A Model Of Biodiversity Taxonomy Database (iTaxo) For Managing And Analysing Kenyir Landscape Biodiversity Data

Salwana Mohamad @ Asmara¹, Mustafa Man², Mohd Tajuddin Abdullah³

¹Faculty of Computing, Universiti Malaysia Pahang, Pahang, Malaysia
²School of Informatics and Applied Mathematics Universiti Malaysia Terengganu, Terengganu, Malaysia
³Institute of Tropical Biodiversity and Sustainable Development, Universiti Malaysia Terengganu, Terengganu, Malaysia

Email: salwanamohamad@ump.edu.my

Abstract. Biodiversity data management and analysis is a complex task that must be handled by biodiversity researchers. The large amount of data and complicated interaction of the data are the main problem in the biodiversity field. Data accessibility also becomes the main issue to create a data sharing environment among other biodiversity researchers. This is important element since the data is related to each other. A standardized form of data model for biodiversity data is not well implemented in real world. This study discussed about the database model for biodiversity taxonomy data (iTaxo) in order to manage the biodiversity data. We implement the Kenyir landscape which is located at the state of Terengganu in Malaysia as our case study data since the landscape is rich with biodiversity species and ecosystem. As a result, the model can be applied to manage, analyse and evaluate the biodiversity status at the landscape for conservation effort.

1. Introduction
Biodiversity is a composition of species and genetic variability of plants, animals, and microorganisms [1], [2]. The composition includes entities identification at the level of ecosystem, species and genetic diversity. The interactions between them are known as ecosystems and ecological processes. This means that biodiversity data can be assessed at three different levels which are ecosystem, species and genetic diversity. The data of biodiversity can be categorised into three groups which are taxonomy, geo-spatial and temporal data [3]. The effort to analyse the biodiversity in the context of ecosystem and functional aspects is the main challenge of this study area [4]. In order to perform the biodiversity analysis, the diversity of data must be mapped appropriately by the database system. The data should be analysed and represented in historiographical form in order to contribute the valuable information and understanding of policy making in biodiversity [5].

There are two main factors that lead to difficulty for researchers in biodiversity study to manage the biodiversity data. The factors are the complicated relationship of the biodiversity data and large amounts of the data. The complicated relationship in biodiversity data is caused by the lack of standardised and acceptable relationship of names, taxonomic concept, clades and globally unique identifiers (GUIDs) to
define and build the relationships of these entities [6]. In terms of large amount of data, Porter [7] suggests the database architecture for taxonomic dataset to be designed with practicality for ‘wide’ database instead of ‘deep’ database. This is because of taxonomic dataset contain many different kinds of data and observation which is typically do not have tools for analysis and the tools function mostly on metadata content.

Besides the difficulty to manage the biodiversity data, the accessibility of the data also become the main issue in biodiversity research. The data accessibility effected the data sharing environment and decision making [8]. Data sharing is the ability to share the same data resource to others and reuse the data for analysis [9]. By sharing the data, every researcher should aware what data that have been discovered by other researchers in order to enable the data comparison and prevent the redundant research work. The data storage for the sharing mechanism must be persistent and sustainable for long term use.

The standard of data model for the storage and management of the biodiversity data is hard to find as it is not well implemented as a practical database [5]. This is because of the incompatibility between the logical view of data model design and the application view from the real scenario of end user. The complexity of the logical design has led to the impracticality of database model for the biodiversity research [10].

This work describes about the database model of iTaxo with reference to a case study in the Kenyir landscape. The landscape situated at the state of Terengganu in Malaysia. Tasik Kenyir covered by more than 340 islands spread out in water catchments area of 38000 hectares. Therefore, Tasik Kenyir is rich with biodiversity sources that cover the terrestrial and aquatic area. The objective of this database model is to create a platform of an integrated directory for all known species that will be updated by all biodiversity researchers especially in Tasik Kenyir landscape. While other biological taxonomy databases have focused their content on collecting and store the data, this proposed database model aims to emphasis on providing the analysis of biodiversity and geospatial data. The geospatial data management of the biodiversity data by implementing in-field data collection may contribute to the species distribution mapping, data analysis and modelling. It may become as a resource for studies on biodiversity that will contribute to the decision making in conservation effort.

