPROACH
IN DEVELOPING HOST-PARASITE DATABASE SYSTEM

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:

Designation:

ABSTRACT

Biodiversity is a very broad area to explore. The developments of research have resulted in increasing the number of biodiversity data in various formats. The ability of a database in dealing with biodiversity data is essential for the effective management and information dissemination process. The capabilities of a database system rely on the formula or approaches used in operating the system. The development of On-Line Analytical Processing (OLAP) technology from time to time expands the use of this approach which previously was dominated by the business field. The effectiveness of structuring data into cube form technique allows data modeling and analysis in multi-dimensional view has drawn the attention of researchers from a variety of other fields and to adopt this method in producing a more conducive database system. The efficiency of online analysis as well as information retrieval which can be manipulated by the user in order to get the required information is seen as one of the factors making OLAP approach very useful in building a database, particularly in biology. A few existing biological databases that used the OLAP approach have been successful in helping researchers to manage data as well as perform relevant analysis, for instance in medical data management, gene expression and molecular sequence analysis. In this study we used OLAP in developing a host-parasite database. This thesis describes the process of parasite-host data collection, data digitization and data cleaning, followed by the construction of relational database system using online analytical processing approach. This system is named as Parasite Information Network System (PINS). PINS

is designed and developed with the intention to help researchers manage data in order to preserve valuable biodiversity information by documenting the data into database storage structure. PINS is also aimed to assists researchers and students to obtain data about parasites and hosts taxonomy, biology, geography or publication of data resources information provided and use them in their studies.

ABSTRAK

Biodiversiti merupakan bidang yang sangat luas untuk diterokai. Perkembangan penyelidikan menghasilkan data-data biodiversiti yang semakin banyak dalam pelbagai format. Kebolehan pangkalan data mengurus data-data biodiversiti ini amat penting bagi pengurusan data dan penyebaran maklumat yang berkesan. Keupayaan sesuatu sistem pangkalan data bergantung kepada formula atau pendekatan yang digunakan dalam mengendalikan sistem. Perkembangan teknologi On-Line Analytical Processing (OLAP) dari semasa ke semasa memperluaskan lagi penggunaan pendekatan ini yang sebelum ini banyak diaplikasikan dalam bidang perniagaan. Keberkesanan penggunaan teknik penstrukturan data ke dalam bentuk kiub yang membolehkan pemodelan data dan analisis dalam pelbagai dimensi menarik perhatian para penyelidik dari pelbagai lapangan kajian yang lain untuk menerima pakai kaedah ini dalam menghasilkan satu sistem pangkalan data yang lebih kondusif. Kecekapan analisis atas talian serta pencarian maklumat yang boleh dimanipulasi oleh pengguna untuk mendapatkan informasi yang diperlukan dilihat sebagai salah satu faktor yang membuatkan pendekatan OLAP adalah sangat penting dan wajar dalam membina pangkalan data, terutamanya dalam biologi. Beberapa pangkalan data biologi sedia ada yang menggunakan kaedah pendekatan OLAP telah berjaya membantu penyelidik menguruskan data dan melaksanakan analisis yang berkaitan, contohnya dalam pengurusan data perubatan, analisis ekspresi gen dan turutan molekular. Dalam kajian ini kami menggunakan OLAP untuk membangunkan pangkalan data hos-parasit. Tesis

ini menerangkan proses-proses pengumpulan data parasit-hos, pendigitasian data dan pembersihan data, diikuti dengan pembangunan sistem pangkalan data relational menggunakan pendekatan pemprosesan analitikal dalam talian. Sistem ini dinamakan sebagai Parasite Information Network System (PINS). PINS direka dan dibangunkan dengan tujuan untuk membantu penyelidik menguruskan data dalam usaha memelihara maklumat biodiversiti dengan mendokumentasikan data ke dalam struktur penyimpanan pangkalan data. PINS juga bertujuan membantu penyelidik dan pelajar untuk mendapatkan maklumat parasit dan hos dalam konteks taksonomi, biologi, geografi atau maklumat penerbitan sumber data yang disediakan dan menggunakannya dalam kajian mereka.

ACKNOWLEDGEMENTS

I am very grateful to the Almighty God, the Most Gracious and Merciful for giving his blessings to me to live in peace and giving me the chances to constantly seek knowledge along the journey of life in this world.

First and foremost, I would like to express my sincere gratitude to both my supervisors, Dr. Sarinder Kaur Kashmir Singh and Professor Dr. Susan Lim Lee Hong for their endless patience, encouragement and valuable advices which are highly appreciated. Without them, this dissertation would not have been possible.

Numerous thanks to my fellow lab mates, Dr. Arpah Abu, Farhana, Evelyn, Lee Kien and Aqilah who always share their knowledge, ideas, opinions and give their supports throughout this research. I would also like to thank Aadilah Baharuddin for always helping and sharing information. Thank you also to Mazlina and Madam Kamariah Ibrahim for taking the time to proofread this thesis.

I would like to express my love and deepest gratitude to the most precious persons in my life, my dear father, Shuhaimi Mamat and my lovely mother, Mariana Junid for their love, support and sacrifices.

Finally, supports from my lecturers, family and friends will constantly be an inspiration for me to go through the twists and turns of life in the future. Once again, thank you very much.

TABLE OF CONTENTS

Abstract	iii
Abstrak	v
Acknowledgements	vii
Table of Contents	viii
List of Figures	xi
List of Tables	xiii
List of Acronyms	xv
 Chapter 1 Introduction	
1.1 Background	1
1.2 Parasites	2
1.3 Biodiversity Databases	4
1.4 On-Line Analytical Processing (OLAP)	17
1.4.1 OLAP Cube	20
1.4.2 OLAP Architectures	23
1.4.3 OLAP in Biology	26
1.4.4 OLAP in Biodiversity	29
1.4.5 Summary of OLAP Technology Application in Databases	31
1.4.6 Summary	34
1.5 Problem Statement	34
1.6 Research Objectives	35
1.7 Scope	36
1.8 Justification of the Study	36
1.9 Chapter Organization	37

Chapter 2 Materials and Methods

2.1	Introduction	39
2.2	Project Development Methodology	39
2.2.1	Development Phase	40
2.2.2	Work Flow	42
2.3	OLAP Technology Approach	46
2.4	Data Source	47
2.5	Summary	49

Chapter 3 System Requirements and Analysis

3.1	Introduction	50
3.2	Data and System Analysis	51
3.2.1	Analysis on the Monogenean Data	51
3.2.2	Analysis on the ROLAP System Architecture	53
3.3	System Requirements	55
3.3.1	User of the System	55
3.3.2	Functional Requirements	56
3.3.3	Non-Functional Requirements	62
3.4	Summary	63

Chapter 4 System Design, Implementation and Testing

4.1	Introduction	64
4.2	System Design	65
4.2.1	System Architecture	65
4.2.2	Relational Design	67
4.2.3	Multidimensional Design	70
4.2.4	User Interface Design	72
4.3	Development Tools	74
4.4	System Implementation	75
4.4.1	Data Preparation	75
4.4.2	Database Development	82
4.4.3	OLAP Cube Construction	98

4.4.4	User Interface	106
4.5	System Testing	109
4.6	Summary	113
 Chapter 5 Discussion and Conclusion		
5.1	Introduction	114
5.2	The Realization of the Research	115
5.3	The Importance of the Research	115
5.4	Future Works	118
5.5	Summary	119
 References		120

LIST OF FIGURES

Figure 1.1:	FishBase search page	6
Figure 1.2:	AntWeb result page	7
Figure 1.3:	The Reptile Database search page	8
Figure 1.4:	The Reptile Database result page	8
Figure 1.5:	The search results page in Avibase	9
Figure 1.6:	GMPD primate database searching page	10
Figure 1.7:	GMPD primate database result page	10
Figure 1.8:	The full list of monogenea species page in MonoDb	11
Figure 1.9:	The host data page in MonoDb	11
Figure 1.10:	The search page in Hexabotriidae database	12
Figure 1.11:	Host-parasite database search page interface	13
Figure 1.12:	Result page provided by host-parasite database	13
Figure 1.13:	The star schema structure	21
Figure 1.14:	The OLAP cube structure and components	22
Figure 2.1:	The flow of Systems Development Life Cycle (SDLC) methodology	40
Figure 2.2:	Workflow of the system development process	43
Figure 2.3:	Flowchart of the development process	45
Figure 3.1:	Example of the monogenean data resided in the publication ..	52
Figure 3.2:	The structure of a relational online analytical processing (ROLAP) architecture	54

List of Figures, continued

Figure 4.1:	The overall system architecture for Parasite Information Network System (PINS)	66
Figure 4.2:	Entity-relationship diagram used for developing PINS system .	69
Figure 4.3:	Parasite cube dimensional model	70
Figure 4.4:	Host cube dimensional model	71
Figure 4.5:	Locality cube dimensional model	71
Figure 4.6:	Reference cube dimensional model	72
Figure 4.7:	User interface design of PINS system	73
Figure 4.8:	Parasite-host data digitization in Microsoft Excel spreadsheet ..	80
Figure 4.9:	MySQL Workbench tool used to develop the parasite-host database	83
Figure 4.10:	Example of SQL script review applied during data insertion process	83
Figure 4.11:	Star schema for PINS system	99
Figure 4.12:	The summary of the query processing flow	100
Figure 4.13:	Connection between Mondrian server and the parasite-host database	101
Figure 4.14:	Parasite cube fact, dimensions and measurement	102
Figure 4.15:	The data schema for Parasite cube in XML format	103
Figure 4.16:	The data schema for Host cube in XML format	104
Figure 4.17:	The data schema for Locality cube in XML format	104
Figure 4.18:	The data schema for Publication cube in XML format	105
Figure 4.19:	Pentaho Mondrian screenshot of MDX query editor	106
Figure 4.20:	Screenshot of PINS homepage	107
Figure 4.21:	Screenshot of Browsing PINS page	108
Figure 4.22:	Screenshot of Mondrian page connected to PINS system	108
Figure 4.23:	Screenshot of Search page	109

LIST OF TABLES

Table 1.1:	The summary information on the existing biodiversity databases	15
Table 1.2:	Eighteen rules of OLAP system proposed by Dr. Codd	17
Table 1.3:	OLAP definition presented by Nigel Pendse	19
Table 1.4:	Comparison of OLAP architecture models	24
Table 1.5:	Summary of OLAP system application in databases	32
Table 3.1:	Functional specification of the PINS system	57
Table 4.1:	The description on the data modules used	68
Table 4.2:	Development tools used throughout PINS system development	74
Table 4.3:	Five major categories of data collected and the associated details	78
Table 4.4:	Table definitions for PINS system	85
Table 4.5:	Data field definition of table Species	86
Table 4.6:	Data field definition of table Genus	87
Table 4.7:	Data field definition of table Family	88
Table 4.8:	Data field definition of table Suborder	89
Table 4.9:	Data field definition of table Order	90
Table 4.10:	Data field definition of table Subclass	91
Table 4.11:	Data field definition of table Synonym	92
Table 4.12:	Data field definition of table Host	93
Table 4.13:	Data field definition of table Host Taxon	94
Table 4.14:	Data field definition of table Locality	95

List of Tables, continued

Table 4.15:	Data field definition of table Description	96
Table 4.16:	Data field definition of table Publication	97
Table 4.17:	Testing notation conducted on the functionality of PINS	111

LIST OF ACRONYMS

CHAR	Character String
DOLAP	Desktop On-Line Analytical Processing
ER	Entity Relationship
FAPESP	Foundation for Research Support of the State of Sao Paulo
FASMI	Fast Analysis of Shared Multidimensional Information
FK	Foreign Key
GIS	Geographic Information System
GMPD	Global Mammal Parasite Database
GXDW	GeneExpress Data Warehouse
HOLAP	Hybrid On-Line Analytical Processing
IDE	Integrated Development Environment
INBALUD	Integrating Nature and Biodiversity and Land Use Data
JSP	JavaServer Pages
LTER	Long Term Ecological Research
MDX	MultiDimensional eXpression
MOLAP	Multidimensional On-Line Analytical Processing
MS	Microsoft
MyCHM	Malaysian Biological Diversity Clearing House Mechanism
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing
PDF	Portable Document Format

List of Acronyms, continued

PHP	Hypertext Preprocessor
PINS	Parasite Information Network System
PK	Primary Key
RAM	Random-Access Memory
RDBMS	Relational Database Management System
ROLAP	Relational On-Line Analytical Processing
SDLC	Systems Development Life Cycle
SGMD	Soybean Genomics and Microarray Database
SQL	Structured Query Language
TLD	Tag Library Descriptor
URL	Uniform Resource Locator
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language

CHAPTER 1

INTRODUCTION

1.1 Background

Efforts to develop biological databases are needed in order to assist biologists or researchers to organize and manage their data systematically as well as collecting and sharing information with other researchers. Currently, various methods and techniques are being used to develop the database and there is a variety of software technologies designed to help the process of information analysis which is parallel to the ongoing growth of a more complex data especially in biology. As of now, there are still lots of biological data in the conventional storage system and developing an information storage management system is sorely needed and should be intensified immediately. As has been noted, database development in various biological disciplines is able to promote the accretion of biological information resources. Besides developing a database to manage and structure the complex biological data, it is highly important to include data analytic methodologies which can be easily used by biologists in order to make use of the available data. Hence, in this thesis a database on host parasites was developed using Relational Database Management System (RDBMS) along with the On-Line Analytical Processing (OLAP) approach for data analytics.

1.2 Parasites

According to Bush, Fernandez, Esch, & Seed (2001) in the book entitled 'Parasitism: The diversity and ecology of animal parasites', parasite is an organism that inhabit on or in another larger animal called host which is classified into several groups; protozoans, platyhelminthes, acanthocephalans and nematodes. Parasites are dependent on the host, which means the host supply food and provide a habitat for them. Most parasites are typically harmful to the host, but parasites do not always cause damage to the host (Rohde, 2010). In general, parasite can be divided into two main categories: ectoparasites and endoparasites (Cheng, 2012). Endoparasites are parasites that live in the body of the host whereas ectoparasites are parasites that live on the body surface of the host. Other than that, parasites can be clustered according to their life cycle behavior, whether it is direct or indirect life cycle. Direct life cycle describes that the parasites require only one host to complete their whole life cycle, whereas indirect life cycle describes that the parasites need more than one host to complete their life cycle (Bush, Fernandez, Esch, & Seed, 2001). The symbiotic relationships between parasites and hosts can be classified into three groups: mutualism, commensalism and parasitism (Campbell & Reece, 2002). Symbiotic relationship shows how parasite and their host interact with each other. There are three possible conditions happening within parasite-host relationships; either both sides will get benefits, parasite causes damage to the host, or there is no effect to both parasite and host.

As mentioned by Bush, Fernandez, Esch, & Seed (2001), parasites can be divided into several classes of species. Protozoa is categorized as unicellular parasites while helminths and arthropods are multicellular parasites. Protozoa can be grouped into three

different phyla: Sarcomastigophora, Ciliophora and Apicomplexa (Hickman, Roberts, Larson, I'Anson, & Eisenhour, 2006). Helminths are worm-like parasites which have complex life cycle. Helminths can be clustered into nematodes and platyhelminths. Nematodes are normally recognized as roundworm whereas platyhelminths are usually known as flatworm. Platyhelminths can then be separated into four different classes: Monogenea, Turbellaria, Trematodes and Cestodes (Cullen, 2009).

Parasite of the class Monogenea is going to be the main focus in this study. According to Huyse, Audenaert, & Volckaert (2003), the Monogenea is one of the largest groups of Platyhelminthes. Studies have shown that monogenean parasites infest mainly on marine and freshwater fishes in which these monogeneans are typically found on host's gills and skins (Sinnappah et al., 2001). According to Sinnappah et al. (2001), monogenean is divided into two suborders; Monopisthocotylea and Polyopisthocotylea. Monogenean suborders were divided further into two subclasses of Polyonchoinea and Oligonchoinea. However, subsequent studies have discovered a new subclass and named it as Polystomatoinea. Besides fishes, a few monogeneans were also found on amphibians and reptiles.

Southeast Asia countries especially Malaysia is a country rich with diverse sources of flora and fauna. Scientific studies involving the natural wildlife have been conducted over the years, including studies concerning monogenean parasites. There is an estimation of the possible numbers of monogeneans that could be present on or in fishes and turtles in Peninsular Malaysia which indicates that only 8% of the monogeneans are presently known (Lim, 1998). Although the percentages shown in this study is very low, but there is a wealth of information available from taxonomic data (Lim, 1998). A study

that was conducted at the University of Florida also shows that there are more than 100 families of monogeneans found on fishes of the world, in fresh and salt water, and at a variety of temperatures (Reed, Francis-Floyd, & Klinger, n.d.) which reinforces the fact that lots of information on monogeneans exist.

The research on monogeneans in Peninsular Malaysia was done to study parasites diversity and host distribution patterns (Lim, 1990). Studies conducted by Lim (1990) resulted in the discovery of some monogenean genera for example *Dactylogyrus*, *Dactylogyroides*, *Silurodiscoides*, *Cornudiscoides*, *Bifurcohaptor*, *Bychowskyella*, *Trianchoratus*, *Gyrodactylus* and *Malayanodiscoides*. All the findings on taxonomic information, description of biological information and illustration obtained from the research were documented in the form of complete article publications.

1.3 Biodiversity Databases

Biodiversity informatics is an emerging field that applies information management tools to the management and analysis of species occurrence, taxonomic character, and image data (Johnson, 2007). Considering a lot of biological data are still stored in flat files and spreadsheets which is mainly due to current database system lacking key functionalities needed for biological data (Eltabakh et al., 2008). Hence, to overcome this deficiency, Eltabakh et al. (2008) has introduced bdbms, an extensible prototype database engine that is developed to complement the requirements of biological databases. Eltabakh et al. (2008) focuses on three parts, which is annotation and provenance management that provides annotation storing and organizing management, tracking and annotating any changes made to data and also provides a monitoring system called Content-based

Authorization for monitoring the identity of users and the modified data. Consequently, database development in various biological disciplines is able to promote the accretion of biological information resources.

Biodiversity databases store biodiversity information such as taxonomy data and descriptive biological data of living things. Globally, there are lots of biodiversity databases focusing on several specific organisms. Biodiversity database is constructed based on variety of domains such as fishes, birds, insects, reptiles, mammal and parasites.

(i) **FishBase** (<http://www.fishbase.org/search.php>)

FishBase is a web-based global information system on fishes and is collaborated with LarvalBase project which holds extended information on the fish larvae. FishBase system was developed using relational database management model. FishBase was centralized at the WorldFish Centre with the involvement of numerous collaborators over the world. The purpose of this page is to provide information on taxonomy data such as scientific name, family, order, class, genus, synonym and common names, locations of species habitat, ecosystem information, images, and information on various fish-related topics. This database is aimed to provide information to different professionals such as researchers, fishery managers, zoologists and many more. As of September 2012, FishBase contains a total of 32,500 species data, 299,700 common names available in multiple languages, 52,500 pictures and 48,700 references with 2010 collaborators. Figure 1.1 shows the search page in FishBase system.

Figure 1.1: FishBase search page.

(ii) **AntWeb** (<http://www.antweb.org/>)

AntWeb is the largest online database providing information on ants. AntWeb provides knowledge on specimen records, taxonomic data and images of ants. The purpose of this page is to publish high quality images of all ant species for the scientific community. AntWeb integrates information from various sources; taxonomic data in Excel, specimen data in Biota, and images taken using Automontage and a Leica microscope. AntWeb stores data in MySQL database. Tools for submitting images, specimen records, annotating species pages and managing regional species lists are also provided to support the study of ants and enable researchers to share and attain the information. As of March 2013, AntWeb has over 100,000 ant images, of over 25,000 specimens representing over 10,549 species. Figure 1.2 shows the interface of result page in AntWeb.



Figure 1.2: AntWeb result page.

(iii) **The Reptile Database** (<http://www.reptile-database.org/>)

The Reptile Database provides information on reptile species such as snakes, lizards, turtles and crocodiles. This database focuses on providing taxonomic data including species distribution information, literature references, ecological as well as species behavioral information. As of March 2013, The Reptile Database has collected 9,789 species records and 31,315 literature references. This database is maintained by the editor, Associate Professor Peter Uetz from Virginia Commonwealth University, and Jiri Hosek for managing the search engine aspect. Figure 1.3 shows the search page interface and Figure 1.4 shows the result page of the Reptile Database.

You are here » [home](#) » advanced search

[Home page](#)
[Advanced search](#)
[Search tips](#)
[Contact us](#)
[Global Reptile BioBlitz](#)
[reptile-database.org](#)

Advanced search

Please use the following text boxes to conduct your search. To see a complete list of every species in the Reptile Database, leave the text boxes blank and click on 'Search'. To perform an exact match against the parameters you enter, check the boxes beside the fields. More details in [search tips](#)

Search category	Search input	Exact match
Higher taxa (e.g. Crocodylia, Sauria, Viperidae, lizard, snake):	<input type="text"/>	<input type="checkbox"/>
Genus (e.g. Chamaeleo, Oligodon):	<input type="text"/>	<input type="checkbox"/>
Species epithet (e.g. elegans, ornatus):	<input type="text"/>	<input type="checkbox"/>
Subspecies (e.g. Ablepharus bivittatus lindbergi):	<input type="text"/>	<input type="checkbox"/>
Author (e.g. Boulenger, Linnaeus):	<input type="text"/>	<input type="checkbox"/>
Year (e.g. 2006):	<input type="text"/>	<input type="checkbox"/>
Common name or synonym (e.g. Abronia, Amphibolurus):	<input type="text"/>	<input type="checkbox"/>
Distribution (e.g. Madagascar, Florida):	<input type="text"/>	<input type="checkbox"/>
Types (e.g. USNM 6769):	<input type="text"/>	<input type="checkbox"/>
Reference (author or title keyword):	<input type="text"/>	<input type="checkbox"/>

Figure 1.3: The Reptile Database search page.

You are here » [home](#) » [advanced search](#) » [Podarcis muralis](#)

[Home page](#)
[Advanced search](#)
[Podarcis muralis](#)
[Search tips](#)
[Contact us](#)
[Global Reptile BioBlitz](#)
[reptile-database.org](#)

Podarcis muralis (LAURENTI, 1768)






Photo: [David Gregory](#)

[Add your own observation of Podarcis muralis](#)

Find more photos by Google images search: [Google](#)

Higher Taxa	Lacertidae, Sauria (lizards)
Subspecies	Podarcis muralis albanica (BOLKAY 1919) Podarcis muralis breviceps (BOULENGER 1905) Podarcis muralis brongniardii (DAUDIN 1802) Podarcis muralis colossi (TADDEI 1949) Podarcis muralis maculiventris (WERNER 1891) Podarcis muralis muralis (LAURENTI 1768) Podarcis muralis nigroventris BONAPARTE 1838 Podarcis muralis sardus (BONAPARTE 1838)

iNaturalist.org
 Can you confirm these amateur observations of Podarcis muralis?

Figure 1.4: The Reptile Database result page.

