Presenting and Preserving the Change in Taxonomic Knowledge for Linked Data*

Rathachai Chawuthai
King Mongkut's Institute of Technology Ladkrabang,
Thailand
rathachai.ch@kmitl.ac.th

Hideaki Takeda
National Institute of Informatics
Japan
takeda@nii.ac.jp

Vilas Wuwongse
Mahidol Univeristy
Thailand
vilas.wuw@mahidol.ac.th

Utsugi Jinbo
National Museum of Nature and Science
Japan
ujinbo@kahaku.go.jp

ABSTRACT

Linked Open Data (LOD) technology enables a web of data and exchangeable knowledge graphs through the Internet. However, the change in knowledge is happened everywhere and every time. This issue is commonly found in every domain especially in taxonomic knowledge. In fact, biological classification and taxonomic names are not consistent among different databases due to new discovery and various viewpoints of taxonomists. This issue leads to ambiguity in taxon interpretation and imprecise linked data. In this case, the temporal representation of taxa and underlying knowledge of the change in taxonomy are considered to integrate with taxonomic data models as well. For this reason, this work objects to introduce an approach to the preservation and the presentation of change in taxonomic knowledge, and to produce linked data for supporting world-wide knowledge exchange. The main outcome is an ontology including operations for presenting changes, context-based human-readable identifiers indicating revision of taxon concepts, an event-based data model for preserving changes with contextual information, and two linked data models for presenting chronological changes in taxa and their temporal information at a given time point. All of these features are originated to provide a better understanding of organisms to learners. In addition, a prototype is implemented to demonstrate the feasibility of the proposed approach against the real cases from domain experts. The result shows that our approach can support various practical cases of changes in taxonomic knowledge and provides open and accurate access to LOD cloud.

CCS CONCEPTS

• **Computer Science** → *Semantic Web*; Linked Open Data • **Biodiversity Informatics** → Taxonomy; Change in Taxonomy

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04. https://doi.org/10.1145/3184558.3186234

KEYWORDS

Biodiversity Informatics; Change in Taxonomy; Knowledge Representation; Knowledge Exchange; Linked Data; Ontology; RDF; Semantic Web; Taxonomic Data.

1 INTRODUCTION

Linked Open Data (LOD) technology attempts to transform web of documents into web of data and make them become exchangeable through the Internet [1]. Thus, all pieces of knowledge around the world becomes in form of semantic knowledge graphs which are mathematically described by a binary relation, and the advantage of this format is to have data become easily linkable [2]. However, knowledge is generally dynamic and subjected to change over time due to new discovery and new justification from humans. A lot of out-of-date and up-to-date pieces of knowledge are intermixed everywhere among sources of knowledge in the Internet. This issue is one cause of the misinterpretation and misunderstanding of some terms or some concepts from chronological transaction records under the different context of time and across communities, and it results in the inaccuracy of linked data. Integrating an aspect of time in a knowledge graph becomes an appropriate solution, but it cannot be done straightforwardly due to the limitation of the binary relation. For this reason, our work efforts to introduce an approach to preserving and presenting the change in knowledge graphs for linked data, and we take this approach into a biodiversity domain because taxonomic knowledge under this domain is commonly changed and has clearly chronological transaction records documented.

Taxonomy (or biological taxonomy) is the sciences of naming and classifying the group of organisms considered by their characteristic shared. Each group in the hierarchical classification is called "taxon" (plural "taxa") and it must be named. A lot of articles about taxonomic knowledge have been published around the world for 200 years. These pieces of knowledge could be linked for increasing learning capacity. However, there is no precise referenceable key to associate these

pieces of knowledge because there is no globally accepted scientific names and classifications [3]. There are more than one scientific names for a species, and there are many changes in the classifications [4-7].

