Preserving Taxonomic Change and Subsequent Taxon Relationships over Time

Andreas Kohlbecker[1], Naouel Karam[2,3], Adrian Paschke[2] and Anton Güntsch[1]

[1]*Department of Biodiversity Informatics, Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin, Königin-Luise-Straße 6-8 14195 Berlin, Germany*

[2]*Fraunhofer FOKUS, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany*

[3]*Institute for Applied Informatics (InfAI), University of Leipzig, Goerdelerring 9, 04109 Leipzig, Germany*

Abstract

Biodiversity research data often reference individual organisms, populations, or other taxonomic contexts by using scientific names. Scientific names, however, are unstable, ambiguous, and no precise identifier for the specific taxonomic concept that has been used implicitly. Using identifiers for taxonomic concepts instead does not fully solve the inherent semantic problems, since taxon concepts may evolve over time, therefore preserving and representing their relationships is critical for any subsequent analysis. We propose a model to represent and preserve the taxonomic change as Linked Data. The approach aims additionally at preserving the semantic relationships between the resulting taxon concepts as well as the temporal sequence of changes. Our model describes taxon relations as set-theoretical relations and thus makes use of the underlying semantics to enable automatic reasoning over the knowledge base.

Keywords

Taxonomic Change, Taxon Relationships, Biodiversity, Linked Open Data, Semantics, Taxonomy, Taxonomic Concept, Scientific Name

1. Introduction

The rapidly increasing global change causes a dramatic impact on ecosystems, and thus demands regular and timely assessments on the status and trends of biodiversity and ecosystem services, and their interlinkages at the local and global level. One of the keys to understanding the transformation of ecosystems is biodiversity research data. Vast amounts of these data have been made available on the Internet, but sources are disparate and data formats are heterogeneous.

Infrastructure projects like NFDI4Biodiversity [1] are dedicated to facilitating the integration, use, and exchanging of biodiversity data. NFDI4Biodiversity's goal is to make the variety of biodiversity data from a multitude of sources available via unified interfaces. This involves mapping different data schemes to each other or to a common scheme. This however only covers the technical level, at the same time data also needs to be mapped semantically.

For example, an observation of a species in Germany from 1989 refers to a narrower geographical area than an observation of the same species in 1991 that refers to Germany as well. The latter observation is dated to the period after German reunification and thus refers to a

S4BioDiv2021: 3rd International Workshop on Semantics for Biodiversity

a.kohlbecker@bgbm.org (A. Kohlbecker); naouel.karam@fokus.fraunhofer.de (N. Karam); adrian.paschke@fokus.fraunhofer.de (A. Paschke); a.guentsch@bgbm.org (A. Güntsch)

broader area. Both usages of the term ”Germany” need to be related to each other to express that the one is a sub-region of the other. Practically this can be done by providing a terminology backbone in which all semantic relationships are expressed that are relevant for the expected use-cases.

Biodiversity research data mostly are related to a species or subspecies, which are represented by the respective scientific name. A name that taxonomists have given to a group of organisms sharing common characteristics and which in their entirety can be defined as a species, or another unit of classification at a different rank. A species, circumscribed by means of a set of descriptive values that distinguish it from other species is a taxonomic concept tagged with a scientific name. Changing knowledge and insight into the individual organisms and their characteristics that form a taxon concept may modify the set of individual organisms that are covered by that concept, that is its circumscription [2]. A taxon concept can become broader or narrower causing more or fewer individuals to be enclosed in the set. Upon that, a new taxon concept emerge while the scientific name remains the same. Taxon concepts bearing the same name can be overlapping, congruent, included, or conversely, even completely distinct. Even completely different names may refer to exactly the same taxon concept (Fig. 1).

