The potential of using sets of specimens to handle species concepts

Ed Baker

November 2015

Abstract

The use of notation and concepts from mathematical set theory is investigated as a method for describing species concepts, and potentially higher level taxa. These methods may facilitate the easy databasing of species concepts, allowing the concepts themselves to become citable through the provision of unique identifiers. The increase in unique identifiers (such as Life Science Identifiers or Digital Object Identifiers) for biological specimens in recent years may make this approach more feasible than it would have been previously.

Contents

1 Problem

The idea od defining species concepts by collections of specimens has been previously proposed (e.g. [1]). Recent increases in the rate of specimen digitisation in natural history museums (e.g. [2]), combined with persistent identifiers for these specimens [4] allows for robust species concepts defined by a collection (set) of specimens. This paper experiments with using mathematical set notation (rather than the terminology used by [3]) to define operations on groups of specimens that may be considered equivalent to taxonomic and nomenclatural acts. These concepts could be used as basis for a repository of species concepts that are well-defined using specimens.

At the present time insufficient numbers of specimens have been assigned unique identifiers for this solution to be generally practical. It is presented here as an example potential use of unique identifiers, and to encourage discussion as to whether this approach has any merit (it may not).

The mathematics of relational databases is understood in terms of manipulations (relations) of sets. The expression of taxonomic and nomenclatural acts as functions on sets may aid the design of systems that record and track species concepts, which will by necessity need to be stored in databases. This approach is the reverse of [5] who developed an object-orientated model, and later showed it could be made into a relational database.

1.1 Reduction of problem

Species concepts are generally considered to comprise all organisms that are defined by that concept, that is to say that all wild living organisms that match the description and primary types are included. Placing the entirety of a population into a set is impractical, so here concepts are confined to specimens. Specimens are increasingly citable 'objects' and are here considered to include physical specimens, nucleotide sequences, etc. These concept sets may be considered to be a representative subset of the population if philosophically desirable.

This work deals solely with the species concept. Similarly methods (nested sets) could be applied to genus- and family- group concepts if so desired, with modification for the differing method of type-designation at these levels.

This paper does not deal with publications for the purpose of clarity. The association of publications of new species with their type specimens is straightforward, if time consuming for the historical literature. It is hoped that any real world system would have this functionality.

1.2 Nomenclature

The assignment of scientific names to species concepts is a great aid in communicating about organisms. This approach does not change anything relating to the naming of concepts. If anything it may help with algorithmically determining the name of a species concept based on specimens.

2 Introduction

We consider x to be a collection (set) of specimens $(s_1, s_2, s_3, ...)$ including a number of primary type specimens $(t_1, t_2, t_3, ...)$.

$$x = \{s_1, s_2, s_3, ..., t_1, t_2, t_3, ...\}$$

A competent taxonomist takes this set of specimens, and sorts them into groups they consider to represent species. If the group has been recently well studied then the groups may each contain a single primary type.

$$x_1 = \{s_1, ..., t_1\}$$

$$x_2 = \{s_2, ..., t_2\}$$

$$x_3 = \{s_3, ..., t_3\}$$

where $x_n \subset x; x = x_1 \cup x_2 \cup x_3 \cup ...$

Each x_n is a species concept, typified by the specimen t_n when there is a single primary type. No specimen appears in more than one subset.

2.1 Synonymy

If the taxonomist selected set contains multiple primary types this is an indication of synonymy.

$$x_n = \{s_a, s_b, s_c, ..., t_x, t_y\}$$

Define *type cardinality* as the number of valid types in a set.

$$typec(x) = |\{t|t \in x\}|$$

Assuming that t_n are valid types then typification can be resolved by the appropriate nomenclatural code.

1. When $typec(x) = 1$ then the concept can be named by the primary type.
2. When $typec(x) > 1$ then a selection of primary type is needed following the appropriate rules of nomenclature, following the concept of priority.
3. When $typec(x) = 0$ then the concept does not contain a type. The set should be expanded to include an appropriate primary type, or if no appropriate type is available then a specimen from the set should be described as the primary type.

2.1.1 Precedence

Of the valid types the oldest is the one used to formalise the species concept. If information on when the types were scientifically described is available we can select the type with the lowest date value.

Define

$$type(x) = min(\{t_{[date]} | t \in x\})$$

3 Comparison of species concepts

3.1 Identity of species concepts

The identity of species concepts in this scheme occurs when the concepts are sets containing the same specimens. Identity between concepts is potentially not the most useful way of determining if two or more sets are compatible (see Consistency). The mathematical identity of two sets is equivalent to each set being a subset of the other.

$$x = y \iff x \in y \wedge y \in x$$

3.2 Consistency of species concepts

Test to see if two, non-identical, species concepts are compatible.