2. Related Works

There are many projects that applied methods to manage and analyse biodiversity data worldwide. Malaysia Biological Diversity Database (MyCHM) [11] that has been developed by Forest Research Institute Malaysia (FRIM) is a mechanism for exchanging and integrating information on biodiversity. In 2015, MyCHM has been improved by Malaysia Biodiversity Information System (MyBIS) that purposes to be a centralized repository for biodiversity information in Malaysia. Various agencies from the area of environmental, forestry, wildlife and public institutions are the data provider for this system. However, MyBIS does not specify the class of species in the analysis of the status of biodiversity. In analysis map menu, MyBIS has provided the visual map of protected area such as forest reserves and sanctuaries in Malaysia but it does not emphasized the analysis of geospatial data of found species.

Global Biodiversity Information Facility (GBIF) [12] is an international open data infrastructure that allows the public to access data about all types of life on Earth. It provides a centralized access to an enormous number of biodiversity data. The project is also assisting institutions to publish data according to common standards. It also provides a map Application Programming Interface (API). The mapping API is a web map tile service to visualize GBIF data on interactive maps and overlay data from other sources in a simple way.

Integrated Taxonomic Information System (ITIS) [13] is an open access database with information on species names. Every species is described by hierarchical classification with standardized nomenclature. It contains documented taxonomic information of flora and fauna from both aquatic and terrestrial habitats. The system is merely functioning to store and manage the biodiversity data that will allow biological information sharing among researchers and related agencies. Taxonomic Database Working Group (TDWG) [14] are efforts to establish standards and infrastructure for integration and interoperability of data from biological collections. It is also an option of biodiversity applications deals with the management of taxonomic information and geographic distribution of species.
A database system for United Kingdom (UK) biodiversity data has been developed as National Biodiversity Network's Gateway (NBN) [15]. The system is a platform for researchers and public to learn and understand the UK biodiversity data by allowing users to record, share, explore and use the data. The accuracy of species identification is important to maintain the quality of the biological data. Thus, NBN has carried out a verification and validation of the data by expert volunteer verifiers. Users also can explore distributions of species records, whole datasets, protected sites and habitats using the interactive map tool. Despite the complete information that is available in the system, it is lack of statistical analysis for species assessment and the trend of the biodiversity status.

Most of the projects are objectively developed to collect, store and manage biodiversity data. The data standards have been applied to enable the accessibility and distribution of information through the web. iTaxo may have the same objective of assisting biodiversity researchers in managing their data about the found species. However, its main contribution is focusing on Kenyir landscape by detailing the hierarchy of the species taxonomy and integrating several analysis and visualization element in the database. Also, we allowed different access of users depend on the user roles that can be applied for Kenyir landscape researchers, student and public.

3. Database Model for iTaxo
This data model aims to describe the representation of the biodiversity data. The database model is designed by considering the taxonomic classification as in Figure 1.

![Figure 1. Biological taxonomy](image)

The hierarchical biology taxonomy presented within this research is the basis for the relational database model. Biological taxonomy is the science of classification according to a pre-determined system of the biodiversity [16]. It includes the taxonomic group of Domain, Kingdom, Phylum, Class, Order, Family, Genus and Species. A taxon or a group of organisms are given name and a rank of taxonomy hierarchy. With the science of taxonomy, taxonomist will defines what belong or does not belong to a taxonomic group. In biological taxonomy, informatics play an important role to develop databases as the repositories for biological information [17]. This effort can integrates the numerous of biological data that is distributed across the various disciplines of biology.

