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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

1 Problem

The idea of defining species concepts by collections (”sets”) of specimens (e.g. [1, 2]) and using mathematical sets (e.g. [3, 4] reviewed in a historical perspective by [5]) have been previously proposed. Recent increases in the rate of specimen digitisation in natural history museums (e.g. [6]), combined with persistent identifiers for these specimens [7] 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 [8]) 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.

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 [9] who developed an object-orientated model, and later showed it could be made into a relational database. Such a system could also be used to automatically describe relationships between subjects between taxonomic concepts as described by [8].

1.1 What is a specimen?

While most taxonomic databases deal with species scopes in the traditional sense of preserved museum specimens and published works (e.g. [2]), increasingly surrogates of these specimens are used, such as nucleotide sequences [10]. When defining a species concept it is possible to include non-type material that aid in forming a comprehensive knowledge of that species. These may include expanded range information from photographs, observations of living specimens, sound recordings [11], traits [12] and trace evidence [13]. As biodiversity databases expand in scope and their use of stable, persistent identifiers there is scope for bringing these additional data into a digital species concept.

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).

1.2 Competing hypotheses

While existing databases, explicitly or implicitly, handle specimens as sets of specimens, these are not currently interlinked. Whether a separate system is needed to handle concepts, or whether this problem could best be solved by building on top of an existing tool (e.g. GBIF) is open to discussion. Global databases tend to have a single consensus taxonomy to maximise the usefulness of their data to scientists who are not taxonomic experts in a given area. As species concepts are hypotheses and subject to continuous change and reinterpretation the formation of a separate repository for these may be a better choice. Methods for managing competing concepts have been developed (e.g. Prometheus [2]) but do not make use of external identifiers.

1.3 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.

Species This work deals with functions that can be applied to sets of specimens considered to be species. The application of these methods to higher ranks is discussed towards the end of the manuscript.

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.

1.4 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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

Algorthmic codes The development of open 'algorithmic codes', based on the current nomenclatural codes, that can be implemented on top of existing systems, could be a productive project for future biodiversity informatics initiatives. Such a system, particularly if developed collaboratively, could reveal (and perhaps prevent) nomenclatural issues in the global databases that are becoming an increasingly important part of the digital biodiversity landscape.

This preliminary study shows that it is possible to represent at least some nomenclatural functions, such as identifying the primary type of a collection of specimens representing a species, as mathematical functions on sets. Further work in this area could develop algorithmic tools for more common nomenclatural functions.

An ideal situation would be where intervention is needed only when special action has been taken within the scope of the nomenclatural codes (e.g. supression of names, neotype designation). If these acts were to be published in a machine-readable format, systems could be developed that are aware of these acts and take them into account during their processing.

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, and is to be considered representative of a species, this is an indication of synonymy.

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

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.

3.2 Consistency of species concepts

When aggregating multiple biodiversity datasets into a single resource there must be some assurance that concepts of the same taxon between them are consistent (i.e. that Species A in dataset 1 is meant to represent the same taxon as Species A in dataset 2). This problem is discussed in [14, 15].

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

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.

sectionAddition 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.

$$x_B = x_A \cup \{x_4\}$$

4.2.1 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.

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

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.

$$x_{new} = x_A | \forall x \in x_A, x \notin x_B$$

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).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

Proliferation of concepts The ease at which a digital system based on these ideas would allow new concepts to be published could potentially lead to a large number of similar concepts. The development of functions that measure the consistency of species concepts, perhaps the one described here, could be used to identify these concepts. The development of consensus concepts, either automatically or (like at present) groups of experts, would benefit the end users of taxonomic information, such as ecologists.

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 history A database such as that proposed could also be used to determine the identification history of individual specimens.

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.

6.1 Beyond the species

While ranks above species have other concepts as their type, they too can be represented as sets. These could be representated as 'sets of sets' (e.g. a genus is a set of species concepts). As species in this model are treated as sets of specimens and the genus is a set of species, the genus can be regarded as containing not only all of the specimens in its constituent species, but also the metadata about the species themselves (through the hypothetical metadata object, or another abstraction). These sets will require different methods to determine some

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1533v2 | CC-BY 4.0 Open Access | rec: 30 Nov 2015, publ: 30 Nov 2015

of their properties (it is not always possible to algorithmically determine the type species of a genus, for example).