(iv) **Avibase – The World Bird Database** (<http://avibase.bsc-eoc.org/avibase.jsp>)

Avibase is a database system containing records about all birds of the world. Avibase provides a lot of information such as distribution information, taxonomy and synonyms in several languages for about 10,000 species and 22,000 subspecies of birds. This database is hosted by Bird Studies Canada and

managed by Denis Lepage. As of August 2013, Avibase has collected a total of 11, 072, 829 records. Figure 1.5 shows the search results page in Avibase.

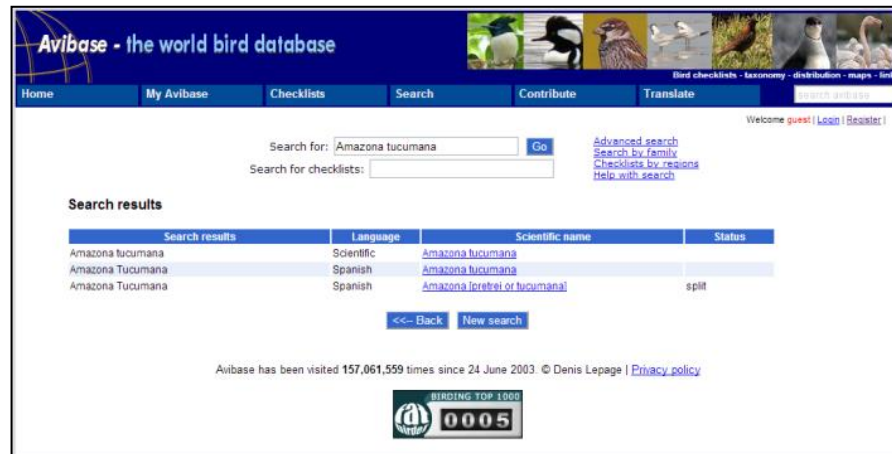


Figure 1.5: The search results page in Avibase

(v) **Global Mammal Parasite Database** (<http://www.mammalparasites.org/>)

The Global Mammal Parasite Database (GMPD) contains information on parasites and their hosts. The records in GMPD are collected from published scientific literature. GMPD consists of five databases developed to provide primates, carnivores, terrestrial hoofed mammals, marine mammals and bats information. Presently, primate, carnivore and terrestrial hoofed mammal databases are the most complete databases maintained by the project organizers. In short, primate database provides 4,000 lines of data on 139 hosts, whereas carnivore database holds 8,266 lines of data on 153 host species along with geographic information for 1,358 lines of data. Terrestrial hoofed mammal database on the other hand contains 6,858 lines of data for 175 host species and 724 parasite species. The provided data in GMPD is available online in which the output is presented in rows and columns format. Figure 1.6 and Figure 1.7

respectively show the interface of search and result page in GMPD primate database system.

GMPD Primate Database Page

Please select a host or hosts

Taxonomy Family: Genus:

Scientific Name

Text based search:

Please select a parasite or parasites

Parasite Type

Scientific Name

Text based search: ☐ Check here to exclude zero prevalence

Please select a location

Geopolitical Nation: Continent:

Text based search:

Figure 1.6: GMPD primate database searching page.

GMPD Primate Database Search Results

QUERY STRING RETURNED 19 LINES OF DATA

Host	Parasite	Location	Reference	Animals Sampled	Is Prevalence Reported?	Is Prevalence Zero?
Ateles fusciceps	Plasmodium brasilianum	PANAMA	Coatney, G. R., W. E. Collins, and W. McWilson. 1971. The primate malaria. Bethesda, National institute of allergy and infectious diseases.	unknown	no	no
Ateles fusciceps	Dipetalonema gracile	Panama	http://www.lpsi.barc.usda.gov/bapcu/index.html	unknown	no	no
Ateles fusciceps	Microfilaria obtusa	Darien Province	McCoy, O. R. 1936. Filarial parasites of the monkeys of Panama. American Journal of Tropical Medicine 16(4):383-403	18	yes	yes
Ateles fusciceps	Microfilaria panamensis	Darien Province	McCoy, O. R. 1936. Filarial parasites of the monkeys of Panama. American Journal of Tropical Medicine 16(4):383-403	18	yes	yes
Ateles fusciceps	Trypanosoma minasense	PANAMA	Sousa, O. E., R. N. Rossan, and D. C. Baerg. 1974. The Prevalence of Trypanosomes and Microfilariae in Panamanian Monkeys. The American Journal Of Tropical Medicine and Hygiene 23:862- 867.	87	yes	no
Ateles fusciceps	Ascaris lumbricoides	Panama	http://www.lpsi.barc.usda.gov/bapcu/index.html	unknown	no	no
Ateles fusciceps	Trypanosoma rangeli	PANAMA	Sousa, O. E., R. N. Rossan, and D. C. Baerg. 1974. The Prevalence of Trypanosomes and Microfilariae in Panamanian Monkeys. The American Journal Of Tropical Medicine and Hygiene 23:862- 867.	87	yes	yes
Ateles fusciceps	Deltaretrovirus sp.	Darien	Kaplan, J. E., M. U. Holland, D. B. Green, F. Gracia, and W. C. Reeves. 1993. Failure to Detect Human T-Lymphotropic Virus-Antibody in Wild- Caught New-World Primates. American Journal of Tropical Medicine and Hygiene 49:236-238.	75	yes	yes
Ateles fusciceps	Oxyurosema sp.	Panama	http://www.lpsi.barc.usda.gov/bapcu/index.html	unknown	no	no
Ateles fusciceps	Buckleyenterobius sp.	Panama	http://www.lpsi.barc.usda.gov/bapcu/index.html	unknown	no	no

Figure 1.7: GMPD primate database result page.

(vi) **MonoDb** (<http://www.monodb.org>)

MonoDb is a web-host for the parasite monogenea. MonoDb provides information on summary data of the species, genus and family of the monogenea as well as the respective host information. Figure 1.8 shows the full list of monogenea species page and Figure 1.9 shows the host data page in MonoDb.

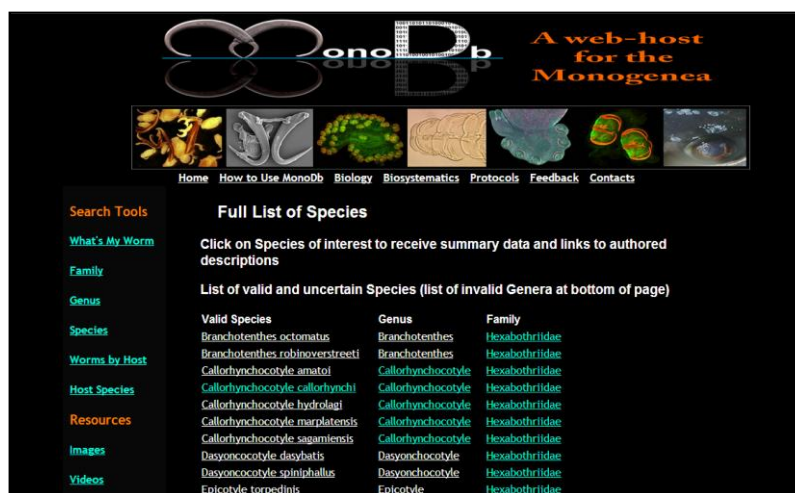


Figure 1.8: The full list of monogenea species page in MonoDb

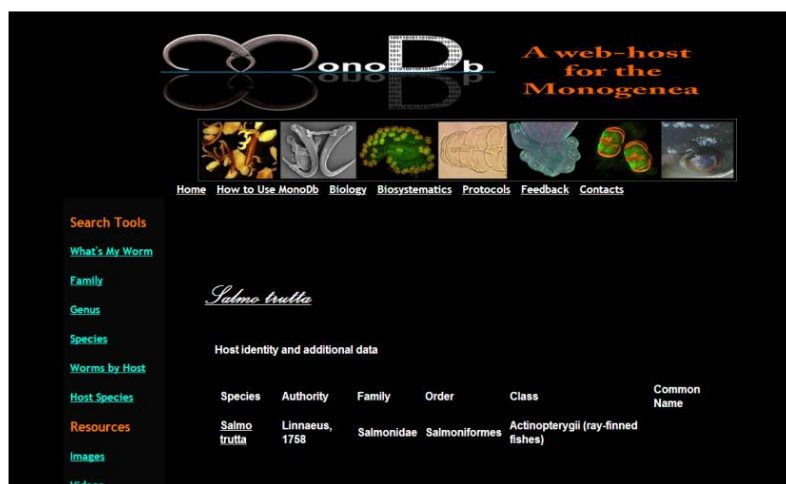


Figure 1.9: The host data page in MonoDb

(vii) **Hexabotrhidae Database** (<http://www.ib.usp.br/~mvdomingues/hexa/#>)

Hexabotrhidae database is a database containing records about hexabotrhids parasite and their hosts. This database provides information on the valid species, synonymies, and literature, as well as the host records. Hexabotrhidae database is supported by the funder of scientific and technological research in Brazil, The Foundation for Research Support of the State of Sao Paulo (FAPESP) and managed by Dr. Marcus V. Domingues from Universidade Federal de Sao Paulo, Brazil. Figure 1.10 shows the search page in Hexabotrhidae database.

Figure 1.10: The search page in Hexabotrhidae database

(viii) **Natural History Museum Host-Parasite Database** (<http://www.nhm.ac.uk/>)

Natural History Museum website provides information on taxonomy, systematic, biodiversity, evolution, planetary sciences and natural resources in various fields of study in biodiversity. One of the databases available is host-parasite database which contains information about parasite species and associate host species as well as locality information extracted from 28,000 references. Host-Parasite database allows retrieval of information by parasite group, hosts, localities and

references as shown in Figure 1.11. Figure 1.12 shows the result page in host-parasite database which displayed information on parasite and host scientific name, location of specimen discovery and complete reference information of the data obtained.

The screenshot shows the 'Host-parasite database' search page. At the top, there's a header with the Natural History Museum logo and navigation links like 'Home', 'Visit us', 'Nature online', etc. Below the header, there's a breadcrumb trail: 'You are here: Home > Research and curation > Scientific resources > Taxonomy and systematics > Host-parasite database > Search database'. The main content area is titled 'Host-parasite database' and 'Search for parasites or hosts'. It contains several search criteria sections: 'Parasite' with dropdowns for Group (Monogeneans), Subgroup (Starts with), Genus (Starts with Wallagothema), and Species (Starts with); 'Host' with dropdowns for Genus (Starts with) and Species (Starts with); 'Other information' with fields for Location (selected), Host state, and Parasite status; and 'Display options' with checkboxes for 'Show parasites', 'Show parasite grouping', and 'Show hosts'. A 'search references' button is located at the top right of the search criteria section. A 'Toolbox' on the right side offers 'Print version' and 'Email this page' options.

Figure 1.11: Host-parasite database search page.

The screenshot shows the 'Host-parasite database' result page. At the top, there's a header with the Natural History Museum logo and navigation links. Below the header, there's a breadcrumb trail: 'You are here: Home > Research and curation > Scientific resources > Taxonomy and systematics > Host-parasite database > Search database'. The main content area is titled 'Host-parasite database' and 'Search criteria:'. It displays the search results for the criteria entered: 'Parasite group: Monogeneans', 'Parasite genus: Wallagothema', 'Parasite species: chauhanii Agarwal & Pandey, 1981', 'Host genus: Wallago', 'Host species: affu', and 'Host www'. Below the search criteria, there's a 'References 1-2 of 2' section. It lists two references: 1. Parasite: Wallagothema chauhanii Agarwal & Pandey, 1981. From: Wallago affu. In the wild. Locality: Oriental. Comments: to Thaparodeidus. Reference: (Lim, L.H.S.) (1996) Thaparodeidus Jain, 1952, the senior synonym of Skurodiscoides Gussev, 1976 (Monogenea: Ancylodiscoidinae). Systematic Parasitology, Dordrecht 35, 207-215, 1 tab. 2. Parasite: Wallagothema chauhanii Agarwal & Pandey, 1981. From: Wallago affu. In the wild. Locality: India. Comments: Not original. Reference: (Lim, L.H.S. & Lersuthichawal, T.) (1996) Monogeneans from Wallago affu (Blotch & Schneider, 1802) of Thailand. Raffles Bulletin of Zoology, Singapore, 44(1): 287-300, 4 figs, 2 tabs. A 'Page: 1 of 1' indicator and navigation buttons (Previous page, Down, Next page) are visible at the bottom of the references section.

Figure 1.12: Result page provided by host-parasite database.

Existing biodiversity databases explained previously shows how the system is used as a data center. Most systems got their data from various sources such as scientific articles, books and websites. In addition, they also provide facilities for any researchers or individuals to contribute and share their data by sending email to the developer. Each data received will be reviewed by the experts prior to uploading to the website. Based on the stated existing database, fish database for instance provides a relatively complete source of information in all aspects compared to parasite database which display quite static and limited information.

Table 1.1 below briefly summarizes the information and features of existing databases as described previously. This description helps in getting a thorough idea to produce a complete database system which provides comprehensive information sources for the user.

Table 1.1: The summary information on the existing biodiversity databases.

Information / Features	Existing Biodiversity Databases							
	FishBase	AntWeb	The Reptile Database	Avibase	Global Mammal Parasite Database	MonoDb	Hexabotrhiidae Database	Host-Parasite Database
Developer	Daniel Pauly and Rainer Froese, Leibniz Institute of Marine Sciences, Germany	California Academy of Sciences	Associate Professor Peter Uetz, Virginia Commonwealth University and Jiri Hosek	Denis Lepage (Bird Studies Canada)	Charlie Nunn, Harvard University Cambridge, Sonia Altizer and project contributors	-	Collaboration between University of Adelaide, South Australian Museum, Universidade de São Paulo	Dr. David Gibson, Natural History Museum, London
System objectives	To provide information on taxonomy data, habitat, ecosystem information, images, and other fish-related topics	To provide high quality images of ant species, distribution maps, and access to the original description	To provide information on taxonomic data, species distribution, references, and ecological and species behavioral	To provide information on taxonomy data, species distribution, and other bird-related topics	To provide information of hosts and parasites in different mammal databases	To provide information on taxonomy data for monogenea as well as the summary information on respective host	To provide information on the valid species, synonymies, and literature of the hexabotrhiids parasite	To provide information on parasitic worms, their host and their primary citations
Main content	Fish (specifically finfish)	Ants	Reptiles	Birds	Primates, carnivores, terrestrial hoofed mammals, marine mammals and bats	Monogenean parasites	Hexabotrhiids parasites	Parasitic worms

Table 1.1, continued.

Information / Features	Existing Biodiversity Databases							
	FishBase	AntWeb	The Reptile Database	Avibase	Global Mammal Parasite Database	MonoDb	Hexabotrhiidae Database	Host-Parasite Database
Database management system	MySQL	MySQL	-	-	MySQL, PostgreSQL	-	-	-
Language	Php	Java	-	Java	-	-	-	-
System-based	Web-based	Web-based	Web-based	Web-based	Web-based	Web-based	Web-based	Web-based
Interface	Simple and user friendly	Simple and user friendly	Simple and user friendly	Simple and user friendly	Simple and user friendly	Simple and user friendly	Simple and user friendly	Simple and user friendly
URL	http://www.fishbase.org	http://www.antweb.org	http://www.reptile-database.org/	http://avibase.bsc-eoc.org/avibase.jsp	http://www.mammalparasites.org/	http://www.monodb.org	http://www.ib.usp.br/~mvdomingues/hexa/#	http://www.nhm.ac.uk/

1.4 On-Line Analytical Processing (OLAP)

One of the technologies which is gaining a place among database developers is the OLAP technology. OLAP is the acronym for On-Line Analytical Processing, a term introduced by Dr. E. F. Codd in a paper entitled “Providing On-Line Analytical Processing to User Analysts: An IT Mandate” (Mallach, 2000). In attempting to clearly define the definition of OLAP system, Dr. Codd proposed twelve rules for the OLAP system in 1993, and added six more new rules in 1995. Dr. Codd then restructured the rules into four categories. Dr. Codd’s 18 rules of OLAP system are briefly explained by Pendse (2008) as described in the Table 1.2 below:

Table 1.2: Eighteen rules of OLAP system proposed by Dr. Codd (Pendse, 2008).

Features	Rules
Basic Features B	<ol style="list-style-type: none"> 1. Multidimensional conceptual view <ul style="list-style-type: none"> • Basic of OLAP system which simplifies dimensional model design and analysis. 2. Intuitive data manipulation <ul style="list-style-type: none"> • Reduces steps performed in obtaining more detailed information. 3. Accessibility <ul style="list-style-type: none"> • OLAP tool as a middleware engine should be able to access various data sources. 4. Client-server architecture <ul style="list-style-type: none"> • Able to perform mapping and data integration from different databases. 5. Transparency <ul style="list-style-type: none"> • Open system architecture allows OLAP to be embedded in any place without impacting the functionality of the host system. 6. Multi-user support <ul style="list-style-type: none"> • Provides concurrent access in every respect including retrieval and update.

Table 1.2, continued.

Features	Rules
Basic Features B	<p>7. Data extraction and interpretation</p> <ul style="list-style-type: none"> Allows the transition from pre-aggregated data to a detail record level, and vice versa. <p>8. OLAP analysis models</p> <ul style="list-style-type: none"> Able to support all analysis models; categorical, exegetical, contemplative and formulaic models.
Special Features S	<p>9. Treatment of non-normalized data</p> <ul style="list-style-type: none"> Integration process between OLAP engines and denormalized source data where any updates in OLAP environment cannot change the data stored in the system. <p>10. Storing OLAP results</p> <ul style="list-style-type: none"> OLAP data changes should be stored separately from transaction data. <p>11. Extraction of missing values</p> <ul style="list-style-type: none"> All the missing values are separated into a uniform representation to distinguish them from zero values <p>12. Treatment of missing values</p> <ul style="list-style-type: none"> Missing values can be ignored, regardless of their sources.
Reporting Features R	<p>13. Flexible reporting</p> <ul style="list-style-type: none"> Reporting facilities should be able to present information in various display view. <p>14. Consistent reporting performance</p> <ul style="list-style-type: none"> Ensure the competency of OLAP performance due to increased dimensions or size of database. <p>15. Automatic adjustment of physical level</p> <ul style="list-style-type: none"> Physical schema of OLAP server should be able to automatically adapt to the type of model, data volumes and sparsity.
Dimension Control D	<p>16. Generic dimensionality</p> <ul style="list-style-type: none"> Each data dimension must be analogous in terms of its structure and operational capabilities. <p>17. Unlimited dimensions and aggregation levels</p> <ul style="list-style-type: none"> Number of data dimensions and aggregation levels are dependent on the system requirements. <p>18. Unrestricted cross-dimensional operations</p> <ul style="list-style-type: none"> No restrictions on data manipulation across dimensions or in the relationship between data cells.

OLAP is also defined as Fast Analysis of Shared Multidimensional Information (FASMI), described by Pendse (2008). He presented OLAP application features without bothering on how the OLAP system is developed. Pendse summarized Dr. Codd's rules and formulated them into five simple definitions as described in Table 1.3 below. The definition on FASMI provides better understanding about OLAP characteristics which is now widely adopted compared to Dr. Codd's rules (Pendse, 2008).

Table 1.3: OLAP definition presented by Nigel Pendse.

Features	Definition
Fast	'Fast' refers to the production of a quick response to the user using simple analysis
Analysis	'Analysis' refers to the accessibility of the system using business logic and relevant application procedures for the users
Shared	'Shared' refers to the necessity of a comprehensive data security system implementation
Multidimensional	'Multidimensional' refers to the fundamental key in OLAP system which provides a multi-dimensional concept that gives support to the OLAP operational features
Information	'Information' refers to the information available that can be accessed from the system

OLAP is an approach used to provide quick answers to the database queries. This approach is widely used in business areas where it meets the needs of business analysts in business management, providing reports on sales or profits yearly, monthly or even daily and also data mining. This approach is not solely bounded to the business field, but also suitable to be applied in different areas such as biology. However, some modifications to the approach should be done to suit the characteristics of data used.

1.4.1 OLAP Cube

OLAP cube is the basic structure of OLAP system and also known as data cube or multidimensional cube. OLAP cube is the main element in the overall OLAP system development process. It is a data structure that conducted the analysis efficiently and is able to overcome few weaknesses in relational database. OLAP cubes can be considered similar to the tables in a relational database system. The cube structure is actually based on the data from the columns and rows in the data sources and is arranged according to specific classifications and hierarchies. The cubes basically consist of dimensions and measures. Dimensions contain members that can be further refined into several levels, hierarchies and attributes. The cube metadata is built through a specific schema. Generally, the schema is named as star schema. The term 'star' used is due to the arrangement of the schema structure which resembles the star shape. But in some cases, star schema is modified by adding on related tables to the dimensions forming linked schema named as snowflake schema. In different cases, the star schema is modified by splitting the fact table into several fact tables forming fact constellation schema.

Star schema is a data structuring model which represents data cube structure in the form of tables consisting of fact table and dimensional tables (Giovinazzo, 2000). Fact table contains fact data, is usually in the center which is linked to numerous related tables known as dimensional tables. A dimensional table consists of members which contains information named as attributes that can be classified into a number of levels. These levels form a hierarchy. Hierarchy enables parent elements to be segregated into individual data or enable children elements to be consolidated into summarized data. Figure 1.13 shows the arrangement of a star schema model. In relational online analytical processing concept, OLAP restructure the relational data structure into dimensional data structure to run the system. Data is stored in a relational database and the schema structure is designed to build up data cubes. As shown in Figure 1.14 below, we can see the transformation from the star schema to a data cube structure where the fact table in schema is forming the measures on data cube and dimensional tables in star schema forming dimensions on the data cube.

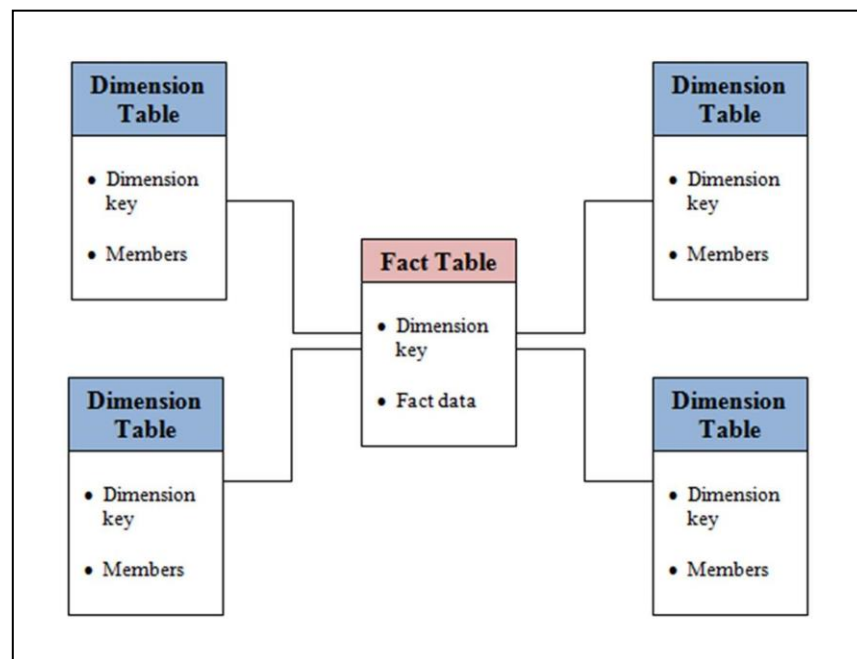


Figure 1.13: The star schema structure.