For example, two genera of snowy owls, *Bubo* and *Nyctea*, have been merged under the name *Bubo* since 1999 [5]. According to the zoological nomenclature rule [8], the species *Nyctea scandiaca*, which was under the genus *Nyctea*, have been transferred into the newly accepted genus *Bubo* with the new name *Bubo scandiacus* [5]. There are a lot of online documents about this snowy owl but there are detailed with these two names separately. For having knowledge exchange, changes in taxonomy of these species must be records with precise context in a standard data model for linked data.

For this reason, we conduct this research whose objectives are to: (1) preserve the change in taxonomic knowledge, and (2) present and publish taxonomic knowledge as linked data.

As we studied from other taxonomic databases [9-17], the online linked taxonomic knowledge is needed, and those works do not support our research objectives.

To accomplish these two objectives, we implement an approach named Linked Taxonomic Knowledge (LTK) based on the idea of two previous works, Contextual Knowledge for Archives (CKA) approach [18] and the Meta-Ontology of Biological Name (TaxMeOn) [19]. We also reuse some taxonomic terms from Linked Open Data for ACademia (LODAC) [20] and Simple Knowledge Organization System (SKOS) [21] to manage the relationships between concepts, and publish data to the Linked Open Data (LOD) Cloud [1]. The LTK and its prototype will be described hereafter.

2 A LOGICAL MODEL FOR LINKING TAXONOMIC KNOWLEDGE

The model of LTK is proposed based two points: (1) the model can preserve the changes as an event along with aspects of time and provenance, and allows tracing background knowledge behind the change; and (2) the model can make linked data be human- and machine-readable, and support the chronological and temporal data presentation.

All data models are designed to be compatible with Semantic web technology, so all of them can be represented by the Resource Description Framework (RDF) format [22]. RDF data are in the form of a subject–predicate–object expression, known as a triple, and many triples generate an RDF graph which becomes knowledge graph as well.

The detail our work is illustrated as follows:

2.1 Structure to Represent Change in Taxonomic Knowledge

On the basis of these changes analyzed from actual use cases [9-17], we first categorize changes in taxonomic knowledge into three mains types as shown in Fig. 1. Change in nomenclature is the change in the name of taxa, change in a taxon concept is the change in circumscription or classification scope of a taxon, and change in relationship is change in association or link between two taxon concepts. There are about 15 terms in the leaf nodes that will become predefined operations representing changes under our LTK ontology.

In practice, some following operations are often used. *Synonym* is a link between different names of same taxon. *Merging* is to lump two or more taxa into one taxon. *Splitting* is to separate one taxon into two or more taxa. *Changing higher taxon* is to transfer a lower taxon to another super taxon. For merging and splitting, the names of taxa before the change become obsolete but not the selected names after the change.

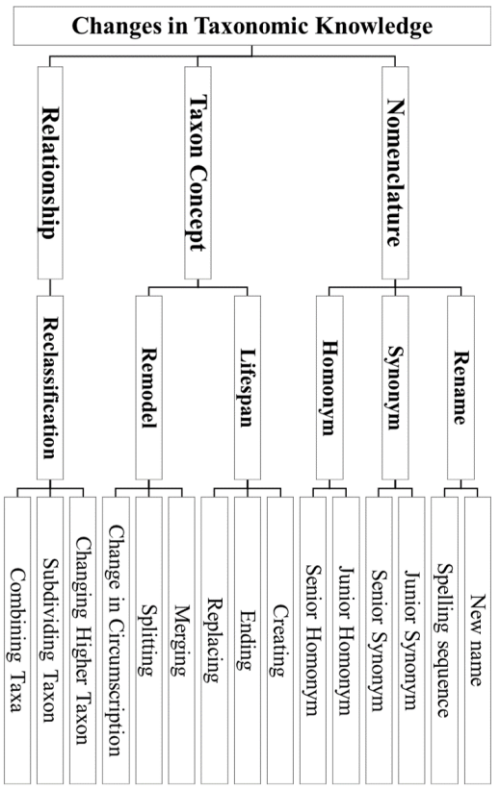

Figure 1: Overall data models of LTK

2.2 Entities for LTK

An entity is a set of names represented by Uniform Resource Identifier (URI). For example, the URI of the genus Bubo is *http://rc.lodac.nii.ac.jp/taxon/genus/Bubo*; and it can be shorten to *genus:Bubo* according to user-defined prefixes. There are three types of entities.