Hence, biodiversity data labeled with the same scientific name do not necessarily refer to the same taxon concept. For a complete semantic mapping of biodiversity data the following requirements have to be fulfilled:

1. stable and reliable taxon concepts, whereas each taxon concept is assigned with a persistent identifier.
2. Names used in biodiversity data need to be related to their taxon concepts that have originally been used, when the data has been created. This is done by annotating data sets with the taxon concept identifiers. Preferably this is done at creation time, or afterwards as in most cases.
3. The relationships between taxon concepts need to be expressed semantically.

2. Context and Motivation

Data networks like NFDI4BioDiversity aim at accumulating huge amounts of environmental data sets from heterogeneous resources and thus create a data space in which taxon concepts from multiple checklists of a wide range of different time periods are referred to. In order to apply inference on these data, it is essential to have knowledge of the relationships of the taxon concepts being involved. Without this knowledge, it impossible for example to correctly assemble all data sets from heterogeneous sources which refer to the same taxon concept. Therefore it should be possible to express the taxonomic relations between all taxon concepts and their change over time in each relevant checklist and between checklists.

In previous works, several approaches to semantically model taxon concept relationships have been published. While Franz et. al [3] discuss the general importance of using taxon concepts and to express their relationships for data integration in taxonomy, phylogenetics, and biodiversity research, Michel et. al [4] propose a model to express the relations between taxon concepts and scientific names of different checklists. Among other reasons, their approach has shortcomings for the use case of data networks like NFDI4BioDiverity, since the temporal aspect of taxonomic changes cannot be expressed. The Linked Taxonomic Knowledge (LTK) model by Chawuthai et al. [5], which we will discuss in more details below, is sufficient in terms of representing the historical dimension of the taxon concept evolution.

Figure 1 shows the timeline for the reclassification of the Baltimore oriole (Icterus galbula Linnaeus, 1758) and the Bullock's oriole (Icterus bullockii Swainson, 1827). In 1964, Sibley and Short argued that these two species should be merged into a single one [6]. In 1995, DNA sequencing of the two species led to the splitting of Icterus galbula into Icterus galbula and Icterus bullockii again [7]. As a consequence, data recorded between 1964 and 1995 about Icterus galbula may contain useful information about Icterus bullockii, yet a search for Icterus bullockii will not lead to any records in this interval.

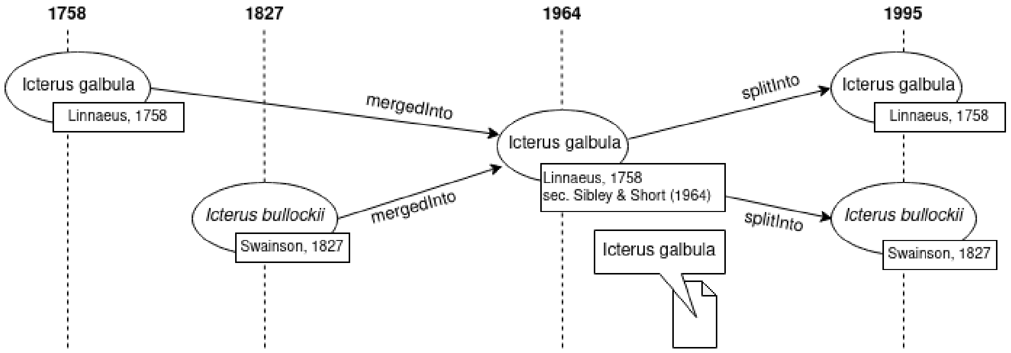


Figure 1: Timeline of the taxonomic change of the Baltimore oriole (Icterus galbula Linnaeus, 1758) and the Bullock's oriole (Icterus bullockii Swainson, 1827)

In order to preserve the taxonomic change information, we need to keep track of the temporal aspect of the change as well as the implications in terms of extensional definitions. The taxonomic concept of the species Icterus galbula from 1964 includes instances of Icterus bullockii and thus is more general than the Icterus galbula from 1995. Our model should in consequence be able to: (1) differentiate between the two taxon concepts, (2) keep track of taxonomic changes including provenance and temporal aspects and (3) preserve the semantic relation between different versions of the same concept and to other related concepts.