6.1.1 Gregg's Paradox

J. R. Gregg [4] showed that treating taxon concepts as sets poses an issue for monotypic taxa within the Linnean hierarchy. If a species is the sole representative of a genus, then these two ranks are defined by the same set of specimens (if the set is considered to extend to all organisms defined by that taxon). Even in the more limited scope of museum specimens used here this problem is likely to still arise. The problem was termed Gregg's Paradox by Buck and Hull [16].

Various solutions to Gregg's Paradox have been proposed that preserve the set theoretic nature of his work [16, 17, 18]. Buck and Hull [16] propose defining sets based on their characters, and arbitrarily defining different levels of the Linnean hierarchy for monotypic taxa. Sklar [18] promotes the inclusion of a unique 'index object' in each set, so that nested monotypic taxa are unique on the basis of the inclusion of additional index sets as the taxonomic rank increases.

If a database based on specimen species concepts is created, it is almost inevitable that these concepts will have associated metadata (authorship, date created, etc). Gregg's Paradox could therefore be resolved through the use of a hypothetical 'metadata object', analogous to the 'index object' of Sklar. This hypothetical object could be used to distinguish between the nested sets of monotypic taxa where this is useful, and ignored when it is not. The metadata described in [2] for the "Nomenclatural Taxon" would provide a suitable metadata object.

7 Acknowledgements

Initially I had based this idea just on species concepts, being somewhat aware that expanding up the Linnean hierarchy would introduce complexities. Rod Page was the inspiration for taking this part of the article somewhat further.

Vince Smith, taxonbytes and Robert Morris provided comments on an earlier draft of this manuscript.

References

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

[2] Martin R Pullan et al. "The Prometheus Taxonomic Model: a practical approach to representing multiple classifications". In: *Taxon* (2000), pp. 55–75.

[3] John R Gregg. "Taxonomy, language and reality". In: *American naturalist* (1950), pp. 419–435.

[4] John R. Gregg. *The Language of Taxonomy.* 1954.

[5] Charissa Sujata Varma. "Beyond Set Theory: The relationship between logic and taxonomy from the early 1930 to 1960." PhD thesis. University of Toronto, 2013.

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

[7] 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.

[8] 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.

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

[10] Sujeevan Ratnasingham and Paul DN Hebert. "BOLD: The Barcode of Life Data System (http://www. barcodinglife. org)". In: *Molecular ecology notes* 7.3 (2007), pp. 355–364.

[11] Edward Baker et al. "BioAcoustica: a free and open repository and analysis platform for bioacoustics". In: *Database* 2015 (2015), bav054.

[12] Cynthia S Parr et al. "TraitBank: Practical semantics for organism attribute data". In: *Semant Web–Interoperability, Usability, Appl an IOS Press J* (2014), pp. 650–1860.

[13] Edward Baker and Yoke-shum Broom. "Natural History Museum Sound Archive I: Orthoptera: Gryllotalpidae Leach, 1815, including 3D scans of burrow casts of Gryllotalpa gryllotalpa Linnaeus, 1758 and Gryllotalpa vineae Bennet-Clark, 1970". In: (Submitted).

[14] David Thau, Shawn Bowers, and Bertram Ludäscher. "Towards best-effort merge of taxonomically organized data". In: *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on.* IEEE. 2010, pp. 151–154.

[15] Nico M Franz et al. "Taxonomic provenance: two influential primate classifications logically aligned". In: *arXiv preprint arXiv:1412.1025* (2014).

[16] Roger C. Buck and David L. Hull. "The Logical Structure of the Linnean Hierarchy". In: *Systematic Zoology* 15.2 (1966), pp. 97–111.

[17] A. F. Parker-Rhodes. “Review of: The Language of Taxonomy”. In: *Philos. Rev.* 66 (1957), pp. 124–125.

[18] A. Sklar. “On category overlapping in taxonomy”. In: *Form and strategy in science*. Ed. by John R. Gregg and F. T. C. Harris. 1964, pp. 395–401.