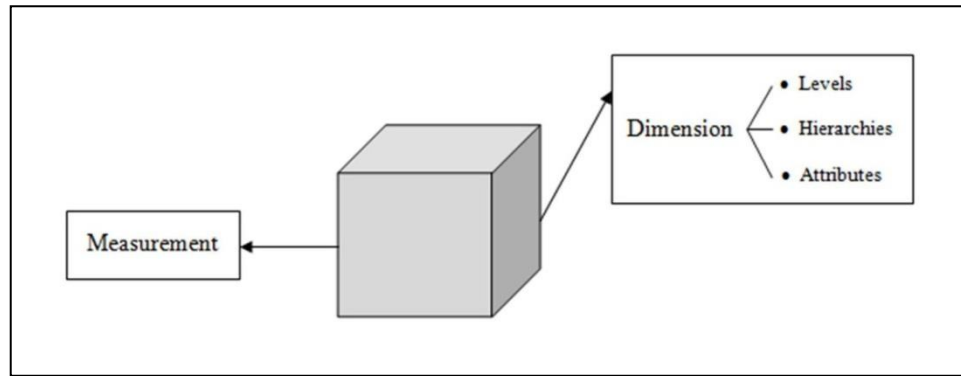


Figure 1.14: The OLAP cube structure and components.

Giovinazzo (2000) described the OLAP cube functionality that operates throughout in two main ways; i.e. slicing and dicing, roll-up and drill down. The cube structure itself connotes a variety of perspectives in which the slicing and dicing operations allow the user to view data from different point of views based on the dimensions provided. Slicing is the process of cutting the cube to focus on more specific data while dicing allows the user to rotate the cube to analyze data across cubes in two or more dimensions. Another function of OLAP cube is roll-up and drill down. The organization of data cube into hierarchical structure as mentioned earlier allows drilling down and rolling up operations. These operations allow the functions to move up or down the hierarchical level to provide users the required summarized data or detailed individual data.

1.4.2 OLAP Architectures

OLAP approach is a methodology used in a system implementation according to a specific architecture model. There are several types of OLAP system architecture models which normally used; MOLAP, ROLAP and DOLAP (Mallach, 2000). However, there is also some system developed using HOLAP model.

MOLAP is the abbreviation for multidimensional online analytical processing where it stores data in multi-dimensional databases. MOLAP is the traditional method used in the application of OLAP approach. The architecture of this model is based on the multidimensional cube structure. MOLAP model is a two-tier client/server architecture (Yoo, n.d.) where the MOLAP server connects directly with the user in the presentation layer. MOLAP executes data pre-calculation for storing information in the cube. MOLAP data cubes contain possible answers to several questions that will be queried from the system. Palo (www.palo.net/) is one of the examples of open source MOLAP server. It is in contrast to the relational online analytical processing which is known as ROLAP. The structure of ROLAP model is a three-tier client/server architecture (MicroStrategy, n.d.) where the data is stored in relational databases and OLAP operation is performed in ROLAP server which is then linked directly to the presentation layer or user. ROLAP pre-aggregates data to store information in its cube. Mondrian (mondrian.pentaho.com/) is one of the examples of open source ROLAP server. DOLAP stands for desktop online analytical processing. DOLAP is a desktop-based single-tier architecture (Kelly, Rehm, & Barbusinski, 2003) where it creates and transfers multidimensional datasets to the desktop machine and retrieve data from desktop databases. DOLAP uses a lot of memory spaces rather than other models.

HOLAP is the abbreviation of hybrid online analytical processing. HOLAP combines features found in MOLAP and ROLAP to form a single architecture model. HOLAP uses ROLAP technique to store data as well as technique from MOLAP to store the aggregations.

Table 1.4 below listed out the comparison of advantages and disadvantages of OLAP architecture models mentioned above.

Table 1.4: Comparison of OLAP Architecture Models.

Architecture	Advantages	Disadvantages
MOLAP	<ul style="list-style-type: none"> • pre-aggregation data process enable data to be extracted effectively • good performance in terms of retrieving information on summarized data • successful in handling numeric data 	<ul style="list-style-type: none"> • incompetent when handling large amounts of data • provides relatively slow performance in handling textual data • sometimes cause data redundancy
ROLAP	<ul style="list-style-type: none"> • capable to handle large amounts of data • efficient in handling both textual and numerical data • therefore, it is also better in handling non-aggregatable facts (textual form) • utilize the functionality available in relational databases 	<ul style="list-style-type: none"> • show a slow performance compared to MOLAP • limitations of SQL function as it is depending on the relational database query statement

Table 1.4, continued.

Architecture	Advantages	Disadvantages
DOLAP	<ul style="list-style-type: none"> • easy to set up • does not require high costs 	<ul style="list-style-type: none"> • have a relatively limited functionality compared to other architecture models
HOLAP	<ul style="list-style-type: none"> • flexibility in accessing data sources 	<ul style="list-style-type: none"> • slow performance in the process to get the detail data

OLAP provides many benefits to the users. The diversity and complexity of biological data lead to the demand of developing an efficient database that is capable of helping users in handling and managing biological data as well as the ability in performing data analysis. In general, OLAP allows users to run complex dimensional queries using efficient operations with query executions in a short time. With the capabilities in such ways, OLAP approach should be applied in various fields of study including biology. The robustness of OLAP system should be fully utilized in providing facilities and benefits to researchers not only in terms of data preservation, but also extended to the effort of processing information to produce useful knowledge which benefits not only the researchers and students, but also facilitates the global community in acquiring knowledge.

1.4.3 OLAP in Biology

OLAP has been widely used in various fields, particularly in business. The simplicity of this technology makes it a preferred option in developing analysis system in many other fields, including biology. There are several studies on the OLAP development that have been conducted in developing various systems including engineering, medical and biological related application system.

Niemi, Nummenmaa, & Thanisch (2001) states that queries are made against multidimensional cube, called OLAP cubes. Niemi, Nummenmaa, & Thanisch (2001) introduced the technique to build OLAP cubes based on the queries from the user with the assumption that a new OLAP cube is constructed for new analysis purposes. The aim in this study is to optimize cube by improving the structure of existing cubes based on posed queries, and to reduce the level of technical knowledge that a user should know in constructing the required OLAP cube. By this method, the system suggests cube design once the user submits their queries. Based from the suggested cube design, the user is able to accept or modify the cube structure by giving more queries. The sets of queries use Multidimensional Expressions (MDX) programming language to extract information from the cubes. Niemi, Nummenmaa, & Thanisch (2001) also presented three main criteria as benchmarks in OLAP cube design: completeness and minimalism of cube design, correctness of aggregations and minimal sparseness.

Bihua, Jian, Haode, & Bing (2010) proposed a web-based On-Line Analytical Process (OLAP) structure to be used in managing business data for petroleum industry. Due to some weaknesses of traditional database management system, OLAP becomes a

technology used to accomplish these requirements. Bihua, Jian, Haode, & Bing (2010) presented a multidimensional model of web-based OLAP drilling analysis system design to overcome the weaknesses of OLAP system within Client/Server structure. This subject-oriented system developed into a three-layer structure consisting of storage layer, application layer and the browser. Various OLAP analysis operations were implemented via web browser by linking Office Web Components with the analysis server. Two kinds of office web components have been inserted; PivotTable component and Chartspace component which are then linked to the analysis server. In short, user input is analyzed and operated by the OLAP analysis engine through changing input command into SQL statement before returning the results to the user interaction layer.

In other example, Qian & Qing (2009) designed and established the highway management data warehousing system model and on-line analytical processing (OLAP) system. Several information systems have been built for the highway management such as highway fees collection and inspectorate management system, road maintenance management system, vehicles management system, and bridges and tunnels management system. This resulted to the implementation of a supporting system to connect all those Online Transaction Processing System (OLTPs) to ease the data mining process and development of decision-making system. This study emphasizes on the architecture of data warehouse, data warehouse model designing and OLAP technology construction. The architecture of data warehouse is designed as layered structure containing three parts: information basis (data sources) systems, data management and analysis systems and end-user/applications. Qian & Qing (2009) used Analysis Services of Microsoft SQL Server 2005 as the data warehouse server, Visual Basic 6.0 to develop the analysis modules, and multidimensional expression language

(MDX) to query datasets with the aim of providing supporting system for manager decision-making.

Alkharouf, Jamison, & Matthews (2005) used OLAP technology to mine data from soybean genomics and microarray database (SGMD). Data collected from the experiments are stored in relational database tables in gene expression database within SGMD warehouse. The purpose of the system developed is to store information and use OLAP application to extract the time-course experiment to find SCN-infected genes on soybean roots. These findings help scientist to conduct further research in producing SCN-resistant soybean cultivars. Alkharouf, Jamison, & Matthews (2005) used Analysis Services 2000, Microsoft SQLServer2000 to build multidimensional cubes of gene expression experiment, and multidimensional extensions syntax to query cubes. OLAP report provided focuses on bringing out relevant information on number of defense genes and pathways triggered in soybean. Besides that, an almost similar research has also been done by Markowitz & Topaloglou (2001) where the application of data warehouse and OLAP concepts are used and developed to mine gene expression data on experimental animal model and cellular tissues from the GeneExpress Data Warehouse (GXDW).

In other studies, Dzeroski, Hristovski, & Peterlin (2000) used OLAP technology to identify patterns related to the removal of Y-chromosome in patients. Database of published Y-chromosome deletions contains 382 patient records which 177 different markers tests done. 364 of the entire record indicate the occurrence of chromosome deletions. Data were obtained from 34 published papers from MEDLINE bibliographic database and are stored in MS Access table. Dzeroski, Hristovski, & Peterlin (2000)

used two methods; clustering, and decision trees and OLAP. Clustering method is used to categorize patients into deleted chromosome and undeleted chromosome groups. Decision tree constructed to distinguish clusters and OLAP is used to study patient's characteristics in details within each cluster. Deletion of Y-chromosome causes infertility in males. Analysis system on Y-chromosome deletion is important for carrying out etiological diagnosis of male infertility to avoid transmission of mutations to offspring through assisted reproduction techniques.

In addition, Malmstrom, Nordenfelt, & Malmstrom (2012) built OLAP system to analyze mass spectrometry-based proteomics data. The system, named Xplor is built based on three components; protocols, OLAP model and viewer. Malmstrom, Nordenfelt, & Malmstrom (2012) used relational OLTP database model to store the original data, processed data and combine some online sources. OLAP model which was constructed consists of a subset of available data from OLTP databases and the model aggregated data in various ways to obtain quick analysis results. The third component in Xplor can network the first and second components by manipulating OLAP models and interact with the original data in OLTP model. Xplor system developed using perl and R programming language, and the cubes and tables are stored in MySQL 5.5.

1.4.4 OLAP in Biodiversity

Biodiversity databases nowadays are extensively developed. This shows a positive benchmark to the progress of biodiversity information. As can be seen in examples presented in section 1.4.3, OLAP approach is mostly adapted in genomics or medical-

related databases. There is still paucity of OLAP technology application developed in biodiversity databases especially in parasitological domain. INBALUD and LTER projects are some examples of biodiversity-related databases systems that were developed using the OLAP technology.

INBALUD project is one of the biodiversity projects developed using OLAP technology. INBALUD stands for Integrating Nature and Biodiversity and Land Use Data, is a project led by GeoVille, a private sector enterprise based in Luxembourg and Austria (INBALUD, 2012). INBALUD project was conducted to gather heterogeneous biodiversity data sources and develop an information central that can be accessed by the users (Milego, 2012). In INBALUD system, the input data is converted into raster format before it is combined into the INBALUD OLAP cube. OLAP cube processes the data and provides results in the form of maps, graphics and statistics.

LTER project refers to the Long Term Ecological Research project. LTER project was conducted to develop an information management system for the coastal waters at Kenting, Taiwan (Chang, Lee, & Lai, 2008). This project implemented three information technologies; web database management system, web geographic information system (GIS), and data warehousing. LTER project used Microsoft Access to store information on fish, coral reef and alga; and Autodesk MapGuide software to establish a web GIS. Microsoft SQL Server 2000 enterprise edition is used to develop the data warehouse and OLAP service for knowledge management execution.

1.4.5 Summary of OLAP Technology Application in Databases

Table 1.5 shows a summary of the literature review done on several applications of OLAP technology in databases which were developed for the use in various fields including biology and biodiversity domains.

Table 1.5: Summary of OLAP system application in databases.

Reference	Research Objectives	Implementation
Niemi, Nummenmaa, & Thanisch, 2001	To build and optimize OLAP cubes by improving cube structure based on the queries from the user	Cube design is suggested from user's queries and user can accept or modify cube structure before using MDX to extract information
Bihua, Jian, Haode, & Bing, 2010	To build a web-based OLAP structure in managing data for petroleum industry	OLAP analysis engine analyze and operate input command into SQL statement and return the result to user interaction layer
Qian & Qing, 2009	To design and establish highway management data warehousing system model and OLAP system.	Connect OLTPs into highway warehouse using Analysis Services of Microsoft SQL Server 2005 as data warehouse server, and Visual Basic 6.0 to develop the analysis modules
Alkharouf, Jamison, & Matthews, 2005	To mine data from soybean genomics and microarray database (SGMD) and use OLAP to extract the time-course experiment to find SCN-infected genes on soybean roots	Use Analysis Services 2000 to build multidimensional cubes of gene expression experiment

Table 1.5, continued.

Reference	Research Objectives	Implementation
Dzeroski, Hristovski, & Peterlin, 2000	To identify patterns related to the removal of Y-chromosome in patients	Data were stored in MS Access table. Two methods used; clustering to group patients according chromosome deletion and decision tree construction to distinguish clusters and OLAP to study patient's characteristics in details within each cluster
Malmstrom, Nordenfelt, & Malmstrom, 2012	To build Xplor system to analyze mass spectrometry-based proteomics data	Xplor is developed using perl and R language based on three components; protocols, OLAP model and viewer
Milego, 2012	To gather heterogeneous biodiversity data sources and develop accessible information central	Input data is converted into raster format and combined into INBALUD OLAP cube which then provide results in maps, graphics and statistics format
Chang, Lee, & Lai, 2008	To gather information on fish, coral reef and alga, and develop the information management system	Data were stored in Microsoft Access, and Autodesk MapGuide and Microsoft SQL Server are used to establish the web GIS and develop data warehouse and OLAP analysis system

1.4.6 Summary

Biodiversity database development has tremendously grown as proven today. Many databases provide a wide variety of information resources covering all aspects in biodiversity. However, biodiversity database development in Malaysia is relatively slow although undoubtedly many scientific studies have been conducted by experts here. Thus, the need to discover new data has increased and this requires the effort to build a data repository system to collect and manage information in a proper way. Studies found that there were difficulties in obtaining resources such as websites that provide parasite database which can be used to mine information and run analysis. Therefore, the construction of host-parasite database using OLAP is proposed in this study. OLAP is used to analyze data from the database to help taxonomists in carrying out their research.

1.5 Problem Statement

Globally, there are lots of biodiversity databases that have been developed regarding specific organisms such as fish, ants, reptiles, amphibians, mammals, parasites and plants. Most of these kinds of databases were developed by researchers at educational institutions or through research institutes. Most of the developed parasite databases provide limited accessible information and some of them provide information only on taxonomy data. Very few databases provide detailed information on the species morphological information, location of species discovery and references. In Malaysia, the progress of biodiversity information system development is still underdeveloped and most of the systems are not easily accessible (Napis, Salleh, Itam, & Latiff, 2001). For

example, according to the Malaysian Biological Diversity Clearing House Mechanism (MyCHM) Webpage (<http://www.chm.frim.gov.my/Bio-Diversity-Databases/>) there are several working databases related to flora, fauna and fungi have been developed and uploaded in the website. Nevertheless, there is still lack of developed specific-databases which cover other types of organisms such as database on parasitology as there is no parasite-related database provided here. Most of the information in the databases is available in the form of checklists and is incomplete. A comprehensive parasite database should be developed because it is very useful not only for the studies in parasitology field but also beneficial to the research in other domains of studies.

1.6 Research Objectives

The aim of this study is to develop a parasite-host database using OLAP analysis approach. This study is carried out to develop Parasite Information Network System or known as PINS. PINS is developed as a practical web-based information system regarding parasite-host information which is expected to become a resource centre to assist biodiversity data preservation process as well as serves as information sharing platform.

The objectives of this study are:

- (i) To develop a parasite-host relational database system that provides information on parasites and their respective hosts
- (ii) To perform data analyses using the online analytical processing (OLAP) approach.

- (iii) To serve as a platform of a comprehensive host-parasite information hub in Malaysia which can further expand the study in other related fields such as in medical and human health research.

1.7 Scope

This study involves the development of Parasite Information Network System (PINS). This study focuses on the data preparation process on parasite-host related information, design and implementation of the parasite-host relational database, and design and construction of the on-line analytical processing analysis system.

1.8 Justification of the Study

This thesis sets out as a proof of concept where it shows how a relational database can be built to manage and analyze information on biodiversity, using the host parasites as an example. This thesis also showed that OLAP is a useful approach in providing data analytics capabilities. Data analysis application using online analytical processing (OLAP) approach provides an analysis system where this approach runs the analysis using multidimensional data modeling structure or commonly known as OLAP cubes. The developed OLAP cubes will pre-aggregate and organize data into a hierarchical level structure to allow the analysis of data from a summarized data to a detailed data and vice versa using certain operations. This analytical approach produces a flexible and user friendly system.

PINS project produces a comprehensive parasite-host information system based in Malaysia. Such system enables easy, quick and safe data gathering and information dissemination process to be done. Indirectly, this study will contribute to the progress of information resources in biodiversity domain in Malaysia as well as promote research in relevant fields. PINS system will also be essential as a source of references for teaching, learning and research process.

1.9 Chapter Organization

Chapter 1 describes the details of the scientific data on parasitology that is used as data source in the development of the information system in the study. This chapter also elaborates on the literature review related to the biodiversity information systems and the development of OLAP analysis system. Problem statement, research objectives, scope and justification of the study were described at the end of this chapter.

Chapter 2 is concerning project development methodology used in this study. The explanation on the system development phases and prescribed procedures in the implementation process were described. This chapter also elaborates on the analytical approach and data sources used.

Chapter 3 explains the system requirement and the requirement analysis. This chapter discusses the data used in this study and elaborates more on the system architecture used in the system implementation. This chapter also discusses on the necessary requirements for a prototype system.

Chapter 4 describes the three main steps performed in this study. The system architecture, relational database design, multidimensional cube design and user interface design are described. This chapter also discusses the system implementation and system testing process. The development tools used in this research are mentioned as well.

Chapter 5 is the last chapter in this study. The discussion in this chapter covers all aspects about this whole project including the discussion on system advantages, system limitations and future enhancements.

CHAPTER 2

MATERIALS AND METHODS

2.1 Introduction

This chapter elaborates on the methodology and data source used in this study. The purpose of this chapter is to clarify steps proposed for developing the system to achieve the fundamental objectives.

2.2 Project Development Methodology

In this study, Systems Development Life Cycle (SDLC) methodology is used to accomplish the development of an information system. SDLC methodology is a conceptual model that describes stages involved from the beginning phases of the study until the end of the system development process. In general, SDLC model involves the study of the existing system, identifying system requirements, preparing proposed system design, implementing the system and evaluating or testing the system to determine the effectiveness of the system as shown in Figure 2.1. This study uses the methodology as described in SDLC model, i.e. to do research on the current system in both biology and other fields, identify the particularly system requirements to meet the needs of biological data, design the appropriate system to suit the data used, implement the proposed design as well as perform the testing procedure to the developed system.

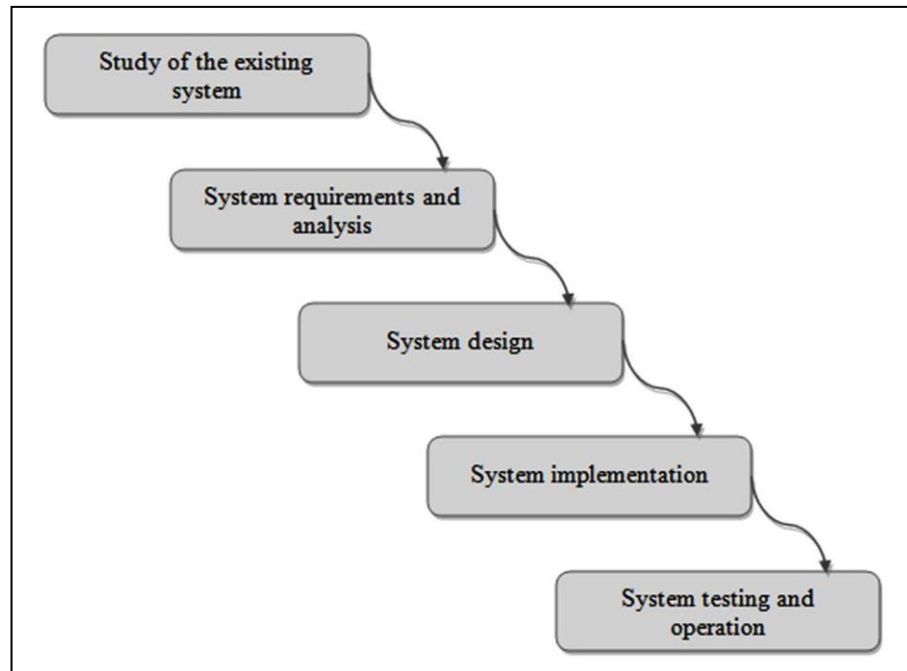


Figure 2.1: The flow of Systems Development Life Cycle (SDLC) methodology.

2.2.1 Development Phase

In order to implement this study, the system development procedures are carried out in five major phases according to the SDLC methodology models described above. The five major phases are the information gathering and literature review phase, identification of the system requirements phase, designing proposed system phase, system implementation phase and finally system testing phase. The explanations of each phase are as follows.

(i) Information gathering and literature review phase

Literature review is the preliminary stage in this study. It is very important to gather information on the OLAP approach applied in any existing biological database systems especially in biodiversity databases. Information gathering regarding previous and ongoing studies is essential to identify the necessary

scope of the study as well as the importance of the study in order to achieve the objectives. The information gathering involves gaining information on the approach used, the advantages and disadvantages of a system through the online articles resources, reference books, research paper presentations, and using the website. This phase is also significant in order to solve the problem statement.

(ii) Identification of the system requirements phase

During this phase, the project requirements analysis and research on the project basis architecture is performed. Based on the findings acquired during the literature review phase, the system requirements should be identified of which indirectly it allows the production of a proposed system solutions design to answer any arising problems. Suggestions for solving the problem are given in the form of logical modules to be studied before the implementation.

(iii) System design phase

System design phase is a very crucial stage where during this phase, the structure of system design is sketched and the identification of the design process is carried out. The system design includes the creation of several logical modules which covers the entire system together with respective description for every designed module.

(iv) System implementation phase

Following the system design phase is the system implementation phase. In this stage, logical modules which have been designed during the previous phase are translated into physical modules structure. The physical structure reflects the

actual design of the system which is then transformed into database structure format.