3.1. Data Structure of iTaxo
The data structure of database model for iTaxo is defined by applying the notation of set theory which may contain repeating elements. It is a basic representation of data structure to develop list theoretic relational databases and operations on them. The data structure contains twelve sets of data entities which are Species, Genus, Family, Order, Class, Phylum, Kingdom, Domain, Wildlife_Status, Type_of_Forest, IUCN_Status and Location. The sets and the elements are notated in Table 1 and Table 2:

| Entities          | Set Notation |
|-------------------|--------------|
| Species           | A            |
| Genus             | B            |
| Family            | C            |
| Order             | D            |
| Class             | E            |
| Phylum            | F            |
| Kingdom           | G            |
| Domain            | H            |
| Wildlife_Status   | I            |
| Type_of_Forest    | J            |
| IUCN_Status       | K            |
| Location          | L            |

**Table 1.** Sets notation

| Set Elements      | Notation |
|-------------------|----------|
| ID_Species        | $a_1$    |
| Name_Species      | $a_2$    |
| references        | $a_3$    |
| notes             | $a_4$    |
| photo             | $a_5$    |
| ID_Genus          | $b_1$    |
| Name_Genus        | $b_2$    |
| ID_Family         | $c_1$    |
| Name_Family       | $c_2$    |
| ID_Order          | $d_1$    |
| Name_Order        | $d_2$    |
| ID_Class          | $e_1$    |
| Name_Class        | $e_2$    |
| ID_Phylum         | $f_1$    |
| Name_Phylum       | $f_2$    |
| ID_Kingdom        | $g_1$    |
| Name_Kingdom      | $g_2$    |
| ID_Domain         | $h_1$    |
| Name_Domain       | $h_2$    |
| ID_Location       | $i_1$    |
| Location          | $i_2$    |
| GPS_Coordinate    | $i_3$    |
| Year              | $i_4$    |
| ID_Wildlife       | $f_1$    |
The intersection and union among the sets of data entities represent the relationship of the database model. Every set share some of the same elements to relate each other is denoted as:

i. \( A \cap B = \{ b_1 \} \)
ii. \( B \cap C = \{ c_1 \} \)
iii. \( C \cap D = \{ d_1 \} \)
iv. \( D \cap E = \{ e_1 \} \)
v. \( E \cap F = \{ f_1 \} \)
vi. \( G \cap H = \{ h_1 \} \)
vii. \( I \cap A = \{ j_1 \} \)
viii. \( J \cap A = \{ k_1 \} \)
ix. \( K \cap A = \{ l_1 \} \)
x. \( L \cap A = \{ i_1 \} \)

The set of Species share the same elements of ID_Genus, ID_Type, ID_Wildlife, ID_IUCN, ID_Location from sets of Genus, Type of Forest, Wildlife Status, IUCN Status and Location is denoted as:

\[ A \cup B \cup L \cup I \cup J \cup K = \{ b_1, i_1, j_1, k_1, l_1 \} \]

Every shared element is an identifier of the values from other set or entities that will describe the overall elements or attributes of the related set.

3.2. Entity-relationship diagram of iTaxo

The entities and relationships represented in the taxonomy must be evaluated to design the database model. Figure 2 shows the entity-relationship diagram of the taxonomy database.
The main entities are defined based on the biological taxonomy hierarchy which also create the main data tables which are Domain, Kingdom, Phylum, Class, Order, Family, Genus and Species. Entities of Wildlife_Status, Type_of_Forest and IUCN_Status are defined to describe the species status and the habitat. An entity of Location is created to link the geospatial information to every species found.

The entities relationship of the taxonomy hierarchy has been designed by defining the cardinality of the relationship as one to many (1:n). For example, one type of Genus can have many type of Species. The hierarchical structure of the taxonomy identified the parent child relationship amongst the entities. The eight domains and subcategories in the taxonomy will function as parent and child entities. Every entities of the taxonomy hierarchy are linked by defining the foreign key (FK) of the child entities that comes from the primary key (PK) of Genus as shows in Table 3 and Table 4. The ID_Genus is defined as the foreign key for Species entity.

![Entity-relationship diagram of taxonomy database](image)

**Figure 2.** Entity-relationship diagram of taxonomy database

The main entities are defined based on the biological taxonomy hierarchy which also create the main data tables which are Domain, Kingdom, Phylum, Class, Order, Family, Genus and Species. Entities of Wildlife_Status, Type_of_Forest and IUCN_Status are defined to describe the species status and the habitat. An entity of Location is created to link the geospatial information to every species found.