Nominal Entity (***ltk:NOM***) is a set of URIs represented taxonomic concepts (taxon concept) which can be either names or identifiers used in a taxonomic information system.

Simple Nominal Entity (***ltk:SIM***) is a subset of *ltk:NOM* which contains only human-readable URIs such as *genus:Bubo*.

Contextual Nominal Entity (***ltk:CON***) is a subset of *ltk:NOM* which presents a version number in a URI. For example, *genus:Bubo_1999* has been a version of *Bubo* since the year 1999. This entity is generally used in the LTK ontology, and it always links to the *ltk:SIM* for linking to an LOD cloud.

2.3 Operations representing Changes

According to Section 2.1, the structure of changes is used to be operations for presenting the change in data models.

Operation of Change (***ltk:OPR***) is a set of types of changes used in the LTK ontology. It includes the followings subtypes.

Operation of Change in Conception (***ltk:OPRC***) is a set of the changes in the circumscription (scope) of taxon concepts such as splitting a taxon into taxa, merging taxa in to a taxon, and replacing a taxon into another one. The URIs for them are *ltk:TaxonMerger*, *ltk:TaxonSplitter*, and *ltk:TaxonReplacement*.

Operation of Change in Relation (***ltk:OPRR***) is a set of the changes in relations or associations between two taxon concepts, such as changing a higher taxon to another one, subdividing a taxon into lower taxa, combining taxa into a higher taxon, and linking a taxon to its synonym. The URIs for them are *ltk:ChangeHigherTaxon*, *ltk:SubdivideTaxon*, *ltk:CombineTaxa*, and *ltk:SynonymLink*.

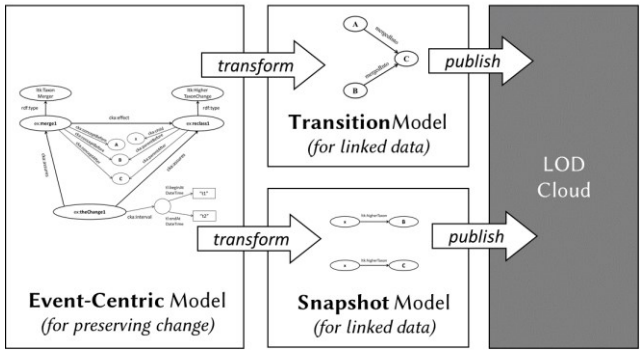

Figure 2: Overall data models of LTK

2.4 Data Models for LTK

There are three data models. We first introduce an Event-Centric model that preserve any changes in taxonomy. This model contains aspects of time and provenance data that uses the idea of *n*-ary relation but presented by the binary relation, so the model becomes complicated by design and it is not suitable for linking to LOD. Thus, we next create two data views of the Event-Centric model; there are a Transition model and a Snapshot model in order make light-weight data models for exchanging knowledge through LOD cloud as shown in Fig. 2.

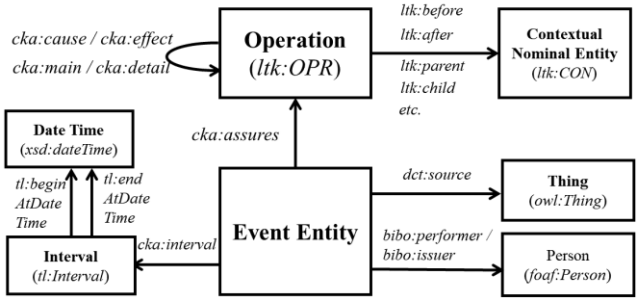

Figure 3: Event-Centric Model Schema

2.5 Event-Centric Model

In detail, the Event-Centric model assures *Operations* of change (by a property *cka:assure*). It also includes provenance data to show the *Interval* time of that change (by a property *cka:interval*) with have beginning time (*tl:begin*) and end time (*tl:end*), to inform *Persons* who perform or issue this change (by properties *bibo:Performer* and *bibo:issuer*), to show relationships and consequences of the change (by properties *cka:main, cka:detail, cka:cause,* and *cka:effect*), and to cite some referent *Sources* (by a property *dct:source*) as shown in Fig. 3.