(v) System testing phase

System testing phase is a process of reviewing the research product. This procedure investigates the functionality of the system. In this study, the system has been tested to check the functional elements and to inspect the level of overall system effectiveness before the system is allowed to be used.

2.2.2 Work Flow

This study was conducted as described by the specific phases mentioned above. The scope of work and total time spent for each phases are also different. SDLC methodology is used as a guide and is changed according to the suitability of the work in this study. Figure 2.2 shows the summary of the three phases conducted throughout the system development process.

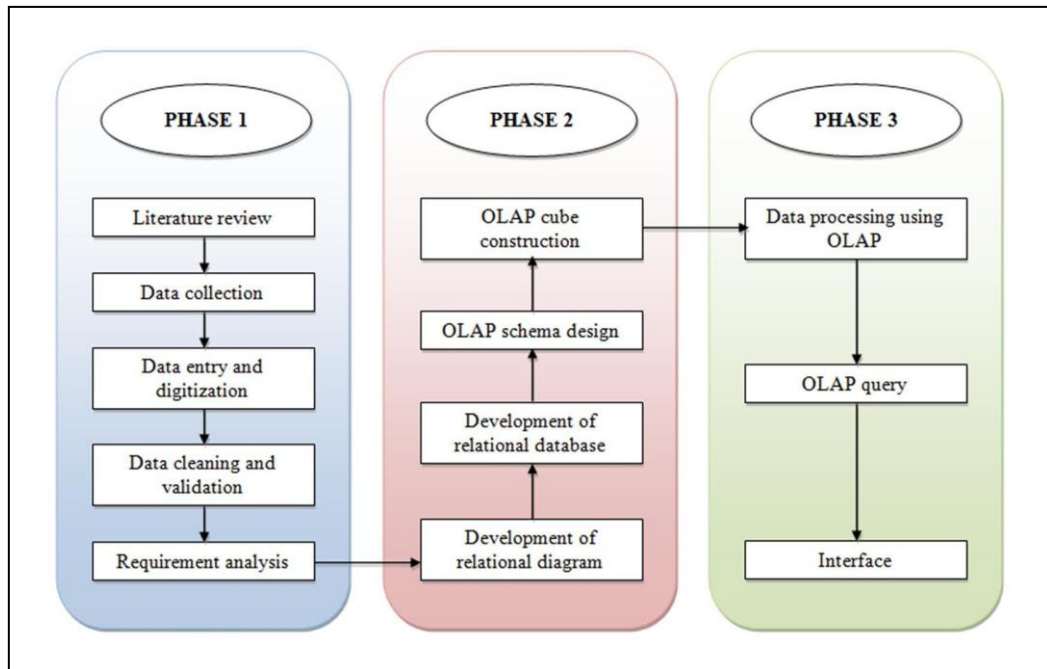


Figure 2.2: Workflow of the system development process.

(i) The collection and preparation of data

This study involves the process of data collection and preparation through two different procedures. The data used in this system is a biological scientific data consisting of a large number and various types of data. Three major steps in preparing the biological data are conducted i.e. the process of searching and collecting data from particular sources, digitizing information obtained and performing data cleaning and validating process. Additionally, the literature review is also conducted to gather related information on the related existing system. Details during literature review phase were obtained from research journal articles, books and information from the website.

(ii) Analysis of system requirements

The system requirements analysis is based on the findings of the literature review and data collection phase. Analysis of scientific data and the structure of

the system development are done. The types of scientific data collected and also the structure of the system development framework are discussed in order to ease the system design process. Besides, functional and non-functional requirements are identified to provide proposed solutions which will be applied when designing the system.

(iii) The system design process

Once all the requirements are identified and prepared, the process of designing the skeleton of the system construction is executed. The design is displayed in an easy to understand diagram accompanied with the required descriptions. In this study, the system design process is carried out by creating the relational database model and designing the OLAP cube schema model which will be used during the system implementation process.

(iv) System implementation process

The completed system design will be used as a basic guideline in building the actual system. The design of the flow of the system will be used in the implementation process of the actual physical form which later can be operated by the users. Conceptually, relational model design is used to develop the parasite-host actual database whereas OLAP cube model is used to create the actual cube in OLAP server.

(v) System testing process

Finally, the system is tested by the user to check and analyze the effectiveness and the functionality of each of the functions available on the system. System

testing process is done where a user is needed to access the developed system and use every function provided in the system. In addition to the system functionality, the purpose of this testing procedure is to ensure that every function is available to deliver information correctly and to get feedback from the user on the suitability of the overall system.

The following Figure 2.3 briefly describes the flow of the system development process.

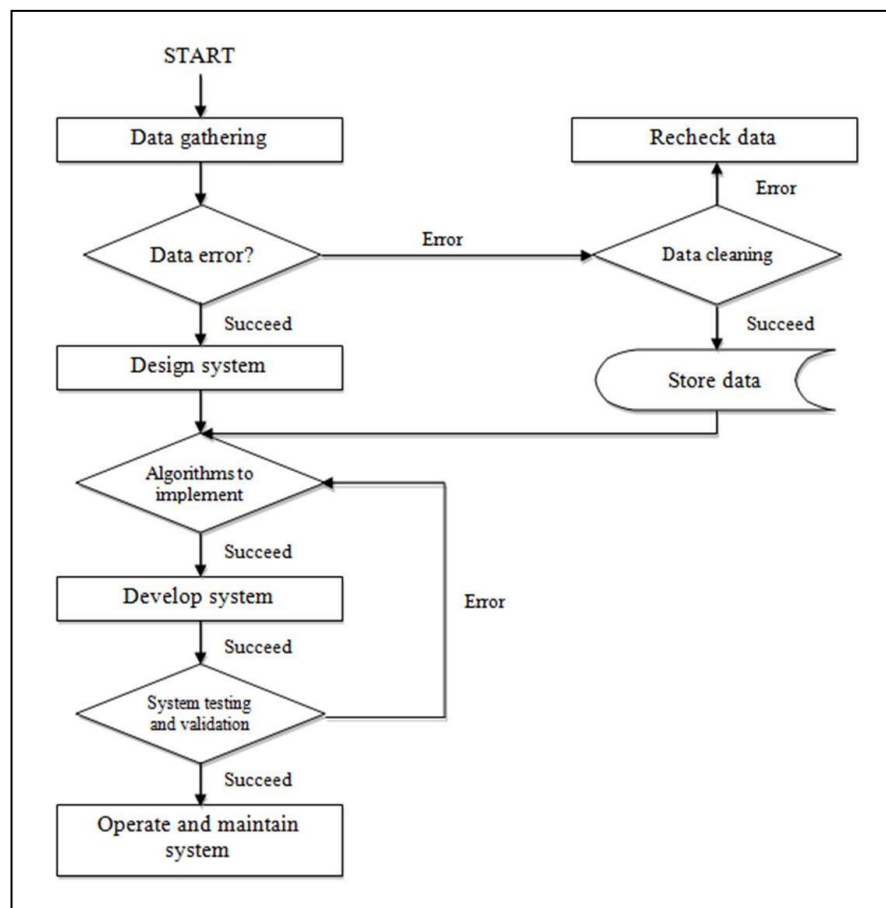


Figure 2.3: Flowchart of the development process.

2.3 OLAP Technology Approach

As mentioned in the previous chapter, OLAP technology approach is one of the approaches used in databases for analysis. This approach is not only applied in business fields, but it has evolved and is being used in the development of biological databases. OLAP technology approach was chosen because of the uniqueness of its multidimensional model which is seen as a successful technique in conducting more efficient and faster data analysis process. This multidimensional cube structure enables data analysis to be performed in different types of operation.

In this study, the OLAP technology is used in data analysis process in the system developed. There are several open-source OLAP engines that can be downloaded, for example The Mondrian-Pentaho OLAP server. Mondrian server connects a relational database with the java-based OLAP front end. The OLAP engine like Mondrian is a java-based server where it is basically working to build the data cube, processing queries and retrieving information. Mondrian supports almost all types of database systems including MySQL. The database system in this study is developed in MySQL server, after which OLAP tools should then be installed and configured to be integrated in the developed system. In short, all required software such as Mondrian OLAP server package, MySQL database server, Apache Tomcat web server and Java Development Kit must be downloaded beforehand. With this preparation, OLAP tools can next be installed and configured.

Mondrian is deployed into a web server by extracting mondrian package and transferring the mondrian.war file into webapps folder in the web server

(Example:/Tomcat 7.0/webapps/mondrian.war/). The mondrian.war file will automatically create a folder named as 'mondrian' in the same webapps folder (Example: Tomcat 7.0/webapps/mondrian/). The mondrian folder contains the required libraries such as JSP files, and jPivot, XSL, XML and TLD configuration files. In order to integrate mondrian in the database system used, the corresponding database driver needs to be downloaded. Mondrian.properties, datasources.xml, web.xml and .jsp files in the WEB-INF folder (Tomcat 7.0/webapps/mondrian/WEB-INF) should be edited to replace the database connection parameters to the correct parameters. For instance, ParasiteHostDb is a MySQL database, and MySQL Java Connector is the correct parameter to be used with the driver class name is *com.mysql.jdbc.Driver* and the connection URL is *jdbc:mysql://localhost:3306/parastehostdb*.

2.4 Data Source

Currently, this study focuses only on the relevant monogenean parasites and their respective hosts data. The data of these monogenean parasites is categorized into four main classifications; parasite information, host information, geographic information and publication information. Parasite information includes full details on taxonomic information and also biological description for each parasite species, whereas host information includes the taxonomic information and details on the specific part on host where the parasite specimen was discovered. Geographic information consists of information on the location where the host species were found, and lastly the publication information describes the complete citation of the publications for each data source.

The taxonomic information of the parasite and host as described above refers to the taxonomic hierarchy data for each parasite and host species. Taxonomic data for both parasite and host are listed based on the class name (group name for the host species), subclass, order, suborder, family, genus and species. For example, the name of class, subclass, order, suborder, family, genus and species for parasite *Sundanonchus foliaceus* is respectively stated as follows: *monogenea*, *polyonchoinea*, *dactylogyridae*, *tetraonchoinea*, *sundanonchidae*, *sundanonchus* and *foliaceus*. Information on the parasites biological description describes in detail the morphological data of the parasite body parts; i.e. the details on the overall body shape and size, haptors, anchors, bars, hooks, copulatory organ, and vaginal system. In addition, the specific site on the host body where the parasite species was found is also noted. For instance, the parasite species was found on the gill of the host body. Geographical information describes the locality data (district, city, country) which denotes the location of the host species discovered. For example, the host for *Sundanonchus foliaceus*, *Channa micropeltes* was found at the Tasik Bera, Pahang, Peninsular Malaysia. Information about the publication provides detailed data related to the resources such as journal articles which are used during information extraction process. The resources information includes the author's name, publication year, article title, journal title, volume and page number as shown in the following example: Tan, W.B. and Lim, L.H.S. (2009). "*Trianchoratus longianchoratus* sp. n. (Monogenea: *Ancyrocephalidae*: *Heteronchocleidinae*) from *Channa lucius* (Osteichthyes: *Channidae*) in Peninsular Malaysia." *Folia Parasitologica* 56(3): 180-184.

All data used are obtained from various sources such as original journals, journals acquired from the internet and also information from the websites. Focusing on the data

collected from the original journals, data digitization process is carried out where the data is extracted into digital format and stored in the form of relational tables in MySQL database. Data cleaning process is then performed to ensure the accuracy of data, no data duplication or data missing happened during digitization process. The parasite-host database contains 861 parasite records derived from 47 publications sources. From this number, 539 parasite species were found to belong to 23 different families with the amount of 273 host species recorded. Each of parasite taxonomic information is included with the complete information on related biological description, location of species discovered, host information and publications information. Parasite-host database is developed to allow the storage of data in order to preserve the information in which most of the resources were stored in the original printed journals. Besides, the developments of the analysis system can be used to mine valuable knowledge and applied for further study.

2.5 Summary

This chapter focused on the materials and methods with explanations on how data was obtained and how this study was conducted. Research methodology is important to ensure that the study is done to achieve the objectives and ensure the implementation process goes on smoothly and systematically. This methodology is a guideline to build the proposed system. This chapter also discusses the workflow throughout this study. The workflow of the research discusses about the data source, online analytical processing technology approach, and the construction of parasite database are done in this study.

CHAPTER 3

SYSTEM REQUIREMENTS AND ANALYSIS

3.1 Introduction

The system requirement and analysis is one of the phases in SDLC methodology that has been described in the previous chapter. This phase plays an important role in conducting the analysis process on some important aspects before developing the system. This requirement analysis is also important to ensure the project development process is able to run smoothly and can be done systematically. System requirement analysis process begins with gathering information on requirements, which can be obtained in various ways from different resources, followed by identifying the importance of the requirements applied during project development, and finally documenting the identified requirements in a specific documentation. The documentation of requirements provides an overview of the interaction between the system and the users in order to achieve the goal of developing the system.

The requirements of PINS system are described in this chapter. In summary, PINS serves as a database system that stores information on parasite-host in Malaysia. To ensure PINS system runs smoothly, the identification of the system requirements in terms of operational facilities and system functionality is very crucial and essential. This

study analyzes the data used, the architectural requirements, and the system functional and non-functional requirements. These requirements are very significant to ensure that the scope of designing the system can be done effectively.

3.2 Data and System Analysis

An analysis of the system was conducted to identify the requirements needed by the system as described above. The purpose of carrying out the analysis on the data was to clarify the process that will be used to gather the scientific data. The architectural requirement was also conducted to identify the appropriate system architecture for developing the system and is explained in the system development process.

3.2.1 Analysis on the Monogenean Data

This study used the data on monogenean to develop a parasite-host database. As mentioned in the previous chapter, a number of relevant databases have been developed such as the Global Mammal Parasite Database (<http://www.mammalparasites.org/>), MonoDb (<http://www.monodb.org>), Hexabotriidae Database (<http://www.ib.usp.br/~mvdomingues/hexa/#>) and Parasite-Host Database by Natural History Museum (<http://www.nhm.ac.uk/>). Most of these databases contain basic information on the parasite and host taxonomy, localities and references. So far, the database observed does not include the biological and diagnostic description. Therefore, this study is looking into developing a database system which provides complete information on taxonomy, host discovery locality, including details on the location of

the discovery of the parasite on the host or referred to as 'site', biological and morphological data, and reference information.

In this study, the aforementioned monogenean parasite information was obtained from published manuscript such as shown in the following Figure 3.1. The information in the reference sources is extracted and digitized into digital form and stored in the database. Data preprocessing is then performed to carry out the data cleaning process. This process is done to ensure the consistency of information between details stored in database and details resided in the original reference sources.



Figure 3.1: Example of the monogenean data resided in the publication (Resource: Lim L.H.S., & Gibson D.I. (2009). A new monogenean genus from an ehippid fish off Peninsular Malaysia. *Systematic Parasitology*, 73, 19)

3.2.2 Analysis on the ROLAP System Architecture

The identification process on the system architecture plays an important role to obtain an appropriate development structure environment for the system construction as well as explanation on matters related to the system development process. In the development of PINS, system structure is very important so that the system can run smoothly while taking into account the appropriateness of the concept of data used. The system development using online analytical processing (OLAP) technology approach can be developed through several different OLAP architecture models; Multidimensional OLAP (MOLAP), Relational OLAP (ROLAP) and Desktop OLAP (DOLAP) as stated in Chapter 1. There are also some systems which are developed using hybrid OLAP architecture called HOLAP. Each model has its own advantages and disadvantages. Model selection in OLAP system development should be suitable with the types of data that will be used to optimize the system efficiency.

In this study, the relational on-line analytical processing (ROLAP) system architecture is used to develop the PINS system. ROLAP is a three-tier client/server architecture (MicroStrategy, n.d.) which is developed based on star schema structure. ROLAP allows the data to be stored in the parasite-host relational database, pre-aggregates data in data cubes, perform OLAP operation in the server and transfer the operation into SQL before proceeding to the presentation layer. Mondrian is selected as ROLAP server used in this study. Figure 3.2 below shows the structure of ROLAP architecture model.

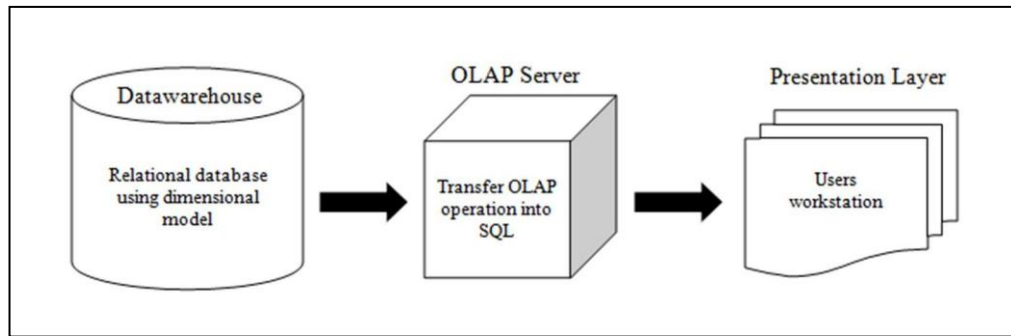


Figure 3.2: The structure of a relational online analytical processing (ROLAP) architecture.

ROLAP model is chosen as it is more suitable for the use in biology. Most biological data are non-numeric of which almost all of them are in textual format plus with other types of data such as an array of sequences as well as images. This requires the right tools to aggregate the data using appropriate functions that can be adjusted to suit the data used (BIOLAP). The technique of storing data in relational databases in ROLAP allows the system to handle large amounts of data. Given the enormous amount of biological data, the system architecture is seen capable of assisting researchers in biology since most of them stored their data in relational databases. Referring to the explanation on OLAP models in chapter 1, ROLAP is chosen instead of MOLAP and HOLAP. As said earlier, MOLAP stores data in multidimensional database and is not suitable for non-numeric data as this study mostly uses textual data. Similarly, the HOLAP model is also not suitable to be used for textual data as this model pre-calculates numeric data and provides slow performance in handling textual data. As compared to the other architectures, ROLAP can work efficiently in handling both numerical and textual data.

3.3 System Requirements

Analysis of system requirements in this study was conducted to determine and analyze the necessary prerequisite of the user and the system before undertaking the development process. Requirements on the system model are analyzed, including the functional and non-functional requirements. Functional requirements describe what needs to be done by identifying the specific tasks and activities while non-functional requirements are listed as the criteria required in an operational system.

3.3.1 User of the System

User requirements were defined through discussions and communications done with the experts in this field to gather the requirements for stakeholders. Through this informal discussion, they indirectly provide information on their needs and requirements that must be provided by the developed system. In addition, the study and observation on the existing systems were also performed to obtain input regarding the design models, procedures and problems encountered during developing the system. Through this study, the requirements regarding design and system development process are identified, including functional requirements and non-functional requirements. The system is tested by a biologist and also a general user.

In this study, the developed system is targeted for the use of parasitologists, taxonomists, biologists, researchers, students and public users. An individual who is responsible on the system administration and management manages the system to ensure the system runs properly. The system is also managed according to the instructions and guidance

from the experts, individuals who have the authority to the data and control the data allowed to be accessed.

3.3.2 Functional Requirements

The functional requirements are described by identifying what are the processes involved and how the system works to get the output when the user generates an input. The functional requirements are also reviewed and are intended to manipulate the database whereas the requirement studies are focused on the interaction between users and functions available on the developed system. Functional requirements describe each function available in the PINS system. The description on the specification of functional requirements for PINS system is described in the following Table 3.1.

Table 3.1: Functional Specification of the PINS System.

Specification	Description
User Login	<p>Outline</p> <p>Users have to register to enable them to use functions provided by the system</p> <p>Processes</p> <ol style="list-style-type: none"> 1. User clicks Login to register 2. Registration form is displayed 3. User selects categories: Admin, Curator or User and fills up the registration form 4. User clicks Submit 5. The system stores the user's information <p>Specific Requirement</p> <p>None</p> <p>Pre-Condition</p> <p>None</p> <p>Extension Process</p> <p>None</p>

Table 3.1, continued.

Specification	Description
Add Record	<p>Outline</p> <p>This function allows users to add new records into the database</p> <p>Processes</p> <ol style="list-style-type: none"> 1. User selects Parasite-HostDb to link to the Manage Database page 2. User selects Add Record 3. Add Record form is displayed 4. User inserts new information 5. User clicks Add button 6. The system will store the information entered by the user into the database <p>Specific Requirement</p> <p>None</p> <p>Pre-Condition</p> <p>Only admin and curators are allowed to add information</p> <p>Extension Process</p> <p>None</p>

Table 3.1, continued.

Specification	Description
Delete Record	<p>Outline</p> <p>This function allows users to delete existing records from the database</p> <p>Processes</p> <ol style="list-style-type: none"> 1. User selects Parasite-HostDb to link to the Manage Database page 2. User selects Delete Record 3. User selects which record is to be deleted 4. User clicks Delete button <p>Specific Requirement</p> <p>None</p> <p>Pre-Condition</p> <p>Only admin and curators are allowed to delete information</p> <p>Extension Process</p> <p>None</p>

Table 3.1, continued.

Specification	Description
Edit Record	<p>Outline</p> <p>This function allows users to edit existing records in the database</p> <p>Processes</p> <ol style="list-style-type: none"> 1. User selects Parasite-HostDb to link to the Manage Database page 2. User selects Edit Record 3. Edit Record form is displayed 4. User inserts new information 5. User clicks Edit button 6. The system will update the information modified by the user in the database <p>Specific Requirement</p> <p>None</p> <p>Pre-Condition</p> <p>Only admin and curators are allowed to edit information</p> <p>Extension Process</p> <p>None</p>

Table 3.1, continued.

Specification	Description
Information Search	<p>Outline</p> <p>This function allows users to search provided information in the database</p> <p>Processes</p> <ol style="list-style-type: none"> 1. User open the main page and selects Search menu 2. User inserts keywords in the search box 3. User clicks Search button 4. The system will perform search process in the database based on the keyword entered by the user 5. Results on related information is displayed <p>Specific Requirement</p> <p>None</p> <p>Pre-Condition</p> <p>None</p> <p>Extension Process</p> <p>None</p>

3.3.3 Non-Functional Requirements

The non-functional requirements were studied from several different points of view and are usually classified according to a certain specific criteria. These criteria play a vital role in operating the system as these non-functional requirements are the additional requirements needed to optimize the PINS system capabilities. In this study, the non-functional requirements should be reviewed prior to implementing the system design and the description encompasses on the functionality and manageability of the system.

The descriptions on the non-functional requirements needed for PINS system are described as follows:

(i) Usability of the system

A good system is a comprehensive system in terms of its information and ease of use. Therefore, PINS system is developed with a bunch of useful information that can be utilized by the user which gives benefits to them. The display of this system is also arranged properly to facilitate user when browsing this database.

(ii) Data Safety

The data collected and mainly used in this database are the researcher's property. Any misuse of this data by any irresponsible parties such as stealing data is an offense to be concerned with. Parasite-host database login system allows only certain accredited users who can access the entire data.

(iii) User friendly interface

A good system uses a simple approach for the user. PINS system built a user-friendly interface where it is provided in a simple and easy-to-understand format. PINS system applied a suitable background, font types and size as well as the proper interface layout structure. This is an important concern so that the system meets the ability of the users without the need for assistance when using the system.

