A Logical Model for Taxonomic Concepts for Expanding Knowledge using Linked Open Data

Rathachai Chawuthai[1], Hideaki Takeda[2], Vilas Wuwongse[3], and Utsugi Jinbo[4]

[1] Asian Institute of Technology, Prathumtani, Thailand
rathachai.c@gmail.com
[2] National Institute of Informatics, Tokyo, Japan
takeda@nii.ac.jp
[3] Thammasat University, Prathumtani, Thailand
wvilas@engr.tu.ac.th
[4] National Museum of Nature and Science, Tokyo, Japan
ujinbo@kahaku.go.jp

Abstract. The variety of classification systems and the new discovery of taxonomists lead to the diversity of biological information, especially taxon concepts. The association among taxon concepts across research institutes is very difficult to establish, because there is no single interpretation of the name of a taxon concept. Owing to this difficulty, further integration of more biological knowledge is very complicated when they deal with many sources of data or depending on different taxon concepts. This research aimes to develop a framework for linking some multiple related taxon concepts across research repositories, and preserving background knowledge of their changes. Therefore, we propose a logical model for taxon concepts in Resource Description Framework (RDF). Herewith, we implement a prototype to demonstrate the feasibility of our approach. It has been found that our model can publish taxon information as linked data and, hence, with additional benefits from Linked Open Data (LOD) cloud.

Keywords. Biological data, Biodiversity informatics, Logical model, Linked data, Ontology, Semantic web, Taxon concept

1 Introduction

More than 1.4 million species throughout the world have been truly described and classified with appropriate naming depended upon their characteristics; such as, morphological characters, living behaviors, DNA sequences, etc. [1-2]. Many taxonomists have dedicated themselves to study living organisms, research, and publish their knowledge for over hundred years. However, their researches have not been completely shared across all researchers around the world. In addition, there is no consensus on classification systems among taxonomists. In other words, taxonomists might have different perspectives to classify and name living organisms. As a consequence, a same species often be classified and named differently [2]. For example, *Papilio*

xuthus Linnaeus, 1767, Chinese Yellow Swallowtail Butterfly, has also been given several names by several taxonomists, such as *xuthulus* Bremer, 1861, *chinensis* Neuburger, 1900, *koxinga* Fruhstorfer, 1908, and *neoxuthus* Fruhstorfer, 1908.

The progress of taxonomic studies frequently causes redefinition of taxon concept, a circumscription of the taxon [2]. For instance, two genera of owls, *Nyctea* and *Bubo*, were merged into the latter genus *Bubo*. Following the change of genera, the scientific name of a snowy owl *Nyctea scandiaca* has been subsequently changed to *Bubo scandiacus* in order to satisfy the convention of scientific name [3]. Thus, a scientific name and a taxonomic concept become lacking of a single interpretation in biological [5-6]. Due to such change of taxon names, one sometimes misses information of this species under the name of the old scientific name when he or she searches information by the new scientific name.

Moreover, some details make researchers be confused when a taxon changes its concept without the change of its taxon name. For example, recently *Picoides tridactylus* (Three-toad Woodpecker) was split into two species, *P. tridactylus* (Eurasian Three-toad Woodpecker) and *P. dorsalis* (American Three-toad Woodpecker) [12]. Although these two species are disjointed, a part of information of *P. tridactylus*, especially recorded before the year 2003, might include details of *P. dorsalis*. One could obtain imprecise information when he or she simply searches information by the name *Picoides tridactylus*. Therefore, a mechanism that enables to link among taxon concepts in the precise context is necessary.

Recently, there was a research about managing the change in scientific conception. The work applied semantic web to develop a meta-ontology of a biological name (TaxMeOn). It provides metadata for representing and managing the temporal change of scientific name from a unit of taxon concept to another unit, and emphasized how the biological names publish [7]. However, the management of name change is not enough for researchers. The correct interpretation with temporal context of concepts and reasons of their changes becomes necessity as well.

The purpose of our research is to formulate a logical model for preserving background knowledge of the change of taxon concepts, and link some related concepts together. We introduced ontology for collecting the change of taxon concepts, cause and effect of the change; and linked data resulting from the change of concepts. We considered to enhance CKA [9] approach to capture the changes of taxon concepts, and their context. We also reused taxonomic terms from LODAC [8], employed SKOS[1] vocabularies to manage the relationship between concepts, and publicized data to Linked Open Data[2] (LOD) Cloud. Moreover, we performed an implementation to prove the feasibility of our proposed model.