For example, the case of change in the taxonomy of snowy owls [5,8] can be represented by the following RDF statements.

```
1  ex:event1999
2    bibo:issuer  pp:Richard ;
3    cka:interval [tl:begin "1999"];
4    cka:assures  ex:mg1, ex:rp1 .
5
6  ex:mg1  rdf:type  ltk:TaxonMerger ;
7    ltk:taxonBefore  genus:Bubo_1805 ;
8    ltk:taxonBefore  genus:Nyctea_1826 ;
9    ltk:taxontAfter  genus:Bubo_1999 .
10
11 ex:syn1 rdf:type  ltk:SynonymLink;
12   ltk:sourceTaxon
13         species:Nyctea_scandiaca_1826 ;
14   ltk:targetTaxon
15         species:Bubo_scandiacus_1999 .
16
17 ex:mg1  cka:effect  ex:syn1 .
```

We use contextual nominal entities for presenting scientific names because non-semantic-web-expert users especially in biological researchers can use less effort on recognizing it. All URIs in the RDF statements are shorten using user-defined prefixes. Lines 1-4 show that the event entity refers to some operations and provenance data. Lines 6-9 show the merging of two genera into the *genus:Bubo_1999*. Lines 11-15 express the synonym of both species. Line 17 presents that the merging of two genera contributes to the synonym of both species.

The *n*-ary data structure of the Event-Centric model is like a data model in a relational database. It is suitable to embed contextual information of change for preserving in a knowledge base and become background knowledge of every change. However, using the binary relation for presenting the *n*-ary relation makes the data model be complex, so its design is not proper for consuming by any LOD applications directly. In this case, the Event-Centric model must be transformed into general knowledge graphs that are the Transition model and the Snapshot model for publishing data to the LOD cloud.

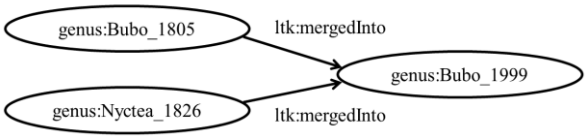

Figure 4: Example data of the Transition Model

2.6 Transition Model

The Transition model is a knowledge graph for presenting the chronological chain of the changes in contextual nominal entities. The model is simpler than the Event-Centric model because its contextual information is omitted. For example, the RDF statements presenting the Event-Centric model of the change in taxonomy snowy owls from Section 2.5 can be transformed into a simple knowledge graph as shown in Fig. 4.

According to that figure, the property *ltk:mergedInto* is appeared to be a link between taxa before and after merging. This property is a linking property and it is bound with the operation *ltk:TaxonMerger* as expressed in the following RDF statements. Our approach restricts that every operation must be declared in the LTK ontology.

```
1 | ltk:TaxonMerger
2 |       ltk:linkingProperty  ltk:mergedInto .
```

This declaration supports the transformation rules to convert the Event-Centric model into the Transition model. The rules are well-defined based on the Semantic Web Rule Language, and it is executed by Semantic Web reasoning engine such as Jena [23].

Properties in the Transition model are link to some properties from well-known ontology for example, *ltk:mergedInto* is a sub-property of *skos:broadMatch*. If users make a query with the property *skos:broadMatch*, they will get the result from the property *ltk:mergedInto* too. It also can trace back to data in the Even-Centric model to view reason behind a particular transition pair. Thus, it becomes advantage to publish the chronological changes in taxa to LOD cloud.