The entities relationship of the taxonomy hierarchy has been designed by defining the cardinality of the relationship as one to many (1:n). For example, one type of Genus can have many type of Species. The hierarchical structure of the taxonomy identified the parent child relationship amongst the entities. The eight domains and subcategories in the taxonomy will function as parent and child entities. Every entities of the taxonomy hierarchy are linked by defining the foreign key (FK) of the child entities that comes from the primary key (PK) of Genus as shows in Table 3 and Table 4. The ID_Genus is defined as the foreign key for Species entity.

| Column               | Data Type | Description               |
|----------------------|-----------|---------------------------|
| ID_Species (PK)      | Int       | Primary key               |
| ID_Genus (FK)        | Int       | Link to Genus type        |
| ID_Type (FK)         | Int       | Link to type of forest    |
| ID_Wildlife (FK)     | Int       | Link to Wildlife Act status |
| ID_IUCN (FK)         | Int       | Link to IUCN Redlist status |
Table 4. Fields of genus data table

| Column       | Data type | Description               |
|--------------|-----------|---------------------------|
| ID_Genus     | Int       | Primary key               |
| ID_Family    | Int       | Link to Family type       |
| Name         | Varchar   | Full name of the genus    |

Table 5. Field of wildlife_status data table

| Column   | Data type | Description            |
|----------|-----------|------------------------|
| ID_Type  | Int       | Primary key            |
| Name     | Varchar   | Full name of the genus |

Table 6. Field of location data table

| Column              | Data type | Description                                    |
|---------------------|-----------|-----------------------------------------------|
| ID_Type             | Int       | Primary key                                   |
| Location            | Varchar   | Full name of the location                     |
| GPS coordinate      | Varchar   | GPS coordinate of the found species           |
| Year                | Int       | Discovered year of the species                |

Beside the taxonomy hierarchy, the foreign key concept is also used to link the species information to other entities such as Wildlife_Status, Type_of_Forest, IUCN_Status and Location. For example, Table 5 and Table 6 show the data table of Wildlife_Status and Location.
4. Result and Discussion

4.1 Taxonomy Information

In the taxonomy information interface, user enable to insert, delete, update and view the taxonomy information of a specimen. Based on the database structure, the taxonomy hierarchy have been defined as a separated entities. The foreign key concept in data attributes is used to relate each of the hierarchy by sharing the unique ID of primary key. Every species may share the same hierarchy of taxonomy with other species. Therefore, we have designed a separate interface to create the flexible environment for storing and managing the taxonomy information. Figure 3 shows the insert new kingdom taxonomy interface.

![Figure 3. Insert new kingdom taxonomy interface](image)

4.2 Species Information

The species information can be inserted through this interface by allowing the user to choose the type of taxonomy hierarchy. Every taxonomy hierarchy that have been created in the taxonomy information interface will be displayed on drop down list as shows in Figure 4.
The discovered information will allow users to enter the location, GPS coordinates, and year of the found species. The habitat of the species can be specified by choosing the type of forest. The status of the species is categorized using two types of standards: Wildlife Conservation Act (2010) status and IUCN Red List status. Photos of the species can also be uploaded in this interface.

The data will be displayed by categorizing it into five types of information: taxonomy and nomenclature, taxonomy hierarchy, discovery information, status, and others. Figure 5 shows the data view of taxonomic biodiversity data.
5. Conclusion

The biological taxonomy developed in this study has been implemented in a database designed for taxonomy biodiversity data of Kenyir landscape. We present a relational database model of iTaxo that combine the field data of taxonomic nomenclature, spatial and temporal data. The primary use of iTaxo is to provide a data repository for users in biodiversity research and public as well. Every user has different authorization of the system access based on their status as a researcher, system administrator and public. The database will have to be updated as new data is validated. The taxonomy and database from this study can promote standardization in the design of taxonomies across a range of biodiversity research.