To begin our approach, the background, the goal, and the related work have been already reviewed in this section. Next, Section 2, we will illustrate some technologies to develop our approach, and introduce the logical model in RDF. Section 3 will present prototype and discuss about its outcome. Lastly, Section 4 will draw conclusions and suggest some future improvements.

---

[1] Simple Knowledge Organization System: `http://www.w3.org/TR/skos-primer/`

[2] Linked Open Data: `http://linkeddata.org/`

2 The Proposed Logical Model

In this section, to achieve our objectives, we introduced a logical model for taxonomic concepts for expanding knowledge using LOD. Here, our model is expressed in ontology named Linked Taxonomic Knowledge (LTK) which was enhanced from several existing approaches.

Firstly, we studied how to classify the change of taxon concept; we found that they are two major categories: the change of name, and the change of classification [2,7,11]. A taxon name is sometimes changed for several reasons. For example, Hoare (2008) established the genus *Kendrickia* (ostracods). Then Kempf (2010) found that this genus was a primary junior homonym of *Kendrickia* Solem, 1985 (gastropods), and proposed *Dickhoarea* as the replacement name for *Kendrickia* Hoare, 2008. It results to the subsequent change of species names; for instance *Kendrickia asketos* had been changed into *Dickhoarea asketos* since Kampf (2010) has been published [2]. Apart from such name change, classifications also may be changed according to the progress of taxonomic researches. For example, the genus *Columba* (pigeons) has been split into five genera: *Patagioenas*, *Chloroenas*, *Lepidoenas*, *Oenoenas*, and *Columba* in the new narrow concept, and then some species of genus *Columba* have been assigned to one of these newly separated genera [12]. For instance, *Columba speciosa* changed to *Patagioenas speciosa* [12]. The analysis of the changes of taxon concept is described by Fig. 1.

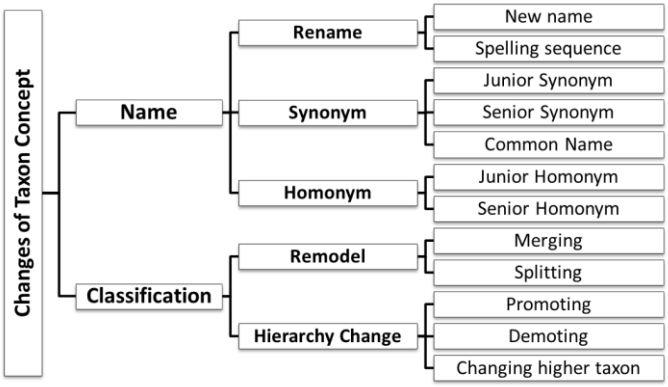

Fig. 1. The analysis of the changes of taxon concept

Secondly, we reviewed ideas in TaxMeOn, to describe concepts in taxonomic field linked to identifiers [7]. In general, when a concept's scope is changed, the changed concept needs to be recognized as new identifier. For instance, the genus *Bubo* before merging with *Nyctea* must not be the same identifier as *Bubo* after merging [2-3]. Thus, an identifier similar to those in TaxMeOn is required to our model. On the other hand, most attributes of the old *Bubo* can be copied to the new *Bubo* definitely, because, the old *Bubo* and the new *Bubo* may share many attributes.

Next, to publish data, we reviewed some standards that can be reused for our model. To model the entities of taxon concepts, we considered reusing some vocabularies from Linked Open Data for ACademia (LODAC), a project to publish a wide range of academic data including species information [8]. For example, a relationship between a species and a genus can be described as RDF using LODAC terms (species and genus are namespaces for species and genus in LODAC, respectively):

```
species:Nyctea_scandiaca  species:hasSuperTaxon     genus:Nyctea .
```