2.7 Snapshot Model

As well as the Transition model, the Snapshot model is designed to be light-weight triples presenting temporal RDF data at a specific time point. For example, if the species *Nyctea scandica* is cited before 1999, it is considered as an accepted species. However, when it is mentioned after 1999, it is a synonymy of the *Bubo scandiacus* as shown in Fig. 5. It means that when we learn about *Bubo scandiacus*, we should link to any information related to *Nyctea scandica* documented after 1999 as well.

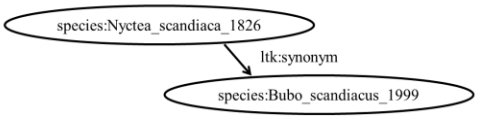

Figure 5: Example data of the Snapshot Model since 1999

The implementation of the transformation process is similar to the Transition model. We firstly declare this operation with a relation property. For example, the relation *ltk:synonym* is bounded with the operation *ltk:SynonymLink*. Then, a proposed rule is able to convert the Event-Centric model into the Snapshot model presenting temporal information of taxa based on a given time point.

3 APPLICATION

To test the feasibility of our approach, a web application is implemented. As shown in Fig. 6., the application has three layers: the first layer contains a web application, web services and a query endpoint; the second layer has a rule engine and ontologies; and the last layer provides RDF data. Thus, users can access the web application at *http://rc.lodac.nii.ac.jp/ltk/*, while other client applications or services can access the system through the web services and the query endpoint. The system also provides dereferenceable URIs that returns a document type according to the mine type in the request header.

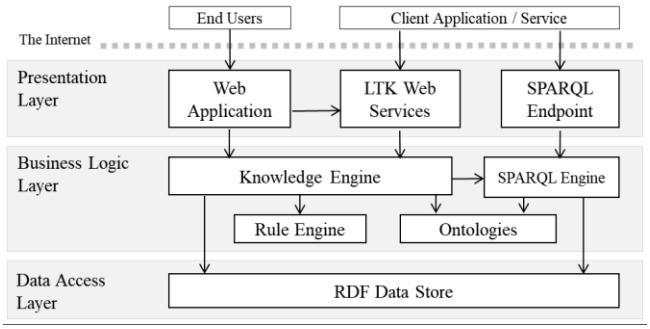

Figure 6: Software architecture of the prototype

We also test our approach with real test cases of Japanese moths under the family Saturniidae published as three checklists (list of names): Inoue in 1982 [24], Jinbo in 2008 [25], and Kishida in 2011 [26]. It has been found that our work can handle all the cases and all data are linked with any Internet resources from well-known ontologies in the LOD cloud. The domain experts from biodiversity informatics and Semantic Web accept that this work gives contribution to the biodiversity domain in terms of knowledge representation, user engagement, and system integration; and it can be applied to any other domains as well.

4 CONCLUSION

Our work introduced an approach to the preservation and the presentation of changes in knowledge for linked data by focusing on the biodiversity domain. Changes in taxonomic classification and names are the key issue of our research. The approach contains the LTK ontology; and it includes the Entities of concept, the Operations for categorizing changes, and the Data models for documenting changes. The primary data model is the Event-Centric model that captures the change with aspects of time and provenance. The secondary data models that are the Transition model and the Snapshot model are initiates to be the light-weight views of the Event-Centric model. The simplicity of the last two data models are proper for being used by LOD applications. In addition, the prototype has been developed to tested with the real cases of changes in taxonomic knowledge. The outcome indicates that our work is one contribution to empower the capability of knowledge management for global accessibility across communities and time using linked data.

REFERENCES

[1] T. Heath and B. Christian, "Linked data: Evolving the web into a global data space," in *Synthesis lectures on the semantic web: theory and technology*, 2011.

[2] P. Hitzler, M. Krotzsch and S. Rudolph, Foundations of semantic web technologies, CRC Press, 2009.

[3] N. Franz and R. Peet, "Perspectives: towards a language for mapping relationships among taxonomic concepts," *Systematics and Biodiversity*, vol. 7, no. 1, pp. 5-20, 2009.

