An Information Model for the Representation of Multiple Biological Classifications

Neville Yoon and John Rose

University of South Carolina, Department of Computer Science and Engineering, Columbia, South Carolina 29208 USA

Abstract. We present a model for representing competing classifications in biological databases. A key feature of our model is its ability to support future classifications in addition to current and previous classifications without reorganizing the database. Data in biological databases is typically organized around a taxonomic framework. Biological data must be interpreted in the context of the taxonomy under which it was collected and published. Since taxonomic opinion changes frequently, it is necessary to support multiple taxonomic classifications. This is a requirement for providing comprehensive responses to queries in databases that contain data reflecting incompatible taxonomic classifications.

1 Introduction

Biological taxonomy provides the organizational framework by which biological information is stored, retrieved, and exchanged. Electronic databases represent a relatively new medium for the storage of biological information, but the concepts and labels that people use to interact with them are the same taxa and names used in the taxonomic literature. It is therefore necessary for biological databases to accurately represent these taxonomic constructs.

Unfortunately there is no single, correct classification that categorizes all organisms for all time. Because of the continuous nature of evolution, taxon delimitation is largely arbitrary; there can be no "correct" classifications, only more or less useful ones. Opinions as to what is more useful vary and frequently change as new specimens are collected, new characters are examined, and new analytical techniques are adopted. Consequently the classifications by which biological information is recorded are often replaced. Furthermore, at any one point in time there may be several incompatible classifications competing for acceptance. Biological databases should be capable of reflecting these competing taxonomic hypotheses. Databases unable to do so risk obsolescence.

As a result of changing classifications, it is often the case that many different names have been applied at different times to a particular group of organisms. Conversely, a single name may have been applied to different sets of organisms. Consequently much biological information is associated in the literature with names that are considered incorrect under current classifications. In order to interpret this information in a modern context, one must know not only the classification assumed by the original publication, but also the nomenclatural and taxonomic changes that relate that classification to the current one.

V.N. Alexandrov et al. (Eds.): ICCS 2001, LNCS 2073, pp. 937–946, 2001.

Most biological database designs ignore this complexity by using only taxon names to identify taxa [1,2,3,8,10,11,12]. With this approach incompatible classifications cannot be maintained simultaneously because there is no way to distinguish among different taxon concepts with the same name [5,6]. With careful editing by taxonomic experts, such a database can be made to conform to a particular set of complimentary classifications. However, the information stored within cannot easily be made to conform to any conflicting classification, making the database inflexible with respect to changes in taxonomic opinion. Worse, if information is widely compiled from the literature and stored uncritically by name, incompatible taxon concepts will become confounded resulting in a database that does not conform to any classification. A more flexible approach to managing taxonomic information is required.

A system for managing multiple classifications will have to satisfy at least two criteria: First, information must be indexed by specific taxonomic interpretations of names (named taxon concepts) rather than by names alone. Second, there must be a way to represent relationships of shared content among named taxon concepts. This allows information to be aggregated both within classifications through the taxonomic hierarchy and among classifications according to overlap in taxon boundaries.

In the following sections we briefly review four previously proposed designs for the management of multiple classifications in biological databases. Each of these meets the above criteria at least in part. However, none is completely satisfactory for use in databases of descriptive, non-specimen-based biological information intended for non-taxonomists. We derive from them an information model that we believe to be more appropriate for these types of systems.

2 Prior Approaches to Modeling Multiple Classifications

To our knowledge, four information models designed to accommodate incompatible classifications have been published or otherwise made publicly available. These are the Association of Systematics Collections (ASC) draft datamodel [4], the HICLAS model [5,9,15], the International Organization of Plant Information (IOPI) model [6,7], and the Prometheus model [12]. We will discuss each of these using an extended entity-relationship (ER) data modeling vocabulary and notation (Fig. 1).

In the ASC model, both taxon names and taxon concepts are explicitly represented as entity types (Fig. 1). The TAXON CONCEPT (TC) entity type represents a taxonomic circumscription with no inherent name. The many-to-many relationship between names and taxon concepts is resolved through the use of an associative entity type, TAXON-NAME-USE. This traditional ER approach satisfies our first criterion of representation of multiple taxon concepts for the same name. The taxonomic placement of a taxon underneath a superior taxon in a classification is represented by an entity of the recursively associative entity type, TAXONOMIC RELATIONSHIP . This structure implements a network of TCs in which each TC is the root of a taxonomic sub-tree, the leaves of which are species-level

TCs. The circumscription of a TC is the full set of terminal TCs in its sub-tree. This model permits the comparison of TCs according to shared content and thus allows aggregation of information both within and among classifications. However, separating out the complete set of TCs and associated TAXON NAME USEs for a particular classification requires additional constructs not presented in the ASC draft datamodel.