Another issue is to describe changes of concept and associated information on the change. There is an approach named Contextual Knowledge for Archives (CKA) Ontology. It offers a logical model developed using Flouris's theory for presenting the changes of conceptions, such as, merge, replace, and split. It also presents reasons behind the changes, changes of relationships such as the reclassification of a concept, and links between some relevant concepts. The CKA illustrates the change of concepts as dynamic RDF that contains fact and temporal aspect [9-10]. For instance, the following RDF expresses the splitting of a genus *Columba*.

```
ex:change2003 cka:interval          [tl:beginAtDateTime "2003"] ;
              cka:assure            ex:split1 .
ex:split1   rdf:type              ltk:TaxonSplitter ;
            cka:conceptBefore     genus:Columba ;
            cka:conceptAfter      genus:Patagioenas, genus:Chloroenas,
                                  genus:Lepidoenas,  genus:Oenoenas,
                                  genus:Columba_2003.
```

Further, the framework provides a technique to transform this dynamic RDF to static RDF with a given specific time point. For example, after year 2003, relationships among genus:Columba and its allies can be found as follows:

```
genus:Columba ltk:splitInto      genus:Patagioenas, genus:Chloroenas,
                                 genus:Lepidoenas,  genus:Oenoenas,
                                 genus:Columba_2003.
```

Technically, some operations from CKA framework can be extended to record the change of some concept's details, such as, color, size, organ, behavior, etc. It can be done by defining some new operations of change, and then binding the new operations with some related attributes. In addition, this model states one change as one unit. It offers association among related units of some changes by having some properties: cka:caused, and cka:effect to express reason and outcome of a change respectively. For example, Fig. 2 demonstrates the new name *Patagioenas speciosa* and its background. Consequently, we can find out the history of the name "*Patagioenas speciosa*". Then, we can use its background concept, such as the old name "*Columba speciosa*" to explore more information in the public LOD.

Fig. 2. Change of a taxon concept and its background

Lastly, to link data with LOD Cloud, our research proposed some useful operations that specify the change of taxon concepts, the changes of details of a taxon concept, the changes of relationships between taxon concepts, and the background of the change. All operations are defined by extending some vocabularies from the well-known ontology: Simple Knowledge Organization System (SKOS), and some properties from LODAC and CKA. Thus, the data from our framework can definitely be exchanged among other repositories. Example of some properties is shown in Table 1.

Table 1. Example properties from LTK which are derived from CKA, LODAC, and SKOS

Properties	rdfs:subPropertyOf
ltk:higherTaxon	cka:higherClass, skos:broaderTransitive, and species:hasSuperTaxon
ltk:replacedTo	cka:serialLinkTo, and skos:exactMatch
ltk:mergedInto	cka:serialLinkTo, and skos:relatedMatch
ltk:majorMergedInto	cka:serialLinkTo, and skos:closeMatch
ltk:synonym	skos:exactMatch

For example, the genus:Nyctea and genus:Bubo in old concepts have been merged into a new concept with the name *Bubo*. As stated previously, the genus *Bubo* in the new concept should be given a new identifier. In practice, we ended the year when it has been changed, so the new identifier of genus:Bubo may be genus:Bubo_1999. The property named ltk:mergedInto is defined to express a merge of two taxon concepts. The relationship between genus:Nyctea and genus:Bubo_1999 remains to be specified by the property ltk:mergedInto. On the other hand, another special property name ltk:majorMergedInto is introduced to demonstrate the very close relationship of two concepts, such as genus:Bubo and genus:Bubo_1999. As *Nyctea* was merged to *Bubo*, *Nyctea scandiaca*, the only member species of *Nyctea,* is transferred to *Bubo* and change the name to *Bubo scandiacus* [2-3]. In summary, these facts will be presented in RDF that satisfies the logical model of the CKA approach as follows:

```
ex:change1999    bibo:performer       pp:Wing, pp:Heidrich ;
                 bibo:issuer          pp:Richard ;
                 dcterms:source       pub:5224773;
                 cka:interval         [tl:beginAtDateTime "1999"] ;
                 cka:assure           ex:mg1, ex:rp1, ex:ac1 .
ex:mg1           rdf:type             ltk:TaxonMerger ;
                 cka:conceptBefore    genus:Bubo, genus:Nyctea ;
                 cka:conceptAfter     genus:Bubo_1999 .
ex:rp1           rdf:type             ltk:TaxonReplacement ;
                 cka:conceptBefore    species:Nyctea_scandiaca ;
                 cka:conceptAfter     species:Bubo_scandiacus .
ex:ac1           rdf:type             ltk:HigherTaxonAddition ;
                 cka:child            species:Bubo_scandiacus ;
                 cka:parent           species:Bubo_1999 .
ex:mg1           cka:cause            ex:rp1 .
ex:rp1           cka:detail           ex:ac1 .
```