(iv) Data manageability

The data maintenance requirement is needed in any developed system. PINS system provides data managing functions where the addition or deletion of data, updating data, search, and data analysis can be done within this system.

3.4 Summary

This chapter focuses on the analysis of system requirements which listed out the system requirements in terms of architectural, functional, and non-functional requirements. Analysis on system requirements is very important to ensure that research can be done in a good condition and the system can be developed properly. The execution of the requirements specified in this chapter is essential so that the developed system functions properly and meets the user's needs. Results from the study of these requirements will be as a guideline during system implementation process that will be described in the following chapter.

CHAPTER 4

SYSTEM DESIGN, IMPLEMENTATION AND TESTING

4.1 Introduction

Description of system design focuses on the activities involved in the process of designing the system framework. The system design process then translates the framework used for implementing the system into standard diagrams for implementation. The system design includes the compilation of data which is essential to assure the production of a good system. Subsequently, the system is implemented according to the action plan prepared during the system design process. Description of the system implementation focuses on the activities involved for each stage throughout the system development process. In view of that, the system implementation process is seen to have a close relationship with the system design process of which both of these processes are very important procedures in developing a database system. Once the system has been developed, testing process should be proceeded. System testing focuses on checking the interaction between user and the system. Testing was conducted to ensure the ability of each function to perform their tasks in the provided system, including examining the correctness of information accessing flow either from or to the user.

4.2 System Design

System design refers to the activities that focus on the design of the database structure that will be used to store and manage data (Coronel, Morris, & Rob, 2011). In this study, the system design process is described in four parts; overall system architecture design, relational design which is used for data storage purposes, dimensional design that is used for the analysis purposes as a part of the system functionality, and user interface design with the purpose of describing the flows involved in accessing the system. Every explanation of those processes associated in designing system is translated into an easy to understand diagram. The exertion in system design stage is very crucial and important to set up a guideline before producing a complete and functional system.

4.2.1 System Architecture

System architecture describes the overall PINS organization. PINS contain a database named as Parasite-Host. This database stores data related to the parasites and their respective hosts. The data is contributed from the research carried out by the researchers. These data were obtained from various resources such as research articles and research reports. Besides, additional data is also sourced from the Internet. All acquired data is extracted and is then digitized. The digitized data is revised and data correction process will be proceeded to address any erroneous such as data inaccuracies, data loss, or detecting any data errors, for instance spelling mistakes. Parasite-Host database allows users to retrieve information from the data stored. Instead of retrieving information, managing data is also possible where data can be added, removed or edited by authorized user. The process of manipulating the database can be done through a

number of specific functions available from the system. Moreover, the analysis function is also available for users to get more information on a particular required subject. The function of multidimensional analysis provides services on data analysis which is depending on the applied online analytical processing approach. This analysis function allows the construction of multidimensional cube known as OLAP cube. Multidimensional cube starts to process the query from the user and obtains results from the system before displaying the query answer back to the user. Figure 4.1 below shows the summary of the whole organization of PINS system as described above.

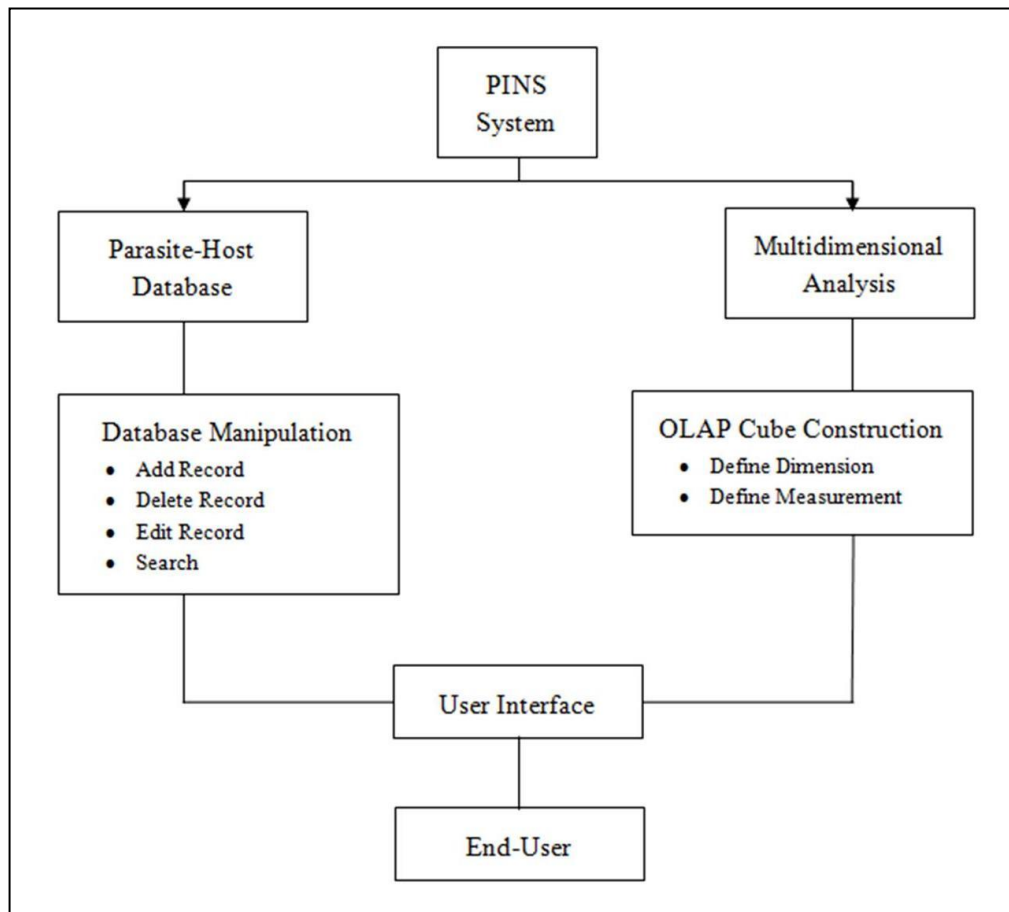


Figure 4.1: The Overall System Architecture for Parasite Information Network System (PINS).

4.2.2 Relational Design

In this study, relational design describes the process of database modeling where it is used for data storage purposes. This project is built using the application of the relational database system where a table is used to store data. Database design is the most critical phase in implementing database because each data element and the relationship between each data should be clearly designed. This is to ensure the smoothness of implementation process that will be developed based on the designs provided. In addition, the use of the correct terms during designing database is also reviewed to ensure that they comply with certain prescribed standards.

The database design in this study is done to correlate several data modules derived from the data preparation process. The data gathered can be grouped according to appropriate data modules such as parasite taxonomy, host taxonomy, biological description, museum collection, parasite ecology, medical information, molecular data and references, as described in Table 4.1 below. These data modules classified the data and grouped them into the same category of information.

Table 4.1: The description on the data modules used.

Modules of Data	Description
Parasite Taxonomy	Consists of parasite taxonomic information such as genus, family, order and class name including the synonym for the species.
Host Taxonomy	Consists of host taxonomic information such as group name, order, family and genus name including current name used for the species.
Biological Description	Describes in detail the characteristics of parasite species, for instance description of the morphological details and diagnosis information obtained through the research.
Museum Collection	Notating ID code for any species in the database that exists and placed in the real museum.
Parasite Ecology	Describes the exact location where the parasite and host specimens have been discovered together with the description of their habitat distribution patterns.
Medical Information	Information related to any disease, symptom and causes of parasite-related diseases.
Molecular Data	Contains information about DNA sequence and genomic location.
References	Hold information on resources materials of the data collected in the database.

In designing the PINS system, the modules mentioned above are used and is structured into an Entity-Relationship (ER) diagram. Entity-Relationship diagram is designed to build a conceptual schema in order to allow us to see how the data are linked as well as to observe the flow of the data theoretically. Conceptually, ER diagram consists of entities, attributes and relationships (Harrington, 2009). Entity is a general subject to a specific type of data or more easily understood as a table with some data in it where the data contained in the entity is referred to as attributes. Each attributes that is represented in the respective entities is associated with a number of relationship methods, such as one to one relationship, one to many, or many to one relationship. The complete ER diagram in this study is shown in Figure 4.2 which has been prepared and will be used in the implementation of the prototype of the actual database.

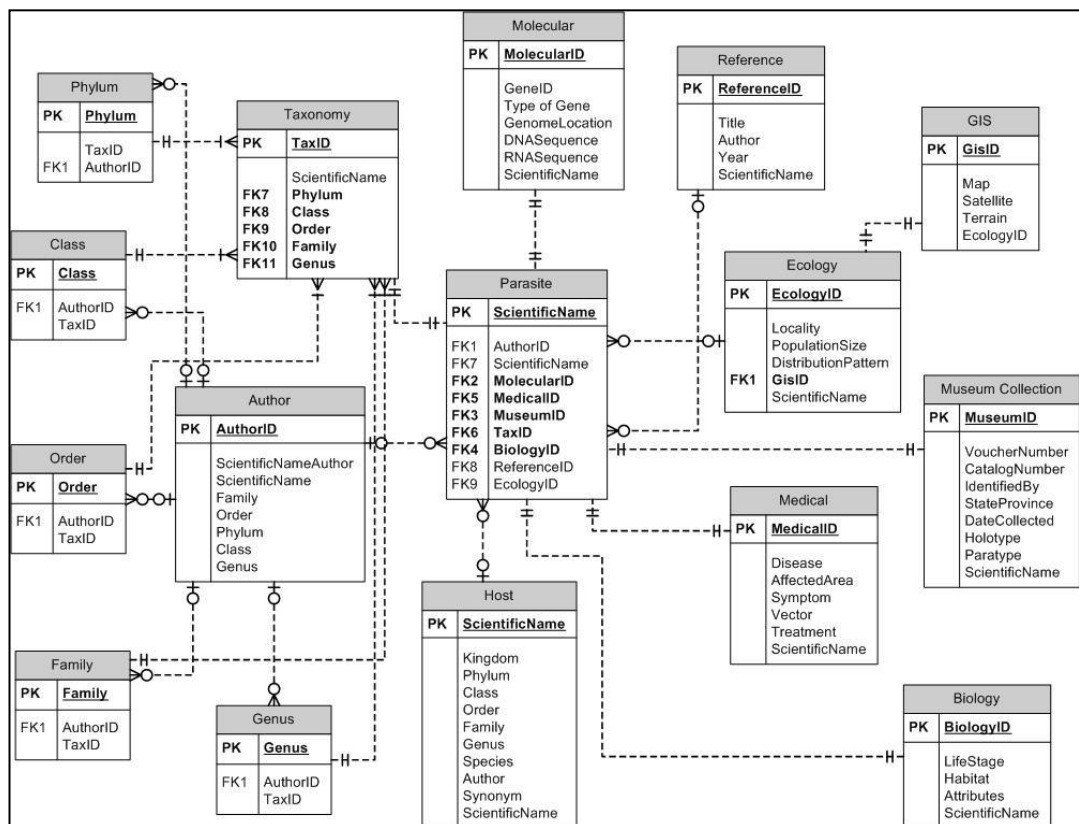


Figure 4.2: Entity-Relationship Diagram used for developing PINS system.

4.2.3 Multidimensional Design

Multidimensional design refers to the process of designing OLAP cube model in a star schema structure which will be used for the analysis part in the PINS system. The multidimensional design for OLAP cube is derived from the relational model. According to Moody & Kortink (2003), deriving dimensional model from a relational model reduces the complexity of the database structure which is easy for end users to understand and write queries against. The analysis process uses the data stored in relational database where the construction of OLAP cubes is done in the ROLAP server.

In this study, cube metadata model is developed into four different fact tables which are connected to five dimensions. These metadata models as shown in Figure 4.3 until Figure 4.6 are derived from the relational structural model and arranged according to star schema structure.

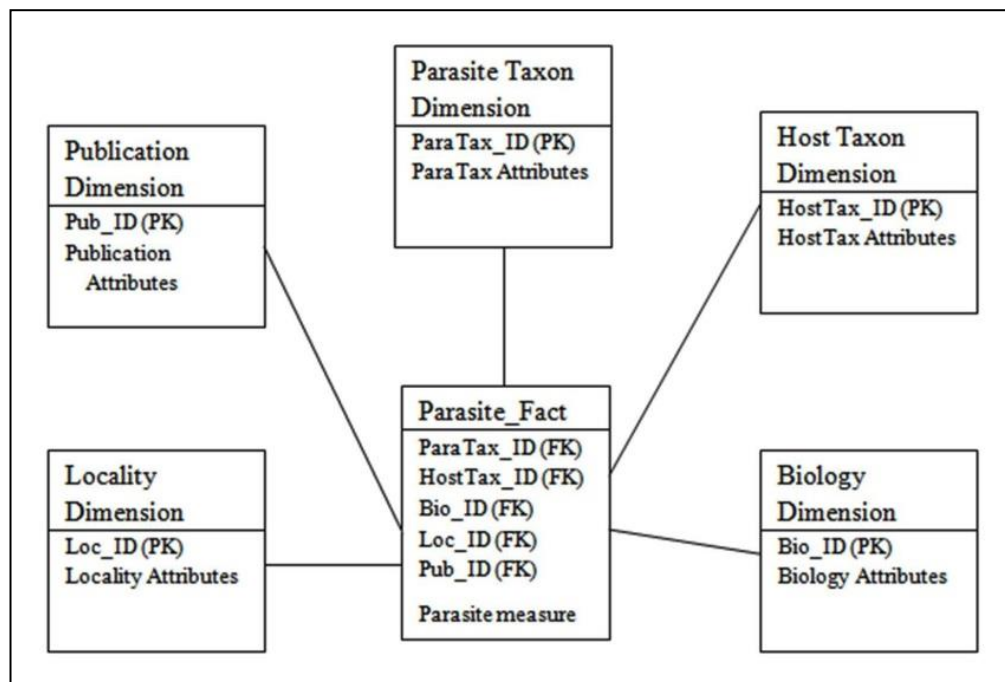
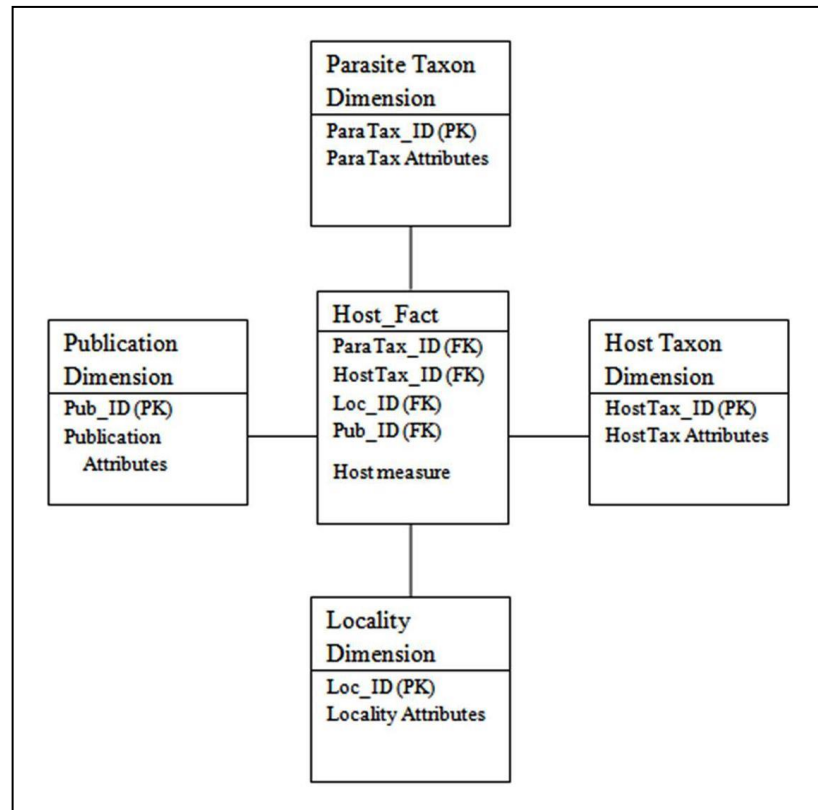
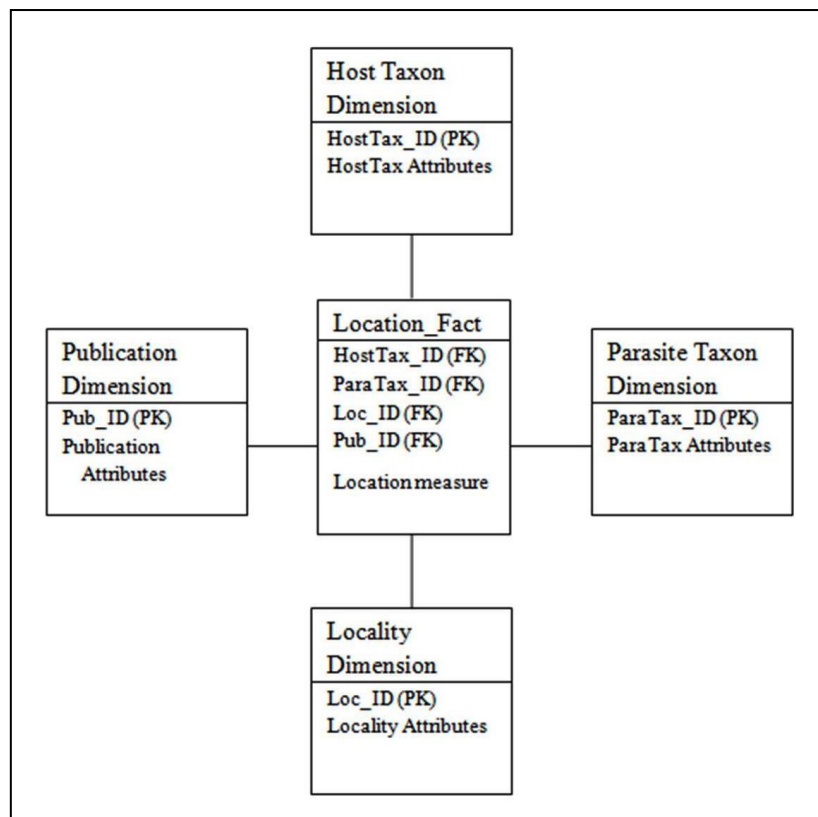
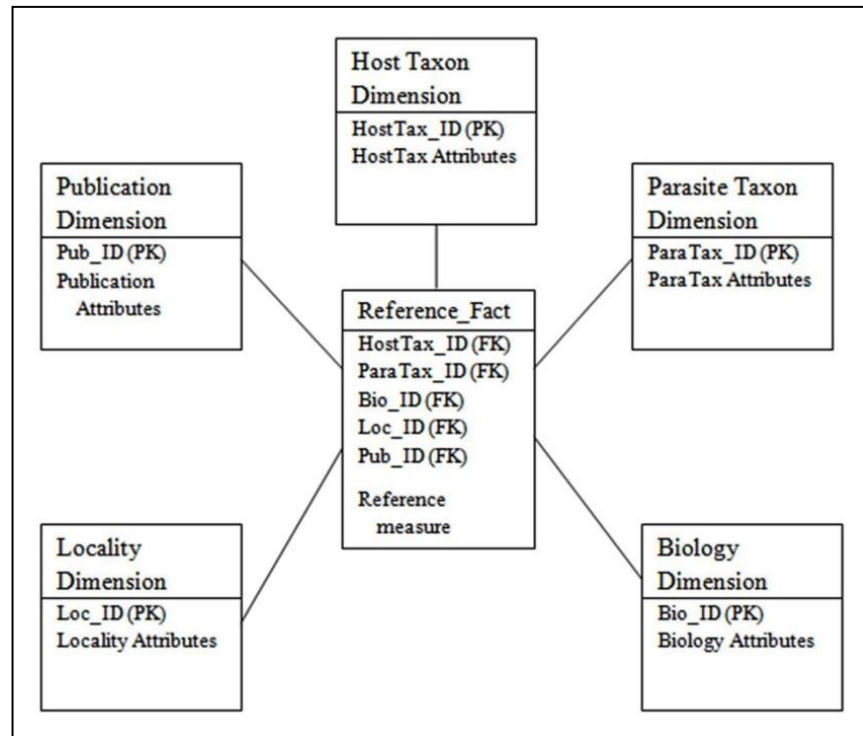


Figure 4.3: *Parasite* cube dimensional model.

Figure 4.4: *Host* cube dimensional model.Figure 4.5: *Locality* cube dimensional model.

Figure 4.6: *Reference* cube dimensional model.

4.2.4 User Interface Design

In this research, user interface design is created with the intention to describe the flow involved in the process of accessing the system. Interface designing task is executed to draw the outline of the visual representation of the system before it is being developed. One of the aims of designing the interface is to attract user to comfortably use the system and to provide a user-friendly system environment. Interface design serves to facilitate the process of developing the real system where every required function is placed in the suitable part on the system. The process of developing the system is already known to be based on the database design, yet it is also developed based on the interface design.

In this study, the user is directed to the home page after logging into the system. From the home page, five links are provided to access to the respective pages. Whenever the user clicks on help or contact button, a new page will appear where the instructions in browsing the system and the contact person details will be displayed respectively. If the user clicks on the parasite-host db and search database menu, a new page will appear. Parasite-Host Database will be linked to the Manage Database page where user can manipulate the data. For instance, at this page, user will be able to add, remove or edit any details of data and save the changes. Search menu will be linked to the Search page where user can enter any scientific name for required species to get information from the system. Browsing PINS menu will be connected directly to the cubes in the Mondrian page where user can do their data analysis. The user interface design structure for PINS system is shown in the following Figure 4.7.

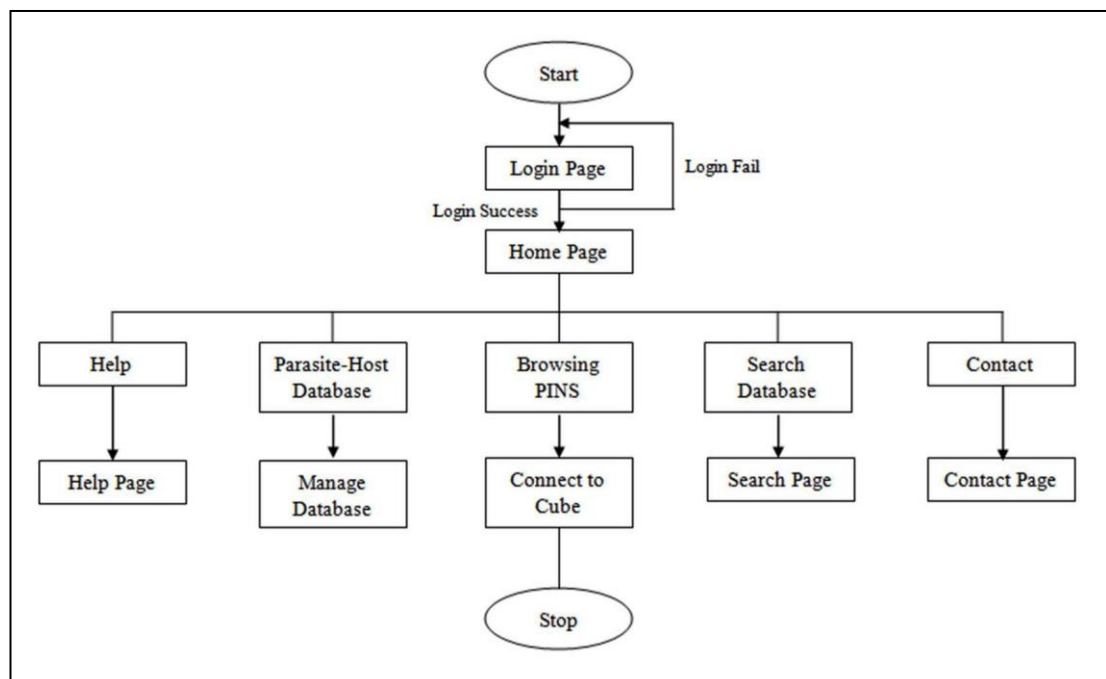


Figure 4.7: User interface design of PINS system.