[4] J. E. Winston, Describing species: practical taxonomic procedure for biologists, Columbia University Press, 1999.

[5] International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature, 4 ed., International Trust for Zoological Nomenclatu History Museum, 1999.

[6] C. G. Sibley and L. L. Short, "Hybridization in the orioles of the Great Plains," *Condor*, pp. 130-150, 1964.

[7] S. Freeman and R. M. Zink, "A phylogenetic study of the blackbirds based on variation in mitochondrial DNA restriction sites," *Systematic Biology*, vol. 44, no. 3, pp. 409-420, 1995.

[8] M. Wink and P. Heidrich, "Molecular evolution and systematics of owls (Strigiformes)," *Owls A Guide to the Owls of the World*, pp. 39-57, 1999.

[9] W. G. Berendsohn, "A taxonomic information model for botanical databases: the IOPI model," *Taxon*, pp. 283-309, 1997.

[10] A. C. Jones, R. J. White and E. R. Orme, "Identifying and relating biological concepts in the Catalogue of Life," *Journal of Biomedical Semantics*, vol. 2, no. 1, 2011.

[11] J. B. Kennedy, R. Kukla and T. Paterson, "Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration," in *Data integration in the life sciences*, Springer, 2005, pp. 80-95.

[12] N. Laurenne, J. Tuominen, H. Saarenmaa and E. Hyvonen, "Making species checklists understandable to machines--a shift from relational databases to ontologies," *Journal of Biomedical Semantics*, vol. 5, no. 1, 2014.

[13] R. D. Page, "Taxonomic names, metadata, and the Semantic Web," *Biodiversity Informatics*, vol. 3, 2006.

[14] D. J. Patterson, J. Cooper, P. M. Kirk and et al., "Names are key to the big new biology," *Trends in ecology & evolution*, vol. 25, no. 12, pp. 686-691, 2010.

[15] I. N. Sarkar, "Biodiversity informatics: organizing and linking information across the spectrum of life," *Briefings in Bioinformatics*, vol. 8, no. 5, pp. 347-357, 2007.

[16] S. Schulz, H. Stenzhorn and M. Boeker, "The ontology of biological taxa," *Bioinformatics*, vol. 24, no. 13, pp. i313-i321, 2008.

[17] N. Ytow, D. R. Morse and D. M. Roberts, "Nomencurator: a nomenclatural history model to handle multiple taxonomic views," *Biological journal of the Linnean Society*, vol. 73, no. 1, pp. 81-98, 2001.

[18] R. Chawuthai, V. Wuwongse and H. Takeda, "A Formal Approach to the Modelling of Digital Archives," in *The Outreach of Digital Libraries: A Globalized Resource Network*, Springer, 2012, pp. 179-188.

[19] J. Tuominen, N. Laurenne and E. Hyvonen, "Biological names and taxonomies on the semantic web: managing the change in scientific conception," in *The Semanic Web: Research and Applications*, Springer, 2011, pp. 255-269.

[20] Y. Minami, H. Takeda, F. Kato and et al, "Towards a Data Hub for Biodiversity with LOD," in *The 2nd Joint International Semantic Technology Conference*, 2013.

[21] "Simple Knowledge Organization System," [Online]. Available: http://www.w3.org/TR/skos-primer/.

[22] G. Schreiber and Y. Raimond, "RDF 1.1 Primer," 2014. [Online]. Available: https://www.w3.org/TR/rdf11-primer/.

[23] "Apache Jena," [Online]. Available: http://jena.apache.org/.

[24] H. Inoue, S. Sugi, H. Kuroko and et al., Moths of Japan, vol. 2. Plates and synonymic catalogue, vol. 2, Tokyo: Kodansha Tokyo, 1982.

[25] U. Jinbo, "List-MJ: A checklist of Japanese moths 2004-2008," 2008. [Online]. Available: http://listmj.mothprog.com.

[26] Y. Kishida, The Standard of Moths in Japan II, vol. 2, Tokyo: Tokyo: Gakken, 2011.