Database Design Specification

for

<Taxonomic Hierarchy Comparator>

Version 0.1.0

Prepared by <Colin>

<2013/7/15>

Content

1 Introduction

1.1 Purpose

This document is about design detail of the database named “TaxonomicForest” which is specific for the web based tool “Taxonomic Hierarchy Comparator (THC)”. It is a reference for the database designer and software developer.

1.2 Background

a. Database name: TaxonomicForest
b. Application to use the database: Taxonomic Hierarchy Comparator (THC) Version 0.1.0
c. This project is proposed in EOL 2013 Rubenstein Program. Colin and Wang Jiangning are responsible for implementation. The database is intended for THC which is a newly proposed tool for analyzing various taxonomic hierarchies. This database will be installed by EOL-China and serve to THC.

1.2 Terminology

All the terms descriptions are excerpted from “methodology for comparing taxonomic hierarchies”.

1. Taxon
   Taxon is a taxonomic unit or group which may be named or not. A taxon encompasses all included taxa of lower rank and individual organisms. Taxonomic rank describes the level of a taxon in a taxonomic hierarchy for nomenclatural purposes. Each rank is either mandatory (e.g. kingdom, phylum, class, order, family, genus, species) or optional (e.g. subkingdom, subphylum, subclass, subfamily, subgenus, subspecies). Nominal taxon is a concept of a taxon which is denoted by an available name based on a name-bearing type.
2. Taxon Concept
   The scope of a taxon may differ from one taxonomists to another, and changes with new data. Each opinion as to what is intended by the name of a taxon is a 'taxon concept'..
3. Taxonomic Hierarchy
   Taxa are arranged hierarchically from high rank to low rank, and compose a taxonomic hierarchy (or “taxonomic tree” and “taxonomic classification”) that reflects the view on classification. As well known, taxonomic hierarchy is usually tree-based and can be regarded as a special case of

mathematical tree technically (Fig 2). Each node represents a taxon concept and its rank determines the vertical position in the hierarchy. Each node has one main label (accepted name) and may have more than one vice labels (synonyms). Definitely, biological taxonomic tree mentioned here is classification rather than phylogenetic tree.

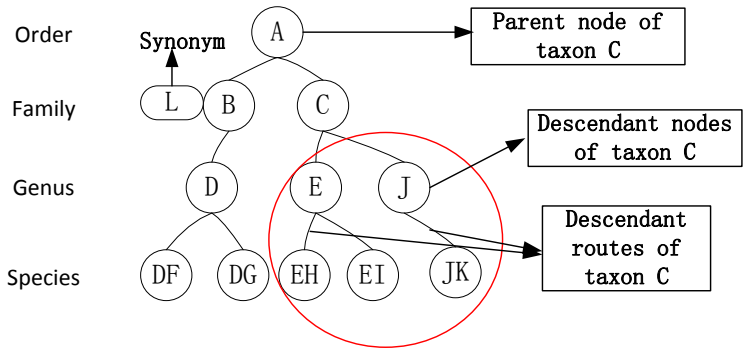


Fig 2 Taxonomic Hierarchy

4. Ancestor Node and Ancestor Path (AP)
   As shown in Figure 2, each node except the root in the taxonomic hierarchy has at least one node above it with higher rank. These nodes are named as Ancestor Nodes. They compose a path from root to the direct parent node called Ancestor Path (AP), e.g. AP_E=A -> C.
5. Descendant Node and Descendant Path (DP)
   For each node except the leaf nodes (the nodes without child taxon) in the hierarchy, there is one or more than one child taxon that is named as Descendant Node (DN). All these nodes compose a Descendant Node Set (DNS). For example, the DNS of taxon C in Figure 2 is:

   DNS(C) = {E, J, EH, EI, JK}

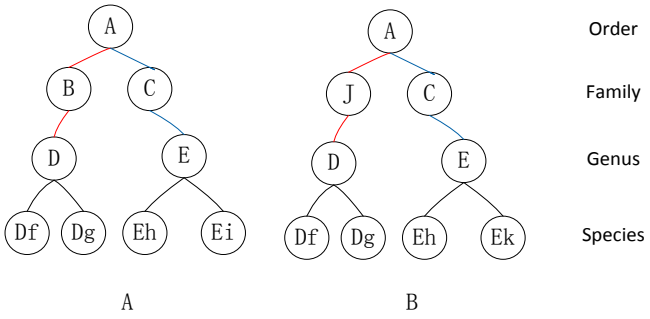
   The Descendant Nodes are arranged into some paths named Descendant Path (DP) ordered by their rank from up to down. These paths compose a Descendant Path Set (DPS). For example, the DPS of taxon C in Figure 2 is:

   DPS(C) = {E->EH, E->EI, J->JK}
6. Nominal Relation (NR)
   Because of the inexact nature of taxonomy, different taxonomies may use different names for the same taxon concepts and the same name for different taxon concepts (Homonym). When compare two taxon concepts from different hierarchies, scientific names including accepted name and synonyms for the concepts usually provide important clues about the relationship between them, and that is named as Nominal Relation. Possible cases are listed in Table 1 for two compared taxon concepts A and B. It is critical that NR is intended for expressing the possible relationships by scientific names and it is hard to discriminate the case of homonym. However, NR can be used as basis for further discrimination by experts or computer.