The outcome of iTaxo is the accessibility of the integrated biodiversity data that will enable data sharing among the researchers. Data retrieval for following analyses can be done by organising primary datasets on biodiversity in a consistent and flexible database. Based on the analyses, it may also supporting decision making among the related users in terms of conservation effort of nature and policy making. Also, the informative knowledge of the database will educate the public about the status of current biodiversity sources that can promote the awareness of preserving our nature ecosystem. The main part of assessing and understanding the status of biodiversity in Kenyir landscape can be done by combining the large amount of biodiversity data in the context of the status of new found species, species threatened and species distribution.

Acknowledgment

Appreciation conveyed to Universiti Malaysia Pahang RDU180363 grant and Kenyir Geopark Development Fund grant.

References

[1] S. Pagiola, J. Kellenberg, L. Vidaeus, and J. Srivastava, “Mainstreaming biodiversity in agricultural development,” Financ. Dev., vol. 35, pp. 38–41, 1998.

[2] C. R. Margules, R. L. Pressey, and P. H. Williams, “Representing biodiversity: data and procedures for identifying priority areas for conservation,” J. Biosci., vol. 27, no. 4, pp. 309–326, 2002.

[3] A. Z. M. Kamruzzaman, H. Selamat, and M. T. Wahid, “Conceptual Design of Biodiversity Data Model (BiDaM) using Object Relational and Event Based Approach,” in Proceedings of the Postgraduate Annual Research Seminar, 2005, pp. 77–81.
[4] T. Lotz, J. Nieschulze, J. Bendix, M. Dobbermann, and B. König-Ries, “Diverse or uniform? — Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research,” *Ecol. Inform.*, vol. 8, pp. 10–19, Mar. 2012.

[5] G. C. Bowker, “Biodiversity datadiversity,” *Soc. Stud. Sci.*, vol. 30, no. 5, pp. 643–683, 2000.

[6] C. S. Parr, R. Guralnick, N. Cellinese, and R. D. M. Page, “Evolutionary informatics: unifying knowledge about the diversity of life,” *Trends Ecol. Evol.*, vol. 27, no. 2, pp. 94–103, 2012.

[7] J. H. Porter, “Scientific Databases,” in *Ecological Data: Design, Management and Processing*, no. 1, W. K. Michener and J. W. Brunt, Eds. Oxford: Blackwell Science, 2000, pp. 48–69.

[8] S. Mohamad, “Generating visualisation for crime scene investigation based on probability result of knowledge-based system,” *Int. J. Comput. Appl. Technol.*, vol. 56, no. 1, pp. 18–26, 2017.

[9] N. Enke, A. Thessen, K. Bach, J. Bendix, B. Seeger, and B. Gemeinholzer, “The user’s view on biodiversity data sharing — Investigating facts of acceptance and requirements to realize a sustainable use of research data —,” *Ecol. Inform.*, vol. 11, pp. 25–33, 2012.

[10] N. F. Johnson, “Biodiversity informatics,” vol. 52, pp. 421–438, 2006.

[11] MyCHM, “Malaysia Biological Diversity Database,” 2008.

[12] GBIF, “Global Biodiversity Information Facility,” 2004. [Online]. Available: http://www.gbif.org. [Accessed: 01-Jun-2016].

[13] ITIS, “Integrated Taxonomic Information System,” 1990. [Online]. Available: http://www.itis.gov/. [Accessed: 10-Apr-2016].

[14] TDWG, “Taxonomic Database Working Group,” 1994.

[15] N. B. N. Gateway, “NBN Gateway,” 2012. [Online]. Available: https://data.nbn.org.uk/. [Accessed: 31-Oct-2016].

[16] M. Farquhar, “Definitions of quality of life: a taxonomy,” *J. Adv. Nurs.*, vol. 22, no. 3, pp. 502–508, 1995.

[17] S. F. Mawatari and H. Okada, “Taxonomy: A Discipline Central to Integrative Biology,” in *Proceedings of International Symposium on "Dawn of a New Natural History - Integration of Geoscience and Biodiversity Studies*, 2004, pp. 29–39.