For the first objective, we plan to assign persistent HTTP identifiers to multiple taxonomic concepts of the same scientific name as this complies with the principles of Linked Data [8]. Additionally, we follow the naming recommendation proposed by Berendsohn [9] for labeling different taxon concepts, which consists on following the full Linnaean name by the term *sec.* (from the latin *secundum*) and the specific author and publication. Our taxon concept from 1964 will be named *Icterus galbula Linnaeus, 1758 sec. Sibley & Short (1964).*

3. Representing Taxonomic Change for Linked Data

A logical model named Linked Taxonomic Knowledge (LTK) for preserving and presenting the change in taxonomic knowledge has been introduced in [5]. The model is based on an ontology of contextual knowledge evolution for representing historical information about taxa and preserving background knowledge of the change. The model covers changes in nomenclature (rename, synonym and homonym), taxon concept (merging, splitting, change in circumscription) and relationship (change in higher taxon). Based on the LTK model, the merge event ex:event1 of Icterus galbula and Icterus bullockii in 1964 will be described as depicted in figure 2.

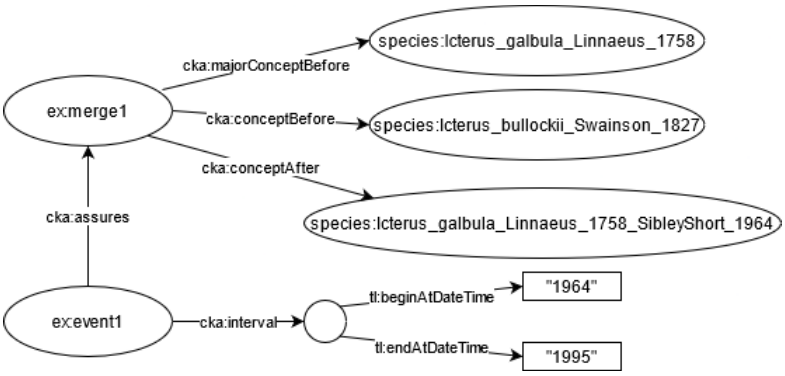


Figure 2: Event-centric model for taxon merge of Icterus galbula and Icterus bullockii

The merge operation *ex:merge1* is assigned the relationships *cka:conceptBefore* and *cka:conceptAfter*, relating it to the concepts to be merged and the resulting concept. The event interval is identified by giving a begin and eventually an end time point.

4. A Model for Representing Taxon Relationships

Once different usages of a name are assigned to their unique taxon concepts, we need to reconnect them and model semantic relationships between them. Five basic set-theoretic relationships, as depicted in figure 3, are fundamental for the description of the connection between two taxonomic concepts [10]: A and B are congruent ($A \equiv B$), A is included in B ($A \subset B$), A includes B ($A \supset B$), A and B overlap ($A \oplus B$) and finally, A and B exclude each other (A ! B). To maximize the expressiveness, these oriented relations should be considered as being mutually exclusive. That is subset relationships should be used exclusively as proper subsets. When $A \subset B$ we can conclude that $B \supset A$, therefore the set of relationships required to describe the set-theoretic connection of taxon concepts can be limited to these four: $\equiv$, $\subset$, $\oplus$, !.

In our example, we would represent the fact that *Icterus galbula Linnaeus, 1758 sec. Sibley & Short (1964)* is more general than *Icterus galbula Linnaeus, 1758* using the inclusion relationship.

We propose to model the relations using description logics (DL) [11] logical axioms. Each relationship is modeled using a DL axiom which interpretation corresponds to its underlying meaning in set theory as follows :

- Congruence: A and B are equivalent ($C \equiv D$);
- Inclusion: A is subsumed by B ($A \sqsubseteq B$); or A subsumes B ($A \sqsupseteq B$);
- Exclusion: A and B are disjoint ($C \sqcap D \equiv \bot$);

where $\bot$ is the bottom concept and is interpreted as an empty set. Due to the open world assumption, two concepts are presumed to be overlapping unless explicitly stated to be disjoint.