Author A: $x_A = \{x_1, x_2, x_3\}$
Author B: $x_B = \{x_1, x_2, x_3, x_4\}$

The species concepts x_A and x_B can be considered to be consistent. An example would be where Author A creates their concept before Author B. Author B later expands Author A's concept with the addition of a new specimen. As Author A has not placed x_4 in any other species concept these two concepts can be considered to be compatible: it is only the fact that Author A was not aware of x_4 that they did not include it in x_A.

Define species concepts are *consistent* when the only specimens not in the intersection of the two concepts are not placed in another concept by either author.

$consitent(x_A, x_B) = \forall x_n \notin x_A \cap x_B$ and x_n not in other concepts by the same authors

4 Expanding scope of a species concept

4.1 Identification of expanded scope

We can test that x_B is an expansion of the scope of x_A by checking that x_A is a subset of x_B and that x_b has a larger cardinality than x_A.

$$x_A \subset x_B \wedge |x_A| < |x_B|$$

4.2 Expansion of scope

It should be noted that both of the following operations will result in the creation of a new species concept.

4.2.1 Addition of a specimen to a species concept

Adding a specimen to an existing species concept can be achieved through the union of that concept with the new specimen.

$$x_B = x_A \cup x_4$$

4.2.2 Synonymy of two previous species concepts

If two concepts, both previously considered to be valid, are found to be synonyms of each other then a new concept can be created that is the union of these two.

$$x_{new} = x_A \cup x_B$$

5 Reducing scope of a species concept

5.1 Identification of reduced scope

Identification of a concept as a reduced set of another requires checking the latter concept is a subset of the former, with reduced cardinality.

$$x_A \in x_B | \, |x_A| < |x_B|$$

5.2 Reduction in scope

It should be noted that both of the following operations will result in the creation of a new species concept.

5.2.1 Removal of specimens from a species concept

The creation of a species concept *de novo* based on those specimens the author wishes to retain is perhaps the easiest way forward, however there may be merit in some instances of explicitly recording the removal of specimens in relation to an existing concept. The specimens to exclude are defined by x_B.

$$x_{new} = \{x_A | x_A \not\subset x_B\}$$

5.2.2 Splitting of a species concept

The same operation as above can be used, with x_B as set representing a species concept rather than any set of specimens.

6 Remarks

The ability to store and manipulate species concepts, whether they be *de novo* concepts (such as from the description of new species) or relative to existing concepts (e.g. the creation or removal of synonymy) can be achieved simply through the use of sets of specimens, assuming that specimens have stable, persistent identifiers. The creation of new concepts becomes resource cheap and objective (at least in the definition of the concept).

It is not currently feasible, and may never be, for the creator of a concept to identify all relevant digitised specimens that may form part of that concept. The creation of an objective list of specimens that were used to form that concept however is certainly feasible. Given the low cost of creating a new concept, it would be easy for other individuals and institutions to create a new concept that expands the original to include additional specimens (e.g. a taxonomist finds the species from a new area; a museum digitises its collection of *Aus bus* and adds those specimens to the existing concept).

Primary types as starting point Given that digitising efforts may focus on the primary types held by an institution, and that this metadata of specimens is almost certain to be recorded the creation of initial 'specimen species concepts' based solely on these type series could be achieved automatically, providing a starting concept dataset.

Taxonomic works If concepts are citable through a unique identifier then the authors of taxonomic works could unambiguously reference the specimens they used in their work. This would clearly indicate which specimens they considered to bea given species during their research.

Specimen metrics The use of specimens in taxonomic works, if they can be cited, could be used as an indication of the taxonomic value of that specimen. This, when expanded to cover an institution, could also give rise to institutional metrics. Such metrics could be used to identify specimens that have not received recent attention and areas of collections that could benefit from increased levels of research.

To create a usable system of species concepts based on specimens a robust citation system for concepts needs to be created, with persistent identifiers. Persistent identifiers would allow the easy expansion and changing of concepts as new specimens are collected or digitised. Life Sciences Identifiers (LSIDs) or Digital Object Identifiers (DOIs) could be used for this purpose.

References

[1] Walter G Berendsohn. "The concept of" potential taxa" in databases". In: *Taxon* (1995), pp. 207–212.

[2] Vladimir Blagoderov et al. "No specimen left behind: industrial scale digitization of natural history collections". In: *ZooKeys* 209 (2012), p. 133.

[3] NM Franz and RK Peet. "Perspectives: towards a language for mapping relationships among taxonomic concepts". In: *Systematics and Biodiversity* 7.1 (2009), pp. 5–20.

[4] Robert P. Guralnick et al. "Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data". In: *ZooKeys* 494 (2015), pp. 133–154. DOI: 10.3897/zookeys.494.9352.

[5] Nozomi Ytow, David R Morse, and David Mcl Roberts. "Nomencurator: a nomenclatural history model to handle multiple taxonomic views". In: *Biological journal of the Linnean Society* 73.1 (2001), pp. 81–98.