4.3 Development Tools

In this study, some tools were used to develop Parasite Information Network System (PINS). The tools used consist of several hardware and software. The software used in developing this system is the open-source community edition software which can be downloaded from the websites. Details on the hardware and software used throughout the system development process are shown in Table 4.2 below.

Table 4.2: Development tools used throughout PINS system development.

Development Tools	Details
Hardware Used	<ul style="list-style-type: none"> • Intel(R) Core(TM)2 Duo CPU processor • 1.00 GB free hard drive space • 2.00 GB RAM • 32-bit operating system
Software Used	<ul style="list-style-type: none"> • Windows Vista with Service Pack 2 • MySQL Server 5.5 • MySQL Workbench 5.2 CE • Eclipse Helios IDE • Apache Tomcat 7.0 • Java(TM) SE Development Kit 6 • Pentaho BI Server CE 3.10.0 • Mondrian Analysis Manager • Mondrian Workbench • Google Chrome • Microsoft Office 2007

4.4 System Implementation

System implementation is one of the phases described in the SDLC methodology. System implementation process is important because it is responsible for the realization of the system as well as to meet the required development criteria. In this implementation phase, several developing activities are carried out. The system implementation activities are dependent on the design of the system which has been prepared before.

In this study, the system implementation process is done through four main procedures; data preparation, database development, construction of multidimensional cubes, and finally the development of the user interface. These four procedures are explained in the following sections.

4.4.1 Data Preparation

Data preparation process is the first step done in the phase of implementing this system. Data preparation process is done through three main steps; data collection, data digitization, and data cleaning. Data preparation process especially during cleaning and validating step is done depending on the instruction and guidance from the experts.

(I) Data Collection

Data collection is the first step to be done in data preparation procedure. This step refers to the process where the required data is identified and collected for

digitization process. In this study, data collection process is done to gather information on monogenean parasites and their respective hosts. The information collected is taxonomic information for each parasite and host species, parasite biological and morphological description, complete information on the reference sources and species discovery location. Parasite species discovery depends on discovery of host species. Based on the records stated in the resources, most of parasite and host species were found in the area of Southeast Asia. But there are also a number of discoveries made in some other areas. Majority of host species found were from the fish family. However, there are some host species found from other classes of animals such as frogs, turtles and bat. All the data and information of the parasite and host were obtained from various available sources such as journal articles, books and research reports. These resources were available from the research writings by local experts and researchers in the field of parasitology and ecology as well as some of collaborations with other researchers from abroad. Besides that, the data is also obtained from websites suggested by the experts.

Data collection is a continuous process of collating and compiling the data. The database should update the current records such as add, remove or edit the existing details, and adding new species when new discovery on parasite species has been published by the researchers. To date, the Parasite-Host database has already documented 861 parasite-host records. Statistically, 539 parasite species are recorded in which all of them belong to 23 different families. Among the total of parasite species, 217 of them have completed details on biological and morphological descriptions while descriptions on other parasite species are not

provided in the respective publications. As said above, most of the host species are fish. All these recorded parasite species were discovered from 273 host species.

(II) Data Digitization

Data digitization and organization process is very important in efforts to manage biological data in a systematic way. This process is also very crucial and should be done carefully to minimize or even prevent any errors. The gathering, digitizing and organizing process will be constantly carried out from time to time depending on the available data or resource materials.

All sources collected are either in the form of journals, books or research reports. All those resources were digitized into electronic form and saved using a standard format, i.e. the portable document format (PDF). Then, all of the information from the digitized resources was extracted and transcribed into digital form. Task of ‘data entry’ is carried out to transfer data from printed form to a computerized format. All the data were documented into Microsoft Excel spreadsheet as shown in Figure 4.8. Data from sources were documented and sorted according to the groups of data listed in Table 4.3 below where those data have been identified and data classifications were determined to divide data into certain particular categories.

Table 4.3: Five major categories of data collected and the associated details.

Categories of Data	Details
Parasite Taxonomy	Contains complete taxonomy data based on what is exactly written in the publications.
	<p>Parasite taxonomy data is then classified further into several detailed categories:</p> <ul style="list-style-type: none"> • Class • Class Author • Subclass • Subclass Author • Order • Order Author • Suborder • Suborder Author • Family • Family Author • Genus • Genus Author • Species • Species Author • Synonym • Synonym Author
Parasite Description	Describes the parasite anatomy details and information related to the specimen finding including other information related to the species.

Table 4.3, continued.

Categories of Data	Details
Parasite Description	<p>Detailed data provided on parasite description:</p> <ul style="list-style-type: none"> • Description • Site • Material Studied • Type specimen • Voucher material • Etymology • Diagnosis • Comments • Remarks
Host Taxonomy	<p>Contains taxonomic data of the host of their particular parasite based on information stated in publications. However, most of the host data is not completely given in the collected sources and most of host data needed has been retrieved from FishBase website (URL: www.fishbase.org).</p>
	<p>Same as parasite data, host data is also classified into several categories as shown below:</p> <ul style="list-style-type: none"> • Group • Group Author • Order • Order Author • Family • Family Author • Genus

Table 4.3, continued.

Categories of Data	Details
Host Taxonomy	<ul style="list-style-type: none"> • Genus Author • Species • Species Author • Current name • Current name Author
Locality	Provides information regarding the location of discovered host species.
Publication	Provides complete information about the publication resources.

	J	K	L	M	N	O	P	Q
	Parasite Family	Parasite Family Author	Parasite Genus	Parasite Genus Author	Parasite Species	Parasite Species Author	Parasite Synonym	Parasite Synonym Author
271	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus legendrei	Pariselle, Lim & Lambert, 2-	-	Ad
272	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus levangi	Pariselle, Lim & Lambert, 2-	-	Ad
273	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus lingmoeni	(Gussev & Strelkow, 1960)	Silurodiscoides li (Gussev & Strelkow, -	
274	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus longicirrus	(Tripathi, 1959) Lim, 1996	Silurodiscoides lc (Tripathi, 1959) Gussev	-
275	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus longitubus	(Gussev & Strelkow, 1960)	Silurodiscoides lc (Gussev & Strelkow, -	
276	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus macracanthus	(Achmerow, 1952) Lim, 19f	Silurodiscoides n (Achmerow, 1952) Gt-	
277	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus macracanthus	(Achmerow, 1952) Lim, 19f	Silurodiscoides n (Achmerow, 1952) Gt-	
278	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus macroclethrius	(Lim, 1986) Lim, 1996	Silurodiscoides n Lim, 1986	-
279	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus macroclethrius	Lim, 1986	Silurodiscoides n Lim, 1986	-
280	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus magnicirrus	(Gussev & Strelkow, 1960)	Silurodiscoides n (Gussev & Strelkow, -	
281	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus magnicirrus	(Gussev & Strelkow, 1960)	Silurodiscoides n (Gussev & Strelkow, -	
282	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus magnus	(Bychowsky & Nagbina, 1f)	Silurodiscoides n (Bychowsky & Nagbi-	
283	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mahakamensis	Pariselle, Lim & Lambert, 2-	-	Ad
284	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus malabaricus	(Gussev, 1976) Lim, 1996	Parancylodiscoid (Gussev, 1976) Dube-	
285	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus markovitschi	(Gussev & Gerashev, 1981)	Silurodiscoides n (Gussev & Gerashev, 1-	
286	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mediacanthus	(Achmerow, 1952) Lim, 19f	Silurodiscoides n (Achmerow, 1952) Gt-	
287	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mediacanthus	(Achmerow, 1952) Lim, 19f	Silurodiscoides n (Achmerow, 1952) Gt-	
288	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus megacephala	(Chen, 1988) Lim, 1996	Silurodiscoides n (Chen, 1988	-
289	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus megagripus	Pariselle, Lim & Lambert, 2-	-	Ad
290	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus megagripus	Pariselle, Lim & Lambert, 2-	-	Ad
291	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mehurus	Pariselle, Lim & Lambert, 2-	-	Th
292	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus microhaptorus	(Lim, 1986) Lim, 1996	Silurodiscoides n Lim, 1986	-
293	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus microhaptorus	Lim, 1986	Silurodiscoides n Lim, 1986	-
294	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus moriformis	(Chen, 1988) Lim, 1996	Silurodiscoides n (Chen, 1988	-
295	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus multispiralis	(Jain, 1957) Lim, 1996	Silurodiscoides n (Jain, 1957) Gussev	-
296	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mutabilis	(Gussev & Strelkow, 1960)	Silurodiscoides n (Gussev & Strelkow	-
297	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus mystusi	(Rizvi, 1971) Lim, 1996	Silurodiscoides n (Rizvi, 1971) Gussev	-
298	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus notopteri	-	-	-
299	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus notopterus	(Jain, 1955) Lim, 1996	Silurodiscoides n (Jain, 1955) Lim & Fu-	
300	Ancyrodiscoidae	Gussev, 1961	Thaparocleidus	Jain, 1952	Thaparocleidus obscurus	(Gussev & Strelkow, 1960)	Silurodiscoides o (Gussev & Strelkow, -	

Figure 4.8: Parasite-Host Data Digitization in Microsoft Excel Spreadsheet.

(III) Data Cleaning and Validation

The last step in data preparation process is the data cleaning and validation. Data cleaning and data validation is done to ensure the data accuracy and also to prevent data error. Data cleaning was performed to examine in more detail each of the data that has been keyed in. The purpose of the data cleaning and validation is to check whether there are any problems encountered such as:

- (i) Spelling errors on scientific names, descriptions, locations and references
- (ii) Data inconsistency which means spotting any differences data stated between excel spreadsheet and the original resource materials perhaps due to carelessness during data transcription process
- (iii) Loss of data during data entry process
- (iv) Duplicated records which may happened when extracting same information from different references as this will cause redundant data problem

Data cleaning is an important fundamental task in any scientific research studies and consume a lot of time to finish. In this study, data taken from the original resources were directly recorded into Microsoft Excel spreadsheet. Thus, the percentage of data duplication that could occur is very high. Apart from the errors caused by human mistakes during the entering of data into the system, a few other cases of errors occurred due to the inaccuracy of data. Another problem which occasionally happened in data preparation phase is when the same species data were recorded in different articles. This happens because the particular species was studied by different researchers and in some cases, same

researcher recorded the same species information in more than one article which may cause data redundancy. In addition, there are also cases where errors happened regarding spelling mistakes or misplace of the correct information between two different species stated in the article itself. In such cases, the process of data checking and verification should be done with the help of experts to ensure the data accuracy. Error is possible and indeed inevitable from happening, therefore the process of data cleaning and validating help in addressing this problem.

4.4.2 Database Development

Database development is the second step in the process of implementing this system. Once the data is prepared as described in the previous section, all the data is transferred and stored in the database. MySQL database management system is used for developing parasite-host database. To realize this, the implementation is carried out using the MySQL Workbench, a tool of designing and developing database as shown in Figure 4.9. MySQL server 5.5 and MySQL Workbench 5.2 CE are set up in the workstation. MySQL Workbench creates schema to develop database. Schema 'parasitehostdb' is created and the next step such as entering data into the database is done as shown in Figure 4.10 below.

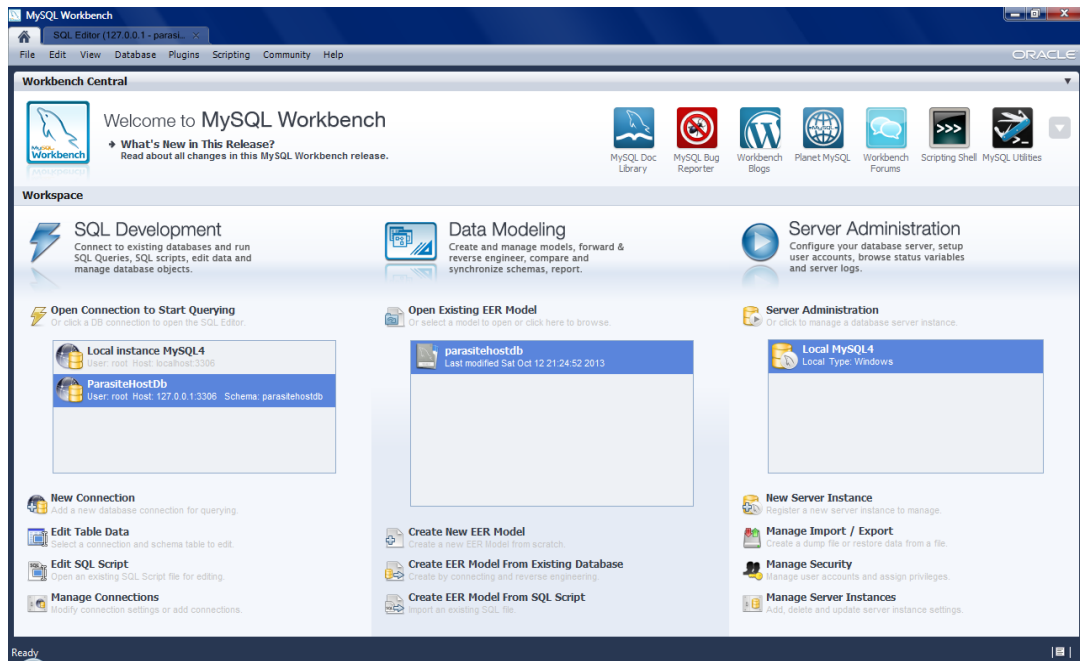


Figure 4.9: MySQL Workbench tool used to develop the parasite-host database

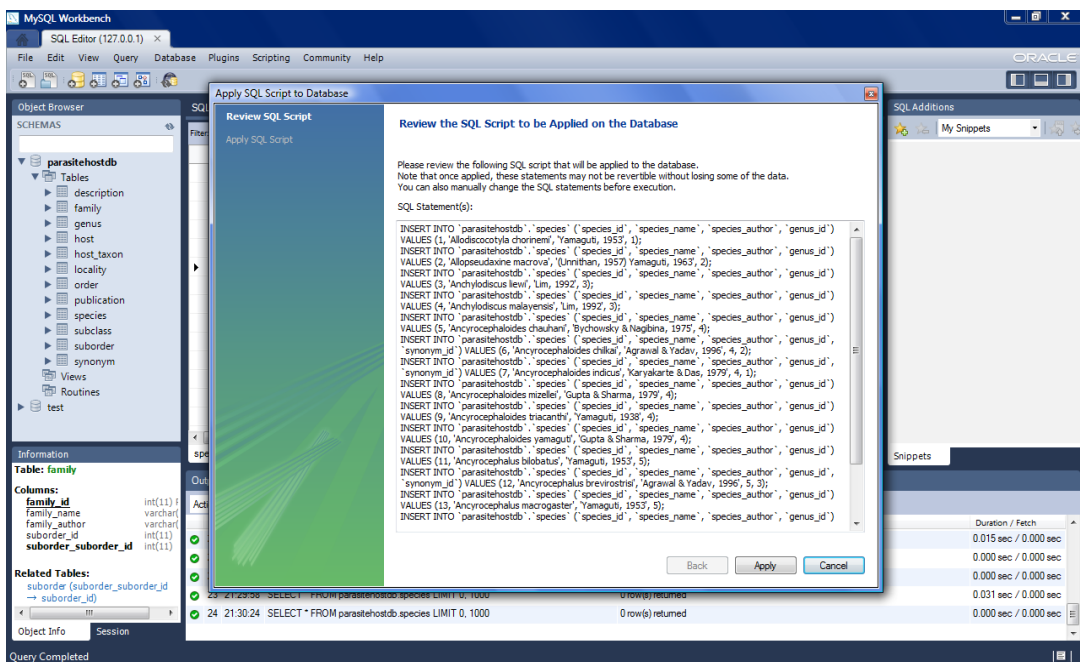


Figure 4.10: Example of SQL script review applied during data insertion process.

In summary, this database contains records of parasite taxonomy, host taxonomic record, biological description for parasite specimens found, records of the location of the specimens' discovery, and records of the data sources. The naming for each attributes is according to the Darwin Core standard data format. The Darwin Core is a standard designed used to facilitate the exchange of information about the geographic occurrence of species and existence of specimens in collections (Wieczorek, 2009). Table 4.4 below shows tables of data which is constructed according to the relationship diagram and briefly described the information provided from these tables. Metadata and data definition is also described in the following Table 4.5 until Table 4:16.

Table 4.4: Table definitions for PINS system.

Table Name	Table Definition
Species	A group of parasites which has been found on a particular animal named host.
Genus	A genus consists of a group of species and formally named following a code of biological nomenclature.
Family	A higher taxonomic category in which a genus has been assigned.
Suborder	A subdivision of an order.
Order	A higher taxonomic category in which a family has been assigned.
Subclass	A subdivision of a class.
Synonym	Other names that have been used for the same species.
Host	An organism which is found as host for a particular parasite species.
HostTaxon	Details on taxonomic classification of a particular host species.
Locality	The exact location in which specimens of a particular taxon have been collected.
Description	Describe the anatomical structure of parasite species in details plus the exact location in which the specimen is found on its particular host (site).
Publication	A citation or reference for data of a particular species

Table 4.5: Data field definition of table Species.

Field Name	Field Definition	PK	FK	Data Type
sp_id	A unique number that has been assigned to every parasite species	/		NUMBER(4)
sp_name	A unique name given to a particular species formed by the combination of genus and species name			CHAR(40)
sp_author	The surname of the person(s) who originally described a species			CHAR(40)
gen_id	A unique number that has been assigned to every parasite genus		/	NUMBER(4)
syn_id	A unique number that has been assigned to every parasite synonym		/	NUMBER(4)
pub_id	A unique number that has been assigned to every publication		/	NUMBER(4)
desc_id	A unique number that has been assigned to every description		/	NUMBER(4)
loc_id	A unique number that has been assigned to every locality		/	NUMBER(4)
host_id	A unique number that has been assigned to every host species		/	NUMBER(4)

Table 4.6: Data field definition of table Genus.

Field Name	Field Definition	PK	FK	Data Type
gen_id	A unique number that has been assigned to every parasite genus	/		NUMBER(4)
gen_name	A unique name given to a particular genus, established following the code of nomenclature			CHAR(40)
gen_author	The surname of the person(s) who originally described a genus			CHAR(40)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)
fam_id	A unique number that has been assigned to every parasite family		/	NUMBER(4)

Table 4.7: Data field definition of table Family.

Field Name	Field Definition	PK	FK	Data Type
fam_id	A unique number that has been assigned to every parasite family	/		NUMBER(4)
fam_name	A unique name given to a particular family, established following the code of nomenclature			CHAR(40)
fam_author	The surname of the person(s) who originally described a family			CHAR(40)
gen_id	A unique number that has been assigned to every parasite genus		/	NUMBER(4)
subord_id	A unique number that has been assigned to every parasite suborder		/	NUMBER(4)

Table 4.8: Data field definition of table Suborder.

Field Name	Field Definition	PK	FK	Data Type
subord_id	A unique number that has been assigned to every parasite suborder	/		NUMBER(4)
subord_name	A unique name given to a particular suborder, established following the code of nomenclature			CHAR(40)
subord_author	The surname of the person(s) who originally described a suborder			CHAR(40)
fam_id	A unique number that has been assigned to every parasite family		/	NUMBER(4)
order_id	A unique number that has been assigned to every parasite order		/	NUMBER(4)

Table 4.9: Data field definition of table Order.

Field Name	Field Definition	PK	FK	Data Type
order_id	A unique number that has been assigned to every parasite order	/		NUMBER(4)
order_name	A unique name given to a particular order, established following the code of nomenclature			CHAR(40)
order_author	The surname of the person(s) who originally described a order			CHAR(40)
subord_id	A unique number that has been assigned to every parasite suborder		/	NUMBER(4)
subcl_id	A unique number that has been assigned to every parasite subclass		/	NUMBER(4)

Table 4.10: Data field definition of table Subclass.

Field Name	Field Definition	PK	FK	Data Type
subcl_id	A unique number that has been assigned to every parasite subclass	/		NUMBER(4)
subcl_name	A unique name given to a particular subclass, established following the code of nomenclature			CHAR(40)
subcl_author	The surname of the person(s) who originally described a subclass			CHAR(40)
order_id	A unique number that has been assigned to every parasite order		/	NUMBER(4)

Table 4.11: Data field definition of table Synonym.

Field Name	Field Definition	PK	FK	Data Type
syn_id	A unique number that has been assigned to every parasite synonym	/		NUMBER(4)
syn_name	A different species name which has been used for the same species			CHAR(40)
syn_author	The surname of the person(s) who originally described a species that has been placed in synonymy			CHAR(40)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)

Table 4.12: Data field definition of table Host.

Field Name	Field Definition	PK	FK	Data Type
host_id	A unique number that has been assigned to every host species	/		NUMBER(4)
host_name	A unique name given to a particular host species			CHAR(40)
host_author	The surname of the person(s) who originally described a host species			CHAR(40)
host_currentname	A different species name which has been used for the same host species			CHAR(40)
host_currentnameauthor	The surname of the person(s) who originally described a host species that has been placed in synonymy			CHAR(40)
taxon_id	A unique number that has been assigned to every host taxon		/	NUMBER(4)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)
loc_id	A unique number that has been assigned to every locality		/	NUMBER(4)
pub_id	A unique number that has been assigned to every publication		/	NUMBER(4)

Table 4.13: Data field definition of table Host Taxon.

Field Name	Field Definition	PK	FK	Data Type
taxon_id	A unique number that has been assigned to every host taxon	/		NUMBER(4)
h_genus	A unique name given to a particular genus of host species			CHAR(40)
h_genusau	The surname of the person(s) who originally described the genus			CHAR(40)
h_family	A unique name given to a particular family of host species			CHAR(40)
h_familyau	The surname of the person(s) who originally described the family			CHAR(40)
h_order	A unique name given to a particular order of host species			CHAR(40)
h_orderau	The surname of the person(s) who originally described the order			CHAR(40)
h_group	A unique name given to a particular group of host species			CHAR(40)
h_groupau	The surname of the person(s) who originally described the group			CHAR(40)
host_id	A unique number that has been assigned to every host species		/	NUMBER(4)

Table 4.14: Data field definition of table Locality.

