Brief Report of Colin's Project in EOL 2013 Rubenstein Program (2013-06/2013-07)

1. Project Introduction

This project is to use the multiple classifications harvested by EOL for analysis to obtain the degree of coverage and congruence among hierarchies and nomenclatures. A tool called Taxonomic Hierarchy Comparator is proposed for analyzing these hierarchies. For more wide and long term use, we extend the task of this project to as following:

a. Propose a reliable method for comparing taxonomic hierarchies.
b. Develop a mature tool basing on comparison method for classifications comparison and visualization.
c. Find out the differences among classifications provided by EOL or from other sources, and propose quantitative indexes for measuring the degree of overlap and congruence among them.
d. Try to explore a possible method for taxonomists to mine differences of taxonomic views and find out potential taxonomic or nomenclatural acts.

2. Plan & Outcomes

Plan in 2013.06 – 2013.07:

Plan	Time line
Requirement analysis; Design the database and tool; Complete the design document.	2013.06.01 – 2013.07.31

Outcomes:

Finished	Outcome
Requirement analysis	"Software Requirements Specification for THC" "Use Cases Specification for THC"
Design the database	"Database Design Specification for THC"
Design the tool	Architecture of THC "UI Specification for THC"
Complete Methodology	Paper about methodology

All the documents can be downloaded from

3. Methodology

A method for analysis on various taxonomic hierarchies was proposed by us, and the detail is being written in a formal paper. We will try to contribute the article to an open access journal as soon as possible, and hope it can be published as an outcome of our project. The original idea can be referred to the project application document "Project Description", but the detail is not described in this

short report.

4. Requirement Analysis

Requirement analysis was finished. The requirements are grouped into three main collections:

1. User management
   User should have an account before using THC. That will help them to create and manage their own taxonomic hierarchies and keep analysis experiment result permanently for reuse. User management is required to manage user account including registration, authorization, log in/out, and account update.

2. Hierarchy management

a. Create hierarchy: hierarchy can be created from different methods. User can upload an Xml in BSBC format or DWCA file, or copy a hierarchy from shared hierarchies' pool. EOL hierarchies will be imported to THC by web API provided by EOL.
b. Edit hierarchy: a simple editor for user to modify the hierarchies. It will help user to edit scientific names, change position of taxon, insert new taxon, and delete taxon.
c. Share hierarchy: user can share his own hierarchies to others for analysis, but others cannot modify the shared hierarchies.
d. Export hierarchy: help users to save their hierarchy as standard DWCA file or BSBC xml.

3. Analysis experiment

a. Create experiment: give an identifier and some descriptions for the experiment; select which hierarchies for analysis. THC will keep the analysis result.
b. Share experiment: result of analysis can be shared with other users
c. Implementation: submit the analysis task to server, and waiting for response message. It is a time consuming process, so task queue is required to deal with multiple analysis tasks. Analysis is based on algorithm proposed by us.
d. Visualization: it is an important function for expressing analysis result. It should show where the congruence and incongruence explicitly is.
e. Computation: base on the result, "intersection" computation between two hierarchies is to extract the common part and "difference" is to produce the incongruence part.

For more detail about requirement analysis, please find in document "Software Requirements Specification for THC".
For use cases of this project, please refer to document "Use Cases Specification for THC".

5. Database Design

MYSQL 5.5 or above version is to be used in this project. All the procedures of designing database were finished. The database for THC is named as "Taxonomic

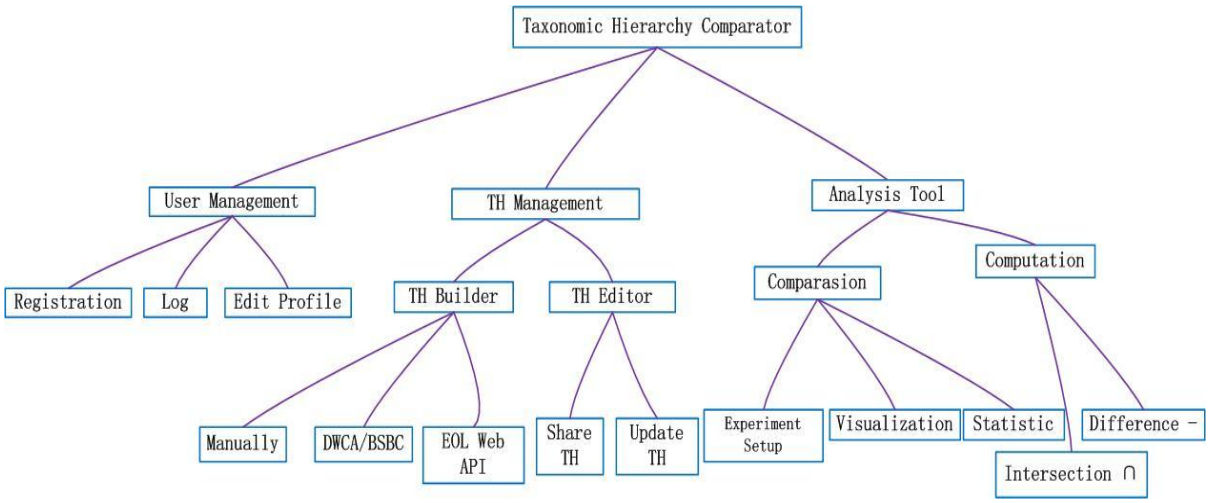
Forest" vividly, for it is to store as many as possible taxonomic hierarchies (or trees named in some other references or websites). The overview of "Taxonomic Forest" is shown in following table.

Table	Type	Cols	Description
experiment	InnoDB	8	This table is for storing detail about experiment, and describes the setup of experiment.
synonyms	InnoDB	6	Table 'synonyms' is to store synonyms of taxon in 'tree'. Synonym is regarded as vice label for taxon.
taxonlinks	InnoDB	13	Table 'taxonlinks' stores the relationships of taxa in compared trees. It is for the result of tree comparison and computation.
tree	InnoDB	10	Each record of 'tree' means a taxon including its propertis such as name, rank, position and so on.
tree_info	InnoDB	8	'tree_info' stores information about tree.
users	InnoDB	6	It is for user management, and records the information for registered users.

For detail about database, please refer to document "Database Design Specification for THC"

6. Architecture & User Interfaces

According to the requirements, we built up the architecture of THC and designed the user interfaces.



Architecture of THC

User interfaces were designed using pure html code. This process is to describe what the THC will be and how the THC will work. It considers and verifies the flow and requirements of THC except the style and outlook. The design is a first draft, and it will be updated during implementation.

For detail about user interfaces design, please refer to document "UI Specification for THC".