Taxon A	Taxon B	Value
Accepted	Accepted	1
Accepted	Synonym	2
Synonym	Accepted	3
Synonym	Synonym	4
Not matched		0

7. Ancestor Relation (AR)

Ancestor Relation is defined to describe the relationship between the Ancestor Paths of two compared taxa. If equal, then AR=1, else AR=0. For taxon D in tree 1 and tree 2, AP_{D1}=A->B, AP_{D2}=A->J, So $AR_{D1\text{-}D2}$=0. For taxon E in tree 1 and tree 2, AP_{E1}=A->C, AP_{E2}=A->C, So $AR_{D1\text{-}D2}$=1.



8. Descendant Relation (DR)

Descendant Relation is intended for assessing the similarity of the branches below the compared taxa. According to the definition of DNS and DPS, there can be two methods to calculate the similarity: (1) similarity based on set DNS called Node Index (NI); (2) similarity based on set DPS called Structure Index (SI).

1) Node Index (NI) and Node Similarity (NS)

Assuming that taxon *a* and taxon *b* are two taxa from compared trees, and *Num ()* is a function to calculate the number of elements of a set, then

$$\mathrm{NI_a} = \frac{Num(DNS_\mathrm{a} \cap DNS_\mathrm{b})}{Num(DNS_\mathrm{a})} \qquad (1)$$

$$\mathrm{NI_b} = \frac{Num(DNS_\mathrm{a} \cap DNS_\mathrm{b})}{Num(DNN_\mathrm{b})} \qquad (2)$$

$\mathrm{NI_a}$ reflects the percentage of common nodes present in $\mathrm{DNS_a}$ as well as the $\mathrm{NI_b}$ does in $\mathrm{DNS_b}$. Node Similarity (NS) can be deduced from NI by the following formula:

$$\mathrm{NS} = \frac{Num(DNS_\mathrm{a} \cap DNS_\mathrm{b})}{Num(DNS_\mathrm{a} \cup DNS_\mathrm{b})} = \frac{\mathrm{NI_a} * \mathrm{NI_b}}{\mathrm{NI_a} + \mathrm{NI_b} - \mathrm{NI_a} * \mathrm{NI_b}}$$

2) Structure Index (SI) and Structure Similarity (SS)

Assuming that taxon *a* and taxon *b* are two taxa from compared trees, and *Num ()* is a function to calculate the number of elements of a set, then

$$SI_a = \frac{Num(DPS_a \cap DPS_b)}{Num(DPS_a)} \quad (3)$$

$$SI_b = \frac{Num(DPS_a \cap DPS_b)}{Num(DPS_b)} \quad (4)$$

SI_a reflects the percentage of common paths present in DPS_a as well as the SI_b does in DPS_b. Structure Similarity (SS) can be deduced from SI by the following formula:

$$SS = \frac{Num(DPS_a \cap DPS_b)}{Num(DPS_a \cup DPS_b)} = \frac{SI_a * SI_b}{SI_a + SI_b - SI_a * SI_b}$$

According to Node index and structure Index, Descendants Relation can be classified into 5 cases (table 2).

Table 2 Cases of Descendants Relation

DR	Conditions	Value
Exclude	$NI_a = 0$ or $NI_b = 0$	0
Congruent	$SI_a = SI_b = 1$	1
Included	$SI_a = 1$ and $SI_b \neq 1$	2
Include	$SI_b = 1$ and $SI_a \neq 1$	3
Overlap	$NI_a \neq 0$ or $NI_b \neq 0$ and $SI_a \neq 1$ and $SI_b \neq 1$	4

9. Taxon Link (TL)

   Taxon Link is defined as a triple tuple (AR, NR, DR). It is the result of taxon to taxon comparison, and intended for representing the combined relationships among compared taxa (Fig3).

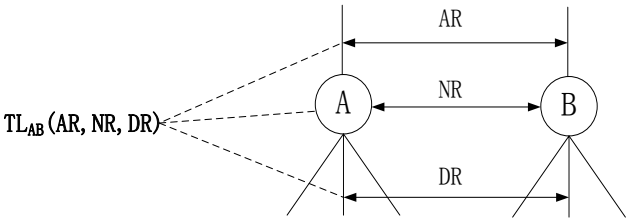


Fig3 Taxon Link

10. Compared Tree and Reference Tree

    Just like subtraction, a tree to be compared on the left of the comparator is regarded as compared tree (CT) and a tree for comparing on the right of the operator is called reference tree (RT).

1.4 References

1. Colin, 2012, Project Description in Application Documents for EOL 2013 Rubenstein Fellows.
2. Congtian Lin, Huijie Qiao, Jiangning Wang, Liqiang Ji*, 2012. Taxonomic Tree Tool for Managing and Comparing Taxonomic Trees (Abstract). 2012 TDWG Conference in China.