Field Name	Field Definition	PK	FK	Data Type
loc_id	A unique number that has been assigned to every locality	/		NUMBER(4)
locality	Specific geographic area in which the samples were collected			CHAR(40)
state	State of the country's name in which the samples were found			CHAR(40)
country	Name of the country in which the samples were found			CHAR(40)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)
pub_id	A unique number that has been assigned to every publication		/	NUMBER(4)
host_id	A unique number that has been assigned to every host species		/	NUMBER(4)

Table 4.15: Data field definition of table Description.

Field Name	Field Definition	PK	FK	Data Type
desc_id	A unique number that has been assigned to every description	/		NUMBER(4)
description	The explanation of anatomical description of parasite species			CHAR(40)
site	The spot where the parasite species was found either on or in the respective host			CHAR(40)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)
pub_id	A unique number that has been assigned to every publication		/	NUMBER(4)

Table 4.16: Data field definition of table Publication.

Field Name	Field Definition	PK	FK	Data Type
pub_id	A unique number that has been assigned to every publication	/		NUMBER(4)
title	Name of the publications for the references of relevant data			CHAR(40)
author	Author(s) for particular references			CHAR(40)
year	The year in which the particular reference is published			NUMBER(4)
sp_id	A unique number that has been assigned to every parasite species		/	NUMBER(4)
desc_id	A unique number that has been assigned to every description		/	NUMBER(4)
loc_id	A unique number that has been assigned to every locality		/	NUMBER(4)
host_id	A unique number that has been assigned to every host species		/	NUMBER(4)

4.4.3 OLAP Cube Construction

OLAP cube construction process is the third phase of system implementation carried out in this study. The OLAP cube construction is done based on the dimensional model framework provided as previously described in section 4.2.3. With this multi-dimensional cube, the data can be comprehensively analyzed.

In this study, the system operates using Relational On-Line Analytical Processing (ROLAP) model. The basic ROLAP feature is ROLAP data storage is done in relational database and provides a multi-dimensional environment for analysis processes. This feature is the basic features that distinguish ROLAP from MOLAP as MOLAP data storage is done in multidimensional databases. ROLAP directly access data from a relational database and is designed to allow analysis of data using a multidimensional data model. OLAP cube was built in the ROLAP server where they do data pre-aggregation or summarization based on the determined dimensional models. As we know, biological data is vast in amount and most of the data are in textual descriptions. ROLAP model is used in handling this kind of biological data which is much more practical since ROLAP is considered to be more scalable in handling large data volumes and it is also better at handling non-aggregatable facts as most of the collected data are in textual format.

OLAP cube is the main structure of OLAP system. OLAP cube can be considered similar to a table in relational database system. The structure of OLAP cube is built from the combination of two or more dimensions which contain several levels of data with the measurement. Each OLAP cube has its own schema called star schema. A star

schema is defined by a set of joined tables. This star schema consists of a fact table and dimensional tables. Fact table is usually the center point, i.e. middle table which is linked to other related tables called dimensional tables. A dimensional table contains some information which is named as attributes that may be classified into a number of levels. These levels form a hierarchy which enables parent elements to be aggregated into children elements and vice versa. For example, parasite taxonomy dimension consists of five levels of attributes: scientific_name, genus_name, family_name, order_name and class_name. This structure is shown in Figure 4.14 for better understanding. OLAP cube compiled data in star schema where measures are derived from the records in fact table and dimensions are derived from attributes in dimension tables. All the explanation about the OLAP cube's components defines the structure for OLAP star schema. OLAP star schema is a data model explaining the OLAP cube structure which classifies the attributes into facts and descriptive dimensions. Figure 4.11 shows the main star schema used in this study.

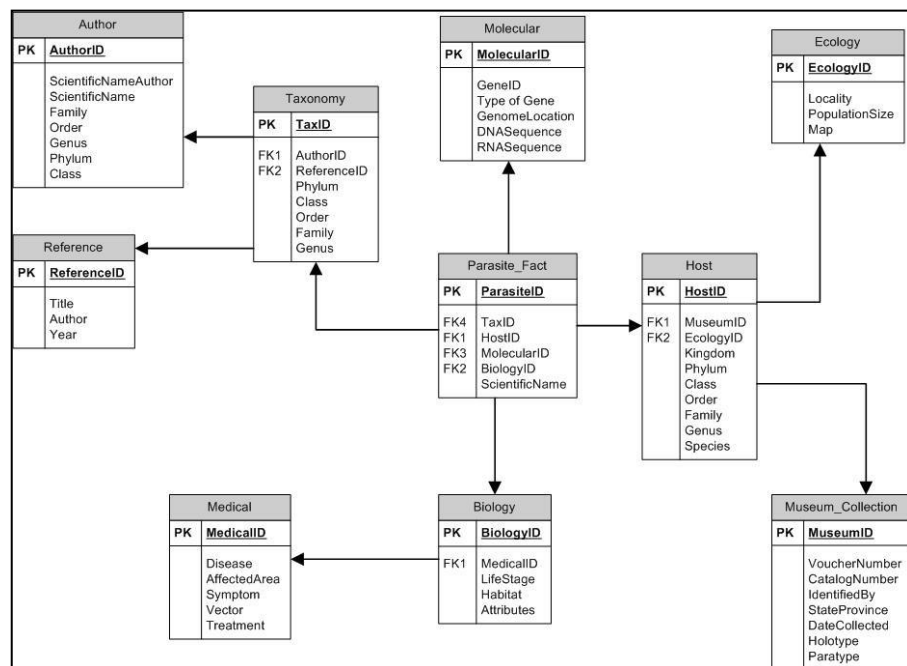


Figure 4.11: Star schema for PINS system.

Mondrian (<http://mondrian.pentaho.com/>) is the analysis manager tool used in developing multidimensional analysis function in PINS system. Mondrian which is also known as Pentaho Analysis Services comes together in the Pentaho server package. Mondrian is an open source relational OLAP engine written in Java. Mondrian supports OLAP for designing, creating and managing the cubes. Basically, when the system receives a query from a user, Mondrian reads data from the relational database and begins to process the multidimensional cubes in the ROLAP server, and finally presents the results in multidimensional formats. Figure 4.12 shows the summary of the query processing flow.

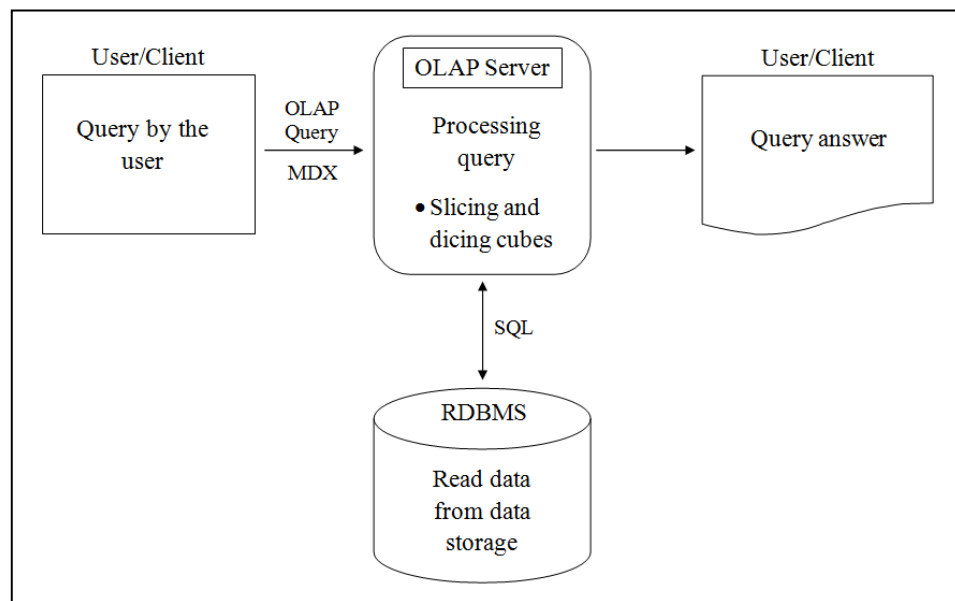


Figure 4.12: The summary of the query processing flow

For a start, the connection between mondrian and the data source, i.e. parasite-host database should be created in advance using MySQL Java Connector as shown in the following Figure 4.13. Next, data from parasite-host database is extracted into Pentaho Mondrian server. OLAP Cube was then built. OLAP Cube allows multidimensional data analysis and preaggregate the data beforehand. For the construction of OLAP

Cube, there are three main elements that should be firstly determined; dimensions, categories and measures. Dimension of OLAP Cube is a category of data that have been classified into certain classes and derived from dimension tables in star schema while measurement is derived from attributes in the fact table. Each data in the dimension consists of several levels mentioned as categories which are arranged into hierarchy structure. The hierarchical structure allows the roll-up and drill-down operations of OLAP Cube to be executed as this hierarchy allows data to be analyzed in summary view or into depth for a detail individual data analysis.

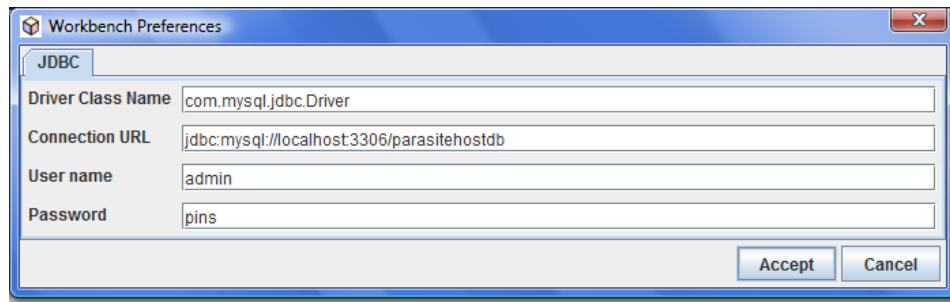


Figure 4.13: Connection between Mondrian server and the parasite-host database

OLAP cube as said earlier was created based on parasite-host star schema. To build the OLAP data structure, dimensions, fact and the members were defined. Dimensions, measure and attributes were defined by selecting certain dimension tables and fact table that need to be used. In this study, four multidimensional cubes were built according to the grouping data based on the star schema models shown in Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6. Each of the cubes contains measure in the respective fact table, consists of more than two dimension tables, with related attributes which are arranged into several levels of hierarchy.



Figure 4.14: Parasite cube fact, dimensions and measurement

The similar procedure is done for all four cubes. For instance, as shown in Figure 4.14, Parasite cube has been created by defining the fact table, measurement, dimension tables, and categories. In this example, the total amount of parasite is defined as the measure and five dimensional tables were selected as dimensions in the cube. Each of ParasiteTaxon, HostTaxon, Locality, Biology and Publication dimensional tables has its own categories which are organized into a hierarchical form that can be further divided to the most basic individual data.

All the listed details as shown above were recorded and saved in XML file format using Mondrian schema workbench tool. The XML schema represents the multidimensional cube in Mondrian. The XML files are ready to be utilized for data analysis purposes. The following Figure 4.15 until Figure 4.18 show the multidimensional cube model in XML schema format in the Mondrian server.

```

<Schema name="PinsSchema">
  <Cube name="Parasite" cache="true" enabled="true">
    <Table name="PARASITE" schema="PS">
    </Table>
    <Dimension type="StandardDimension" foreignKey="paratax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="paratax_id">
        <Table name="PARASITETAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="PARASITE_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="PARASITE_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="hosttax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="hosttax_id">
        <Table name="HOSTTAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="HOST_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="HOST_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="bio_id" name="Description">
      <Hierarchy hasAll="true" primaryKey="bio_id">
        <Table name="DESCRIPTION" schema="PS">
        </Table>
        <Level name="Description" column="DESCRIPTION_DESCRIPTION" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Site" column="DESCRIPTION_SITE" type="String"

```

Figure 4.15: The data schema for *Parasite* cube in XML format.


```

<Schema name="PinsSchema">
  <Cube name="Host" cache="true" enabled="true">
    <Table name="HOST" schema="PS">
    </Table>
    <Dimension type="StandardDimension" foreignKey="hosttax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="hosttax_id">
        <Table name="HOSTTAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="HOST_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="HOST_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="paratax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="paratax_id">
        <Table name="PARASITETAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="PARASITE_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="PARASITE_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="loc_id" name="Location">
      <Hierarchy hasAll="true" primaryKey="loc_id">
        <Table name="LOCATION" schema="PS">
        </Table>
        <Level name="District" column="LOCATION_DISTRICT" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="State" column="LOCATION_STATE" type="String"

```

Figure 4.16: The data schema for *Host* cube in XML format.

```

<Schema name="PinsSchema">
  <Cube name="Location" cache="true" enabled="true">
    <Table name="PARASITE" schema="PS">
    </Table>
    <Dimension type="StandardDimension" foreignKey="paratax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="paratax_id">
        <Table name="PARASITETAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="PARASITE_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="PARASITE_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="hosttax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="hosttax_id">
        <Table name="HOSTTAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="HOST_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="HOST_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="pub_id" name="Publication">
      <Hierarchy hasAll="true" primaryKey="pub_id">
        <Table name="PUBLICATION" schema="PS">
        </Table>
        <Level name="Title" column="PUBLICATION_TITLE" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="PUBLICATION_AUTHOR" type="String"

```

Figure 4.17: The data schema for *Locality* cube in XML format.

```

<Schema name="PinsSchema">
  <Cube name="Reference" cache="true" enabled="true">
    <Table name="PUBLICATION" schema="PS">
    </Table>
    <Dimension type="StandardDimension" foreignKey="paratax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="paratax_id">
        <Table name="PARASITETAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="PARASITE_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="PARASITE_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="hosttax_id" name="ScientificName">
      <Hierarchy hasAll="true" primaryKey="hosttax_id">
        <Table name="HOSTTAXON" schema="PS">
        </Table>
        <Level name="ScientificName" column="HOST_SCIENTIFICNAME" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Author" column="HOST_AUTHOR" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" foreignKey="bio_id" name="Description">
      <Hierarchy hasAll="true" primaryKey="bio_id">
        <Table name="DESCRIPTION" schema="PS">
        </Table>
        <Level name="Description" column="DESCRIPTION_DESCRIPTION" type="String"
          uniqueMembers="true" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Site" column="DESCRIPTION_SITE" type="String"

```

Figure 4.18: The data schema for *Publication* cube in XML format.

Mondrian is used as analysis manager tool for analyzing OLAP cubes data from the database. The Multidimensional Expressions (MDX) language is a query language that is used to extract information from the cubes which operates in the same way as the SQL language used in relational databases. MDX processes the query received from the MDX Query Editor in the mondrian server as shown in Figure 4.19. The required information is extracted from OLAP cubes, i.e. the XML files which were created before to provide the result based on the requested information.

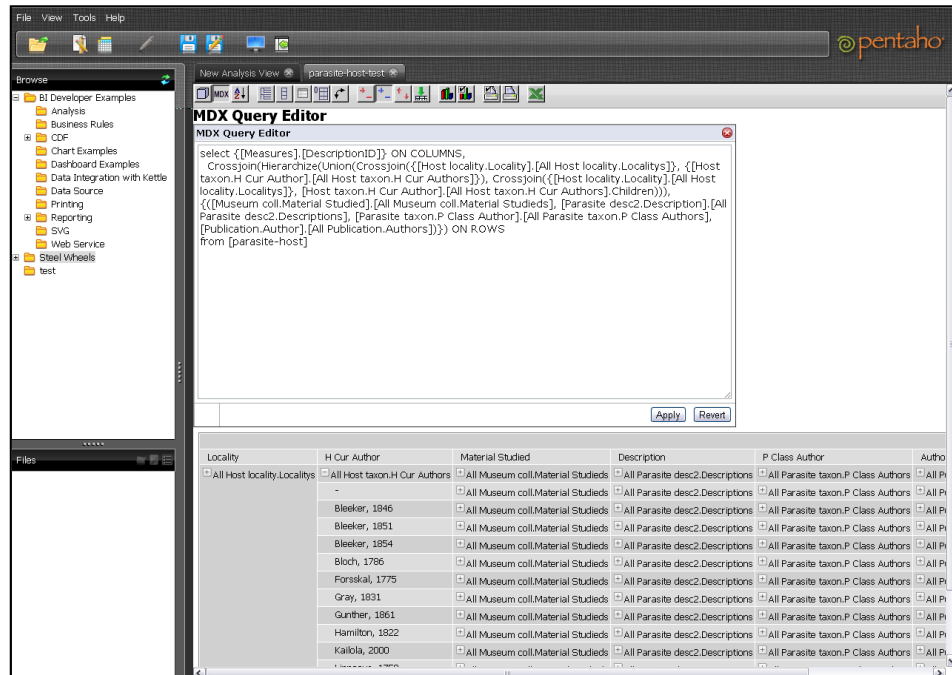


Figure 4.19: Pentaho Mondrian screenshot of MDX query editor.

4.4.4 User Interface

The process of developing user interface is the last process in the implementation phase. The user interface design which has been provided during system design phase is used as a guide in developing the interface for this system. User interface shows the process that occurs when accessing PINS system starting from the home page to the other pages provided.

Figure 4.20 shows the main page of PINS. From the main page, user can click on application menus listed on the left pane to direct to five different pages; Parasite-Host Db, Browsing PINS, Search, Help and Contact.

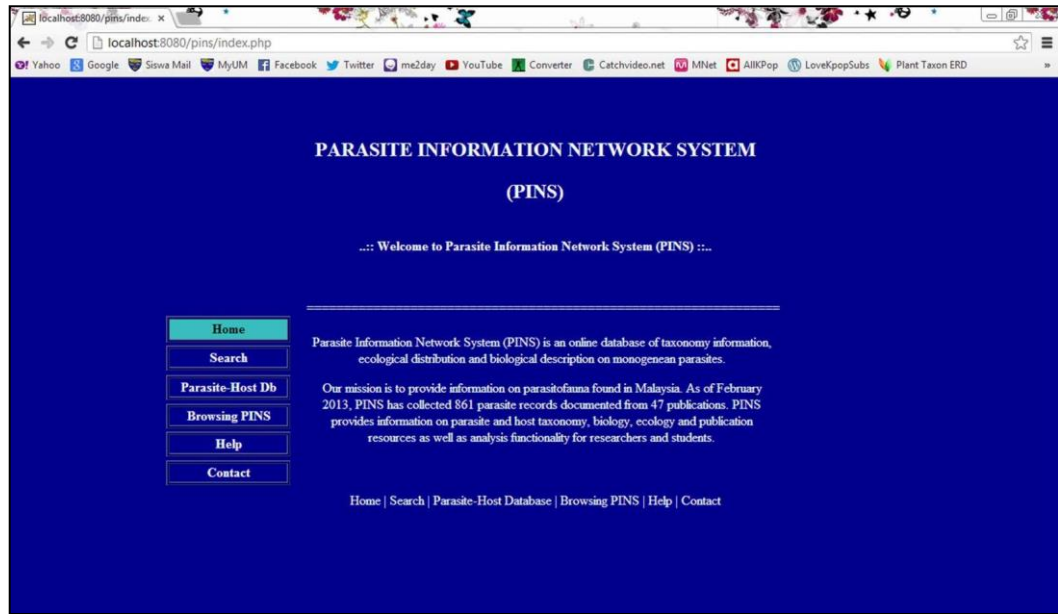


Figure 4.20: Screenshot of PINS Homepage.

Parasite-Host Db is linked to the Manage database page where it is only available to the administrator and authorized curators. Manage database page enables the user to add, delete, update and save their data.

When the user select for Browsing PINS option menu, they will be directed to Mondrian page where real-time analysis can be done. This feature shows the uniqueness of PINS system where PINS give choices for the users to freely browse the cubes themselves and go through the data dynamically. In Browse PINS page, users need to select a data source and the data cube for data analysis as shown in Figure 4.21. This page will link to the Mondrian page after users choose Connect to Cube, where users can start analyzing the data in mondrian as shown in Figure 4.22. Data analysis in mondrian allows data to be viewed in pivot table as well as in the forms of charts.

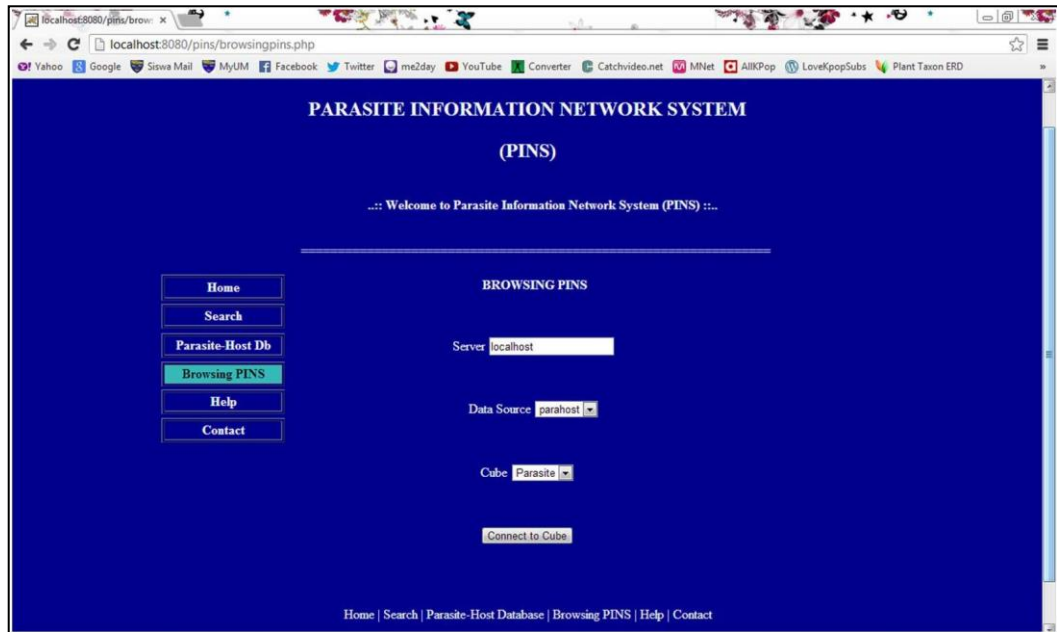


Figure 4.21: Screenshot of Browsing PINS page.

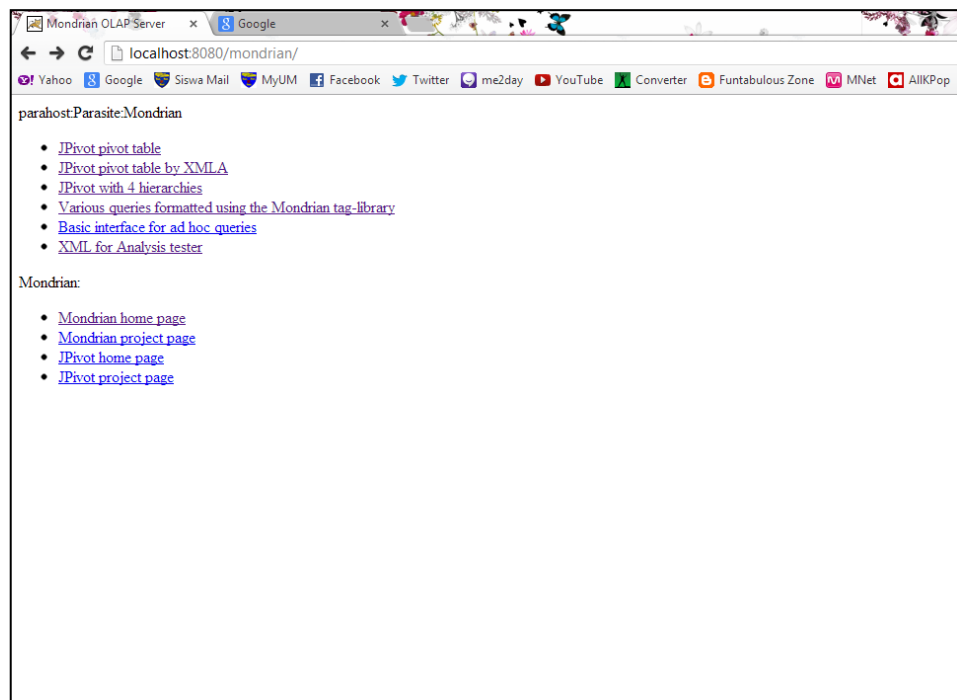


Figure 4.22: Screenshot of Mondrian page connected to PINS system.