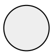
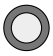
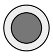
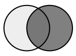
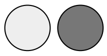
Basic relation	Representation
A and B are congruent A ≡ B x∈ A ⟺ x∈ B	
A is included in B A ⊂ B x∈ A ⟹ x∈ B, ∃y∈ B | y∉ A	
A includes B A ⊃ B x∈ B ⟹ x∈ A, ∃y∈ A | y∉ B	
A and B overlap each other A ⊕ B ∃x∈ A | x∉ B, ∃y∈ B | y∉ A, ∃z∈ A | z∈ B	
A and B exclude each other A ! B x∈ A ⟹ x∉ B	

Figure 3: Basic concept relations (representation after Geoffroy & al., 2003 [2])

Using axioms (1) and (2) below, we can model the relationships between our Icterus species taxon concepts. Going a step further, we can model the merge between the two species as the union of the two sets using axiom (3). Finally, we represent the two original species as disjoint using axiom (4).

```
Icterus_galbula_Linnaeus_1758 ⊑ Icterus_galbula_Linnaeus_1758_SibleyShort_1964 (1)
Icterus_bullockii_Swainson_1827 ⊑ Icterus_galbula_Linnaeus_1758_SibleyShort_1964 (2)
Icterus_galbula_Linnaeus_1758_SibleyShort_1964 ≡ Icterus_galbula_Linnaeus_1758 ⊔ (3)
Icterus_bullockii_Swainson_1827
Icterus_galbula_Linnaeus_1758 ⊓ Icterus_bullockii_Swainson_1827 ≡ ⊥ (4)
```

We represent those axioms as triples in an OWL format. The following RDF statements describe the axioms introduced above.

```
species:Icterus_galbula_Linnaeus_1758
    rdfs:subClassOf  species:Icterus_galbula_Linnaeus_1758_SibleyShort_1964 ;
    rdfs:label ”Icterus galbulaLinnaeus, 1758” .

species:Icterus_bullockii_Swainson_1827
    rdfs:subClassOf species:Icterus_galbula_Linnaeus_1758_SibleyShort_1964 ;
    rdfs:label ”Icterus bullockii Swainson, 1827” .

species:Icterus_galbula_Linnaeus_1758_SibleyShort_1964
    owl:equivalentClass [   rdf:type owl:Class ;
                            owl:unionOf (   species:Icterus_galbula_Linnaeus_1758
                                            species:Icterus_bullockii_Swainson_1827
                                        )
                        ] ;
    rdfs:label ”Icterus galbula Linnaeus, 1758 sec. Sibley & Short (1964)” .

species:Icterus_galbula_Linnaeus_1758
    owl:disjointWith species:Icterus_bullockii_Swainson_1827.
```

Once relationships triples are added to the knowledge base, it is possible to use such information for data access and analysis. For instance, a search for *Icterus bellucki* can lead now to data annotated with `Icterus_galbula_Linnaeus_1758_SibleyShort_1964`, as the information about the species inclusion has been encoded in the knowledge base.

Such triples can be derived from the event-centric model described in Section 3, using rules. For instance, a rule for the ex:merge1 in Figure 2, that can infer an inclusion relationship between the *cka:conceptBefore* and the *cka:conceptAfter* would be defined as follows:

```
[rule_merge:
    (?operation rdf:type ltk:TaxonMerger),
    (?operation cka:conceptBefore ?conceptBefore),
    (?operation cka:conceptAfter ?conceptAfter)
    -> (?conceptBefore rdfs:subClassOf ?conceptAfter)
]
```

5. Discussion

Our approach extends the Linked Taxonomic Knowledge (LTK) model [5] by explicitly expressing the set-theoretic relations that are implicit to data modeled through LTK. Depending on specific use-cases, inferring these implicit taxon relations on demand can be appropriate. For data networks like NFDI4BioDiversity, this information will be requested quite frequently, so that inference on demand may become too costly in terms of time and computing resources. Therefore it will be important to model these relations explicitly to avoid that computing overhead. As the LTK model is based on information of taxonomic changes it can express the historic dimension in one checklist, or of checklists that are historically connected by derivation.