2 Design

2.1 Description of Database

According to the analysis of THC's requirements, the database should consider the following main data requirements:

1. The user information like username, password, and so on
2. The tree information like tree name, creator, created date, description of the tree and so on.
3. The experiment information like name, creator, created date, compare tree, reference tree and so on.
4. The result of comparison like similarity index, taxa relationship and so on

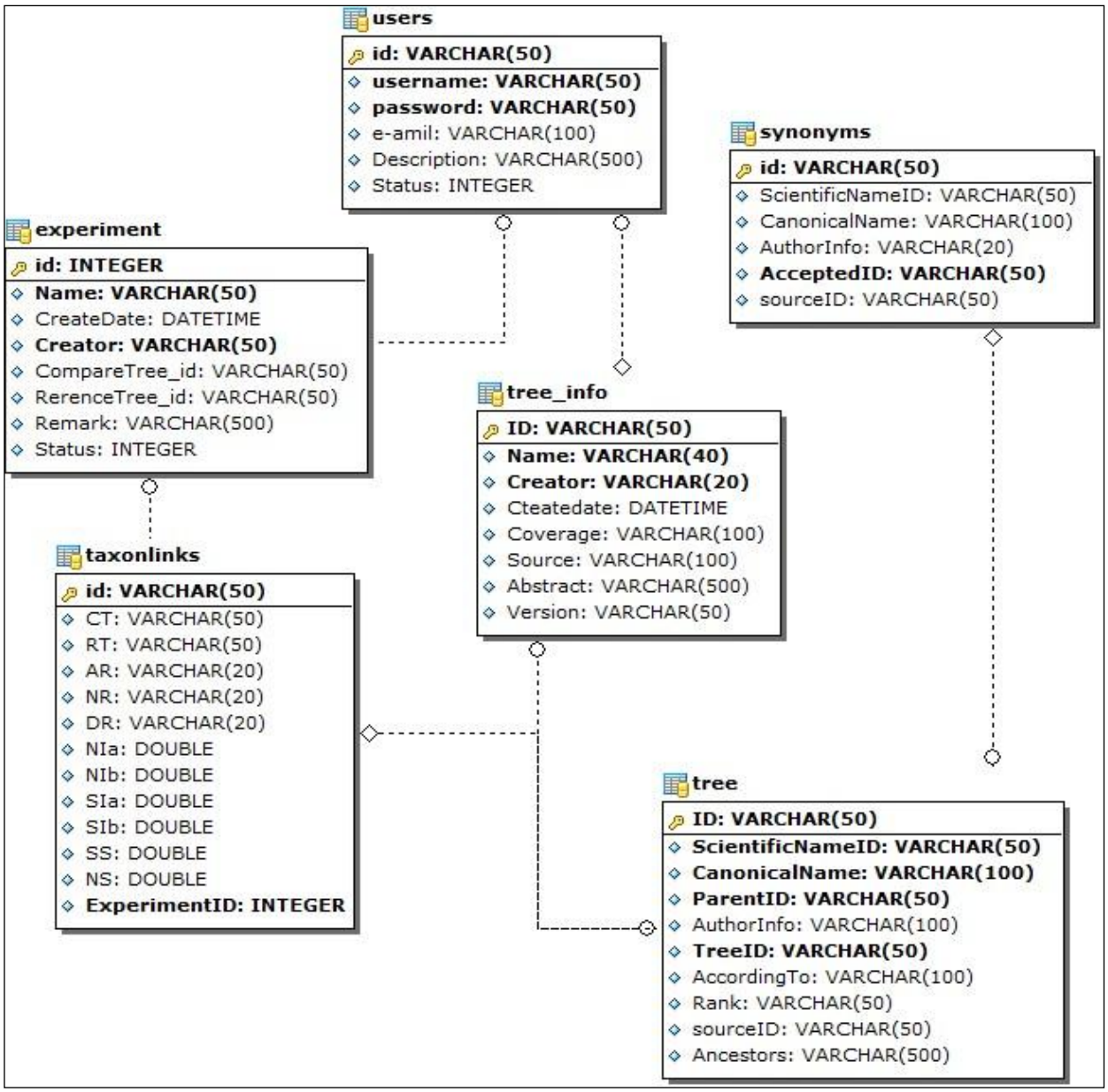
This database will store as many as possible taxonomic hierarchies (or referred as taxonomic trees in some references and web sites) just like a forest, so it is named as "TaxonomicForest" vividly.

2.2 Database Tool

MYSQL, version 5.5 or above.

2.3 E-R Diagram

The tables and relationships among them are described in E-R diagram:



E-R Diagram of Database TaxonomicForest

2.4 Tables & Columns

This section illustrates all the tables and fields in detail. Please refer to /databasereport/Index.html

2.5 Security

1. Automatic backup strategy by proper certain schedule e.g. each time one week.
2. Authorization strategy: Only manager have the full power to access the database.