After that, we apply some rules to transform dynamic RDF data to static form. For example, a rule that infers the merging operation of taxon concepts is expressed along these lines:

```
?change   rdf:type           ltk:TaxonMerger .
?change   cka:conceptBefore  ?before .
?change   cka:conceptAfter   ?after .
-------------------------------------------------
?before   ltk:mergedInto     ?after .
```

This rule and some others rules that infer each operation of change can convert the temporal RDF to be the following result.

```
genus:Nyctea              ltk:mergedInto          genus:Bubo_1999 .
genus:Bubo                ltk:majorMergedInto     genus:Bubo_1999 .
species:Bubo_scandiacus   ltk:higherTaxon    genus:Bubo_1999 .
species:Bubo_scandiacus   ltk:synonym        species:Nyctea_scandiaca .
genus:Nyctea                   cka:expired         "1999" .
genus:Bubo                     cka:expired         "1999" .
genus:Bubo_1999                cka:entered         "1999" .
species:Nyctea_scandiaca       cka:expired         "1999" .
species:Bubo_scandiacus        cka:entered         "1999" .
```

Therefore, clients can query these facts conveniently. For instance, if the users query some genera, which closely match (skos:closeMatch) genus:Nyctea; they will get genus:Bubo_1999. They sometimes query the data with species:hasSuperTaxon and get the result as same as ltk:higherTaxon. They can also find the present-day taxon concepts by inquiring some concepts which do not have a property named cka:expired. Moreover; the client can query more detail about a fact that includes the time when it changed, people who involved, reference documents, and triple data. For example, the replacement of species:Nyctea_scandiaca was caused by the merging between genus:Nyctea and genus:Bubo. In addition, the relationships of concepts can be presented by RDF statements, because the operation ltk:HigherTaxonAddition can establish the associations between concepts by producing some triples with having a property named ltk:higherTaxon. Our work offers some operations binding with properties; such as, dwc:scientificName[3], foaf:depiction[4], species:hasCommonName [8], etc. Thus, the consumers can query temporal information of taxon concepts along with specific time point.

3 Implementation and Discussion

After developing the LTK ontology, we verified the possibility and feasibility of it by implementing a prototype. The prototype is a web-based system that comprises three service layers: web interface, web services, and RDF data store. Firstly, the web interface allows a user to create the knowledge of taxon concepts in RDF. It also demonstrates the temporal context and link of taxon concepts. Further, it presents the reasons and details about changes of them. Secondly, the Java servlet service is made for

[3] Darwin Core Terms: http://rs.tdwg.org/dwc/terms/

[4] Friend of a Friend: http://xmlns.com/foaf/0.1/

managing and computing RDF data by using the performance of Jena[5] reasoning engine. Other clients can access data via this layer. Lastly, we used SESAME[6], a RDF store, to record data. Users can create data which come from some publications or books, and then the data is published to LOD cloud by providing SPARQL endpoint.

In Fig. 3, the left-side screen presents the context of the species:Nyctea_scandiaca (the figure displays as spc:Nyctea_scandiaca) and its linked taxon concepts, and the right-side screen shows information about the reason of changing this species. The web interface allows user to enter URI of concept and a specific time point in order to display the temporal context information as well.