In data infrastructures like NDFI4BioDiversity, it is needed to map taxon concepts of multiple checklists onto each other, which are otherwise completely disconnected. As their relations are not the result of changes in taxonomic knowledge, it is not possible to express them in LTK. The Set-theoretic taxon relations expressed in OWL are thus an appropriate method to model the connections between checklists.

Background information on taxon concepts is often sparse and insufficient to determine the type of concept relations between taxon concepts of different checklists with absolute confidence. Therefore, the set-theoretic taxon relationships will often be tainted with uncertainty, which needs to be expressed in addition. This is only one of the challenges on the way to an information space that allows efficient inference across related taxonomic concepts.

References

[1] F. O. Glöckner, M. Diepenbroek, J. Felden, A. Güntsch, J. Stoye, J. Overmann, K. Wimmers, I. Kostadinov, R. Yahyapour, W. Müller, U. Scholz, D. Triebel, M. Frenzel, B. Gemeinholzer, A. Goesmann, B. König-Ries, A. Bonn, B. Seeger, NFDI4BioDiversity - A Consortium for the National Research Data Infrastructure (NFDI) (2020). URL: https://zenodo.org/record/3943645. doi:DOI:10.5281/zenodo.3943645, publisher: Zenodo.

[2] M. Geoffroy, W. Berendsohn, The concept problem in taxonomy: Importance, components, approaches, in: Schriftenreihe Vegetationsk, volume 39, 2003, pp. 5–14. Journal Abbreviation: Schriftenreihe Vegetationsk.

[3] N. Franz, R. Peet, A. Weakley, On the use of taxonomic concepts in support of biodiversity research and taxonomy, The New taxonomy, systematics association special volume series 74 76 (2006). doi:10.1201/9781420008562.ch5.

[4] F. Michel, O. Gargominy, S. Tercerie, C. Faron Zucker, A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF, in: ISWC 2017 Workshop on Semantics for Biodiversity (S4Biodiv 2017), volume CEUR Vol. 1933, Vienna, Austria, 2017, pp. 1–12. URL: https://hal.archives-ouvertes.fr/hal-01617708.

[5] R. Chawuthai, H. Takeda, V. Wuwongse, U. Jinbo, Presenting and preserving the change in taxonomic knowledge for linked data (extended abstract), in: Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 461–465. URL: https://doi.org/10.1145/3184558.3186234. doi:10.1145/3184558.3186234.

[6] C. G. Sibley, L. L. Short, Hybridization in the orioles of the great plains, The Condor 66 (1964) 130–150. URL: http://www.jstor.org/stable/1365391.

[7] S. Freeman, R. M. Zink, A phylogenetic study of the blackbirds based on variation in mitochondrial dna restriction sites, Systematic Biology 44 (1995) 409–420. URL: http://www.jstor.org/stable/2413601.

[8] Berners-Lee, Linked Data - Design Issues, 2006. URL: https://www.w3.org/DesignIssues/LinkedData.html.

[9] W. G. Berendsohn, The concept of "potential taxa" in databases, TAXON 44 (1995) 207–212. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1222443. doi:https://doi.org/10.2307/1222443. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.2307/1222443.

[10] M. Koperski, M. Sauer, W. Braun, S. Gradstein, Referenzliste der Moose Deutschlands, Schriftenreihe für Vegetationskunde 34 (2000) 1–519.

[11] F. Baader, D. Calvanese, D. Mcguinness, D. Nardi, P. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, 2 ed., Cambridge University Press, 2007. doi:10.1017/CBO9780511711787.