Other than that, Search menu linked to the search page that allows the user to search for information as shown in Figure 4.23. Details of the required parasite or host species entered in the search box are delivered to the user.

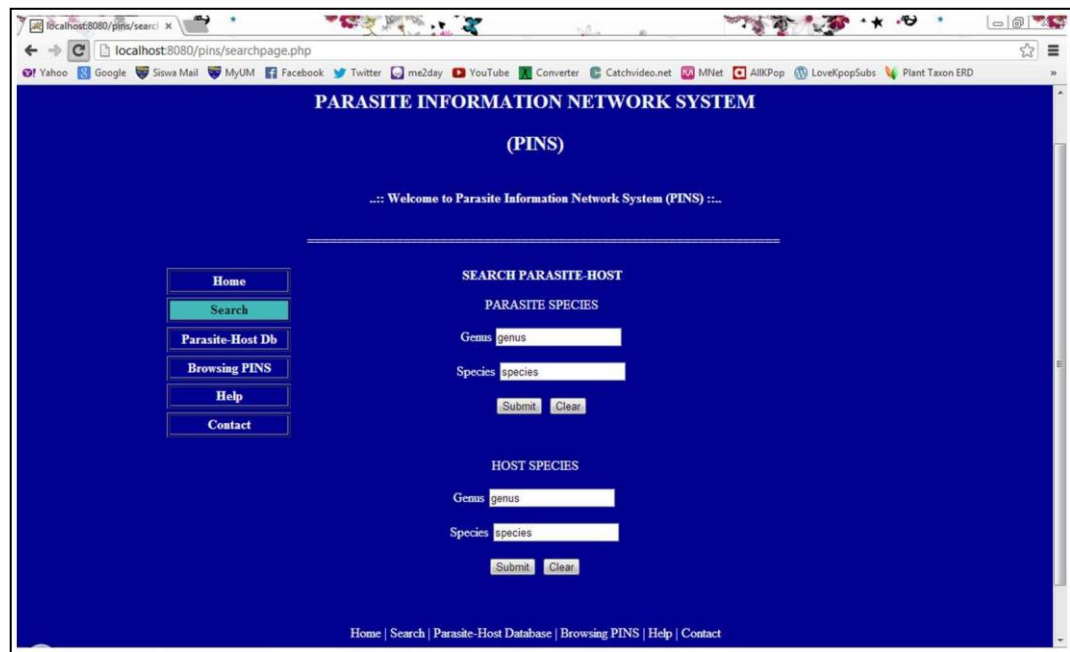


Figure 4.23: Screenshot of Search page.

4.5 System Testing

System testing is the final procedure carried out in the Parasite Information Network System project. In this study, system testing was conducted to check the interaction process between users and the system. The users tested the ability of every function provided in the system to ensure each function is working in carrying out particular tasks.

This system is aimed to be used by the biologists, parasitologists, students and general user. The system is presented with a user-friendly approach where a user does not need

to be an expertise to browse PINS. In this study, a biology student has tested the system functions. The user is required to surf PINS system and trying out every links provided to ensure that every function on the web pages in PINS system are working and can be linked to the next page without any problems.

System testing was conducted through seven test plans which have been prepared, i.e. search species information, add, delete and edit species record, browse PINS, help information, and contact details. Basically, the tester is required to click on each menu displayed on the homepage of this system. Each menu selected should lead the tester to a new linked page. For example, the tester is required to click on the 'Search' menu. Once the 'Search' button is selected, a new search page will be displayed and tester should enter keyword on scientific name of parasite or host species. Tester must click the 'Submit' button and a new page showing information on the species is displayed. To go back to the home page, tester is required to click the Homepage button and continue testing the other menus. System test plan, testing procedures, expected outcome and remarks by the tester is notated in Table 4.17 below.

Table 4.17: Testing notation conducted on the functionality of PINS.

System Test Plan	Testing Procedure	Expected Outcome	Remarks by the User
Search species information	User click Search button, enter species scientific name and click Submit button	Search page and species information page will be displayed	Search page is displayed after selecting Search menu and species information page is displayed after the submit button on the search page is selected. Ok.
Add species record	User click ParasiteHost Db button, click the Add button, enter new data and click Save button	Manage database page and the Add record page will be displayed	Manage database page is displayed after selecting ParasiteHost Db menu and Add record page is displayed when the Add button is selected. Ok.
Delete species record	User click ParasiteHost Db button, click the Delete button, select data, click Remove Data button and click Save button	Manage database page and the Delete record page will be displayed	Manage database page is displayed after selecting ParasiteHost Db menu and Delete record page is displayed when the Delete button is selected. Ok.
Edit species record	User click ParasiteHost Db button, click the Edit button, select data, click Edit Data button and click Save button	Manage database page and the Edit record page will be displayed	Manage database page is displayed after selecting ParasiteHost Db menu and Edit record page is displayed when the Edit button is selected. Ok.

Table 4.17, continued.

System Test Plan	Testing Procedure	Expected Outcome	Remarks by the User
Browse PINS	User click Browse PINS button, select Data Source, select Cube, click Connect to Cube button	Browse page and mondrian page will be displayed	Browse page is displayed after selecting the Browse PINS menu and mondrian page is displayed after the data source and cube were selected. Mondrian allows data to be analyzed through pivot tables and charts. Ok.
Help information	User click Help button	Help page will be displayed	Help page is displayed when the Help button is selected. Ok.
Contact details	User click Contact button	Contact page will be displayed	Contact page is displayed when the Contact button is selected. Ok.

4.6 Summary

This chapter describes system design, system implementation, and system testing. These three aspects are very important in the development of any system. In this study, the system design process is viewed from four different categories; design of the system architecture, relational design, multidimensional design, and finally the user interfaces design. All the processes involved in system design phase should be done properly because they play an important role during the process of system implementation. Next, the chapter describes how the implementation process is executed. Here it classifies the implementation process into four main parts: data preparation, database development, OLAP cube construction and the last one is the user interface development. In sum, PINS system that has been developed is tested through several test plans which aims to detect errors on the data or system functions.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

The collection and management of information has become more advanced and in tandem with the rapid growth of technology. The users of this system consist of researchers and students in related fields. Consequently, PINS is developed to facilitate the collection and information management system, as well as to share information and knowledge via on-line. The knowledge gathered from the system can be used as reference in research and aid the students in their studies.

In summary, this thesis encompasses several aspects; chapter one describes the background of the study, the data and the approach used as well as the literature review; chapter two describes the methodology used; chapter three states the system requirements analysis; and chapter four describes the system design, implementation and testing process. In chapter five, a comprehensive discussion will be done beginning from the initial project planning through to the construction system phase. This chapter discusses on several issues such as the achievement through the development of this system, the importance of the system, the limitations which existed throughout the process when carrying out the project and discusses matters that can be done for the future improvement of the system.

5.2 The Realization of the Research

This study emphasizes the importance of collecting and gathering parasite taxonomic information. Malaysia is the 12th largest in biodiversity in the world (Dhillon, Shuhaimi, Lim, & Sidhu, 2013), and is one of the biodiversity hotspots (Lim & Gibson, 2009) where many species are not yet discovered by the taxonomic researchers. Based on this situation, the collection of parasite information is still relevant and significant in appreciating the researchers efforts in carrying out field studies, as well as providing a system to help them manage their data and preserve the data from any possibility of data loss. The process of data collection, data preparation, and the development of database and analysis system have been described in the previous section. At present, data collection involves only monogenean parasite data. This system has the potential to be a data center that provides information on parasite which could benefit the user. The information obtained from this system is useful for further research in several aspects, academic, economic, ecological, nutritional and treatment of disease, particularly parasitic diseases (Lim & Gibson, 2009). This study is unique because it was not only focused on the development of the parasite-host database, but also on the analysis system development using the on-line analytical processing approach. This is to provide a new dimension in data analysis system for biodiversity information.

5.3 The Importance of the Research

PINS is developed to meet the objectives of this research. There are three main objectives listed at the beginning of this research. The objectives include the development of parasite-host database, performing data analysis using on-line

analytical processing approach, and providing a comprehensive information-hub in Malaysia that will benefit many people, particularly researchers and students.

Data processing plays an important role in maintaining and preserving research resources information in biodiversity field. With the construction of parasite-host database, data which was previously stored in the form of published articles in journals, books and reports documentation, can now be transferred into the database. The storage of data in a database help researchers store and manage their data safely. In addition, the information sharing can be done globally. In short, PINS system is a web-based system allowing the information storage management to be carried out more systematically. The system enables the users to dig out relevant knowledge from the references online and allowing them to access and extract information easily and systematically.

Apart from storing, managing and extracting information, the system is also developed to provide data analysis functions. Analyzing data using OLAP technology approach is capable in helping taxonomists who adopt this system to make their job better. There are several reasons why this approach is highly recommended to be used. Among the most important reason is the basic multidimensional structure as the basis of the analytical process whereby this structure opens up users to the flexible data access procedure. Compared to the two-dimensional model, OLAP model in the form of multidimensional structure allows data to be manipulated from various angles or perspectives. The features available on the basic structure of OLAP cube provide rotation and drill-down functions which allows data to be studied from summary level into more detailed data and vice versa. For instance, the user is able to obtain taxonomic information starting from a class name and eventually explore more into detail order

such as family, genus and scientific name; or from a species discovery location to more general detail such as name of the region and country. This result to a more comprehensive data analysis as OLAP cube structure pre-aggregate the data beforehand for a better query respond. The ability of OLAP system in modeling complex relationships is the key in analytical processing applications where these capabilities cannot be easily done by SQL (Akhramovich, 2011). It is very helpful because the biological data is very complex and interrelating. Another reason than those described above, data analysis using the OLAP approach can be done effectively because the multidimensional model makes the access to information to become more flexible.

The relational on-line analytical processing (ROLAP) system as used in this study is different from the relational database management system (RDBMS). RDBMS stores data in two-dimensional tables. Data is stored in each row and column in the table and the data is static. RDBMS is suitable for data transaction process, but not appropriate for analysis. ROLAP fixes this problem. ROLAP is an OLAP-based system. ROLAP gets data from relational databases and perform data analysis using multidimensional data model. This data model stores data in two or more dimensions. The data is more flexible and is stored in each unit in the cube. Therefore, ROLAP is very suitable to run data analysis process. ROLAP runs metadata-based queries where more complex data analysis can be done using the MultiDimensional eXpression (MDX) language compared to RDBMS since there is some difficulties in translating complex analysis in Structured Query Language (SQL). The analysis of the data in ROLAP allows users to interact with the data directly and performs real-time analysis as well. In addition, ROLAP pre-aggregate data in which this assists in speeding up the process of getting

response compared to RDBMS. Evidenced by facts above, ROLAP is better in handling textual data compared to other OLAP-based systems.

As described in the previous chapters, OLAP technology has been widely applied in the business field. The use of OLAP approach has also been applied in some biological databases (Alkharouf, Jamison, & Matthews, 2005; Markowitz & Topaloglou, 2001; Dzeroski, Hristovski, & Peterlin, 2000; Malmstrom, Nordenfelt, & Malmstrom, 2012). However, in this study, we found that this approach is mostly used in genomic and medical-related fields. There is less biodiversity information system which uses the OLAP applications for developing the system. Therefore, this study was conducted to adopt this approach into the parasite-host database application. Thereupon, the concept of multi-dimensional analysis in OLAP approach has the potential to assist the biodiversity data analysis process.

5.4 Future Works

PINS was created with the intention to develop a database which provides comprehensive information and data analysis functionality to the user. However, this system can be enhanced further. The monogenean parasite data was used in this study. To produce a more valuable system, additional information in terms of a broad spectrum of parasite is important in order to make this system a complete parasite knowledge base. In fact, parasite data related to molecular and medical terms can be added over time. Therefore, the addition of molecular and medical data into PINS system will complement the information provided in this system. On the other hand, parasite is found from their host, that particular host species discovery sites have

been recorded, and the locality data are stored in the database. Thus, the addition of a location map in PINS is seen as a good effort to help user get to know the exact species discovery location. Data privacy and confidentiality is very important in biodiversity research. Therefore, a more holistic data security system needs to be developed in PINS to ensure the database can be accessed without threatening the security of the data stored in it. In addition, more interactive functions can be added into PINS to provide wider access and improve the level of user-friendly display presentation on the pages provided by this system.

5.5 Summary

PINS gather and compile information on monogenean parasite species into an organized system. Almost all the data collected in PINS are found in Malaysia. PINS is designed and developed with the intention to help researchers manage data in order to preserve valuable biodiversity information by documenting the data into database storage structure. Besides, it also assists researchers and students to obtain data about parasites and hosts taxonomy, biology, geography or publication of data resources information provided and use them in their studies.

REFERENCES

- Akhramovich, A. (2011). OLAP Based Analytical Systems: Architecture Overview. Retrieved from <http://www.scientist.by/index.php/data-mining-and-business-intelligence/34-olap-based-analytical-systems-architecture-overview>.
- Alkharouf, N. W., Jamison, D. C., & Matthews, B. F. (2005). Online Analytical Processing (OLAP): A Fast and Effective Data Mining Tool for Gene Expression Databases. *Journal of Biomedicine and Biotechnology*, 2, 181-188.
- AntWeb. Retrieved May 28, 2013 <http://www.antweb.org>.
- Avibase. Retrieved October 5, 2013 <http://avibase.bsc-eoc.org/avibase.jsp>.
- Bihua, X., Jian, L., Haode, L., & Bing, W. (2010, July 16-18). *Research of Data-warehouse-based OLAP Drilling Analysis System*. Paper presented at the International Forum on Information Technology and Applications, Kunming.
- BIOLAP. Retrieved December 21, 2012, from <http://biolap.sourceforge.net/>.
- Bush, A. O., Fernandez, J. C., Esch, G. W., & Seed, J. R. (2001). *Parasitism: The diversity and ecology of animal parasites*. Cambridge, Massachusetts: Cambridge University Press.
- Campbell, N. A., & Reece, J. B. (2002). *Biology* (6th ed.). San Francisco, CA: Pearson Education, Inc.
- Chang, Y.-C., Lee, M.-T., & Lai, K.-C. (2008). Web-based Information Management System for the Long Term Ecological Research Program in Kenting, Taiwan. *Journal of Marine Science and Technology*, 16(3), 174-181.
- Cheng, T. C. (2012). *General Parasitology*. (2nd ed.): Elsevier.

- Coronel, C., Morris, S. A., & Rob, P. (2011). *Database Systems: Design, Implementation, and Management*. (10th ed.): Cengage Learning.
- Cullen, K. E. (2009). *Encyclopedia of Life Science* (Vol. 1): Infobase Publishing.
- Dhillon, S. K., Shuhaimi, N. I., Lim, L. H. S., & Sidhu, A. S. (2013, February 27-28). *Malaysian Parasite Database Infrastructure*. Paper presented at the International Symposium on Biomedical Data Infrastructure (BDI 2013), Kuala Lumpur, Malaysia.
- Dzeroski, S., Hristovski, D., & Peterlin, B. (2000). *Using data mining and OLAP to discover patterns in a database of patients with Y-chromosome deletions*. Paper presented at the American Medical Informatics Association Symposium.
- Eltabakh, M., Ouzzani, M., Aref, W. G., Elmagarmid, A. K., Silva, Y., Arshad, M., Salt, D., & Baxter, I. (2008). *Managing Biological Data using bdbms*. Paper presented at the IEEE 24th International Conference Data Engineering (ICDE 2008), Cancun.
- Froese, R., & Pauly, D. (2013). FishBase. Retrieved May 28, 2013 <http://www.fishbase.org>.
- Giovinazzo, W. A. (2000). *Object-oriented data warehouse design: Building a star schema*. The University of Michigan: Prentice Hall PTR.
- Global Mammal Parasite Database. Retrieved May 27, 2013 <http://www.mammalparasites.org/>.
- Harrington, J. L. (2009). *Relational database design and implementation*. Burlington, USA: Morgan Kaufmann Publishers.
- Hexabotriidae Database. Retrieved October 5, 2013 <http://www.ib.usp.br/~mvdomingues/hexa/#>.
- Hickman, C. P., Roberts, L. S., Larson, A., I'Anson, H., & Eisenhour, D. J. (2006). *Integrated Principles of Zoology*: McGraw-Hill.

- Huyse, T., Audenaert, V., & Volckaert, F. A. M. (2003). Speciation and host–parasite relationships in the parasite genus *Gyrodactylus* (Monogenea, Platyhelminthes) infecting gobies of the genus *Pomatoschistus* (Gobiidae, Teleostei). *International Journal for Parasitology*, 33, 1679–1689.
- INBALUD - Integrating nature and biodiversity and land use data. (2012). Retrieved April 13, 2013, from www.geoville.com/index.php/INBALUD.html.
- Johnson, N. F. (2007). Biodiversity Informatics. *Annual Review of Entomology*, 52, 421-438.
- Kelly, C., Rehm, C., & Barbusinski, L. (2003). What is a DOLAP? Retrieved from <http://www.information-management.com/news/6564-1.html>.
- Lim, L. H. S. (1990). *Silurodiscoides* Gussev, 1961 (Monogenea) from *Pangasius sutchi* Fowler, 1931 (Pangasiidae) cultured in Peninsular Malaysia. *The Raffles Bulletin of Zoology*, 38, 55-63.
- Lim, L. H. S. (1998). Diversity of Monogeneans in Southeast Asia. *International Journal for Parasitology*, 28, 1495-1515.
- Lim, L. H. S., & Gibson, D. I. (2009). A new monogenean genus from an ehippid fish off Peninsular Malaysia. *Systematic Parasitology*, 73, 13-25.
- Lim, L. H. S., & Gibson, D. I. (2009). *Taxonomy, Taxonomists & Biodiversity*. Paper presented at the Biodiversity and Biotechnnology Symposium, Hilton Kuching, Sarawak.
- Malaysian Biological Diversity Clearing House Mechanism (MyCHM). Retrieved October 7, 2013 <http://www.chm.frim.gov.my/Bio-Diversity-Databases/>.
- Mallach, E. G. (2000). *Decision Support and Data Warehouse Systems*. Pennsylvania State University: Irwin/McGraw-Hill.
- Malmstrom, L., Nordenfelt, P., & Malmstrom, J. (2012). Business intelligence strategies enables rapid analysis of quantitative proteomics data. *Journal of Proteome Science & Computational Biology*, 1. doi: <http://dx.doi.org/10.7243/2050-2273-1-5>.

- Markowitz, V. M., & Topaloglou, T. (2001, Nov 4-6). *Applying Data Warehouse Concepts to Gene Expression Data Management*. Paper presented at the 2nd IEEE International Symposium in Bioinformatics and Bioengineering BIBE, Bethesda, MD.
- MicroStrategy. (n.d.). The Case for Relational OLAP. Decision Support Viewpoint White Paper. Retrieved from <http://www.cs.bgu.ac.il/~onap052/uploads/Seminar/Relational%20OLAP%20Microstrategy.pdf>.
- Milego, R. (2012). Biodiversity OLAP Cube in the Framework of the INBALUD project. Retrieved April 13, 2013, from <http://prezi.com/deulp2rsps7l/biodiversity-olap-cube-in-the-framework-of-the-inbalud-project/>.
- MonoDb: A web-host for Monogenea. Retrieved October 5, 2013 <http://www.monodb.org/>.
- Moody, D. L., & Kortink, M. A. R. (2003). From ER Models to Dimensional Models: Bridging the Gap between OLTP and OLAP Design, Part 1. *Business Intelligence Journal*, Summer 2003, 7-24.
- Napis, S., Salleh, K. M., Itam, K., & Latiff, A. (2001). *Biodiversity Databases for Malaysian Flora and Fauna: An Update*. Paper presented at the Internet Workshop 2001, National Institute of Informatics, Tokyo, Japan.
- Natural History Museum Host-Parasite Database. Retrieved May 25, 2013 <http://www.nhm.ac.uk/research-curation/scientific-resources/taxonomy-systematics/host-parasites/database/index.jsp>.
- Niemi, T., Nummenmaa, J., & Thanisch, P. (2001). *Constructing OLAP Cubes Based on Queries*. Paper presented at the 4th ACM International Workshop on Data Warehousing and OLAP, Atlanta, Georgia, USA.
- Palo: Open Source Business Intelligence. Retrieved October 5, 2013, from <http://www.palo.net/>.
- Pendse, N. (2008). What is OLAP? An Analysis of What the Often Misused OLAP Term is Supposed to Mean. Retrieved February 8, 2013, from www.olapreport.com/fasmi.htm.

- Pentaho Mondrian Project. Retrieved March 9, 2012, from <http://mondrian.pentaho.com/>.
- Qian, Z., & Qing, X. (2009). *The Study on Data Warehouse Modelling and OLAP for Highway Management*. Paper presented at the International Conference on Measuring Technology and Mechatronics Automation, Zhangjiajie, Hunan.
- Reed, P., Francis-Floyd, R., & Klinger, R. E. (n.d.). Monogenean Trematodes. Retrieved February 8, 2013, from http://www.simplydiscus.com/library/disease_medications/general_info/monogenean_trematodes.shtml.
- Rohde, K. (2010). Parasitism (An Introduction to Parasitology): How many animal and plant species parasitize hosts, and how? Retrieved from <http://krohde.wordpress.com/article/parasitism-an-introduction-to-xk923bc3gp4-51/>.
- Sinnappah, N. D., Lim, L. H. S., Rohde, K., Tinsley, R., Combes, C., & Verneau, O. (2001). A paedomorphic parasite associated with a neotenic amphibian host: Phylogenetic evidence suggests a revised systematic position for Sphyrnauridae within anuran and turtle Polystomatoineans. *Journal of Molecular Phylogenetics and Evolution*, 18(2), 189-201.
- Uetz, P., & Hosek, J. The Reptile Database. Retrieved May 25, 2013 <http://www.reptile-database.org/>.
- Wieczorek, J. (2009). Historical DarwinCore wiki site. Deprecated. Retrieved June 17, 2013, from <http://wiki.tdwg.org/DarwinCore>.
- Yoo, M. C. (n.d.). MOLAP: Multidimensional OLAP Retrieved May 3, 2013, from http://www.ischool.drexel.edu/faculty/song/courses/info607/tutorial_OLAP/MOLAP_sub.htm.