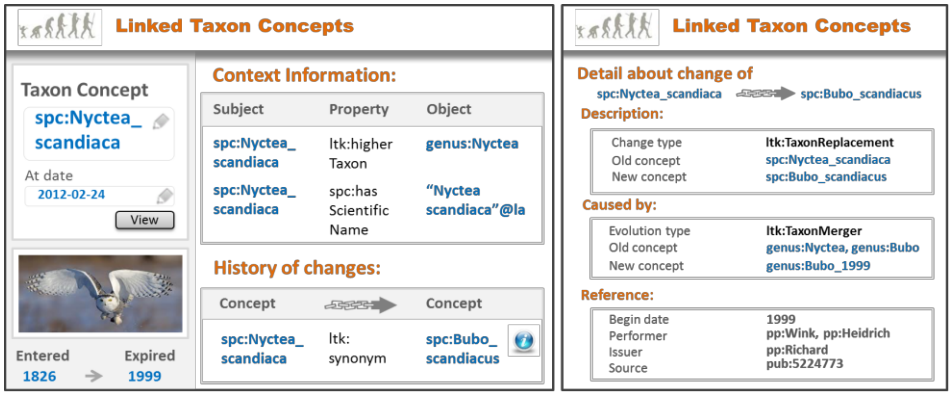

Fig. 3. Example screen of information about the concept species:Nyctea_scandiaca

As example RDF data in section 2, one change consists of many triples. When all changes are recorded, the triple store will manage over billion triples. Thus, it will consume a lot of resources when the service transforms the dynamic data to flat data for every request. However, most of all requests always ask for the present data. The prototype has to prepare current static data every time when each dynamic data is recorded. Then, the service can provide fast responses for the present information.

In summary, the prototype indicated that our approach is possible and feasible to make a real system. Moreover, other services can retrieve this data from LOD cloud.

4 Conclusions and Future work

Our paper presents a logical model and ontology for linking taxon concepts which comprises a series of changes, the diversity of taxonomic classifications, and the variety of naming. For the purpose of linking data, we have developed our model by employing ontology of contextual knowledge evolution together with some widely accepted ontology such as LODAC and SKOS. Therefore, our model can deal with both dynamic and static information represented in RDF and hence can trace the history of

[5] Apache Jena - reasoners and rule engines: `http://jena.apache.org/`

[6] SESAME – a framework for processing RDF data: `http://www.openrdf.org/`

the taxon concept. In addition, we have implemented a prototype which utilizes the proposed model in order to publish the taxonomic information to LOD cloud. As a consequence, other applications that need linked taxon concepts can readily connect to these data. Moreover, we have implemented a knowledge base using Jena's inference engine and SESAME's storage for computing data, and we have provided a web application to record and present the information. The result from our prototype demonstrates that our approach is feasible and suitable for the need of linked taxon concepts across different repositories and relationship backgrounds in order to discover broader knowledge of biology.

However, our approach gives priority to ontology rather than software application; hence the system requires much human effort to import a great number of data. For example, when a genus is split, some species under the genus have to move to new genera. In this case, taxonomists have to analyze and enter data by themselves. Thus, it should have some algorithms to improve the reclassification of some taxonomic ranks by their attributes. Moreover, in the future, when the number of data is over a billion, requesting historical data would be a great challenge because it requires the inference engine to process complex activities that consume very high computing capability. Future research might be focusing on how to improve the computing resources or methodologies for caching time-series of taxonomic data.

References

1. Darwin, C., Peckham, M.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Penn Press, Philadelphia (1959)
2. Winston, J.E.: Describing Species: Practical Taxonomic Procedure for Biologists. Columbia University Press, New York (1999)
3. Wink, M., Heidrich, P.: Molecular evolution and systematics of the owls (Strigiformes). In: A Guide to Owls of the World. Yale University Press, Yale (1999)
4. International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature. The International Trust for Zoological Nomenclature, London (1999)
5. Mallet, J.: Species, Concept of. In: Encyclopedia of Biodiversity. Elsevier, Oxford (2007)
6. Ytow, N., Morse, D., Roberts, D.: Nomencurator: a nomenclatural history model to handle multiple taxonomic views. In: Biological Journal of Linnean Society, pp. 81-98 (2001).
7. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web: managing the change in scientific conception. In: ESWC 2011. LNCS, vol. 6644, pp. 255-269. Springer, Heidelberg (2011)
8. Linked Open Data for Academia, `http://lod.ac/`
9. Chawuthai, R., Wuwongse, V., Takeda, H.: A Formal Approach to the Modelling of Digital Archives. In: ICADL 2012. LNCS, vol. 7634, pp. 179-188. Springer, Heidelberg (2012)
10. Flouris, G., Meghini, C.: Terminology and Wish List for a Formal Theory of Preservation. In: PV 2007. Proceedings, DLR, Munich (2007)
11. Franz, N., Peet, R.: Towards a language for mapping relationships among taxonomic concepts. In: Systematics and Biodiversity, vol. 7, iss. 1, pp. 5-20 (2009)
12. Banks, R.C., Cicero, C., et al.: Forty-fourth supplement to the American Ornithologists' Union check-list of North American birds. In: The Auk, vol. 120, pp. 923–931 (2003)