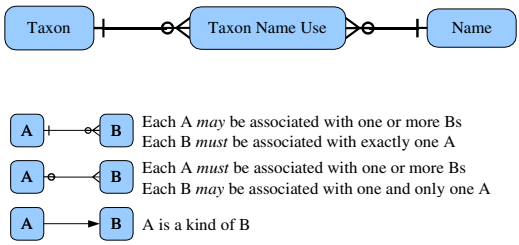

Fig. 1. A partial entity-relationship diagram of the many-to-many relationship between taxa and names

The HICLAS and IOPI models represent circumscriptions implicitly by association with a publication rather than explicitly with an entity type. This approach models taxon concepts as a relationship between a taxon name and a publication and is analogous to the TAXON-NAME-USE entity type of the ASC model without the associated TC. This is called a *taxon view* in the HICLAS model and a *potential taxon* in the IOPI model. We will use the term *name-use* to represent the generalized concept. A name can be used in different contexts in different publications, but by linking information to the name-use, these different taxon concepts are not confused. However, since circumscriptions are not directly represented, the synonymy of different names applicable to the same taxon must be recorded directly rather than by association with overlapping circumscription. The two models take different approaches to this problem.

In the HICLAS model both classificatory relationships and derivational relationships among taxon views are represented. Classifications are represented by trees consisting of taxon views connected by classification relationships. The derivational history of a taxon concept is represented by a set of operation trees that trace the previous taxon concepts that have been split, merged, moved, or accepted to create the taxon view in question. Relationships of shared content can be inferred from these operation trees.

The HICLAS model provides a simple mechanism for the simultaneous maintenance of multiple classifications, for the structural comparison of different classifications, and for tracing the histories of taxon concepts. However, scientific names themselves are treated in a simplified manner that does not allow a complete representation of purely nomenclatural relationships.

The IOPI approach uses an associative entity type called STATUS ASSIGN-

MENT to represent the relationships among potential taxa. Different types of status assignments are used to indicate nomenclatural relationships and relationships of shared content.

The position of a potential taxon in a classification is represented through a simple recursive relationship on the POTENTIAL TAXON (PT) entity type. The restriction that a POTENTIAL TAXON can have only one taxonomic position results in considerable proliferation of PTs.

A sub-tree composed of all of the descendents of a particular PT will often contain PTs representing different circumscriptions for the same taxon name. Berendsohn [6] proposes a simple ranking of taxonomic reference works to resolve these conflicts. When conflicts are found in the generation of a classification tree from the database, those established by the preferred reference works are chosen for presentation. This seems to be a limited and unwieldy method for reconstructing alternative classifications from the database.

The Prometheus model represents biological taxonomy more accurately than any other model published to date. Aspects of biological nomenclature are carefully separated from those related to circumscriptions and classifications to reflect the way that taxa are actually created and named in taxonomic practice.

Taxonomic names are represented by the NOMENCLATURAL TAXON (NT) type. The NT is the combination of a taxon name, a rank, a superior NT for names at species-level ranks, a publication, and a nomenclatural type, which may be a specimen or another NT. Official declarations of nomenclatural status that may affect the priority of a name may also be assigned to NTs. Taxon concepts are represented by CIRCUMSCRIBED TAXON objects which are the combination of either an NT or an informal name, a circumscription, a rank, an author, and possibly a publication.

The Prometheus model is unique in that relationships of shared content are not represented declaratively, but rather are derived from rigorous and detailed representations of taxon content. Taxon content is represented by the CIRCUMSCRIPTION type. A CIRCUMSCRIPTION object specifies a complete set of specimens included in a taxon. For published taxa, the circumscription consists of all the specimens cited in the published description. For experimental taxa, the specimens include all those deemed to belong by the practicing taxonomist. Relationships of synonymy by shared content are derived by directly comparing circumscriptions among CTs. Furthermore, the nomenclatural principles of priority and typification can be applied algorithmically to validate the assignments of taxonomic names to CTs.

This strict, specimen-based approach to comparing taxa is very powerful when complete sets of included specimens are available. In this case all objective relationships of shared content can be found. However, in many cases complete sets of specimens are unavailable or the effort of compiling them exceeds the abilities of a database team. Furthermore, the Prometheus model prohibits the extrapolation of taxonomic inference beyond that directly supported by the specimen content of circumscriptions. The result of these restrictions is that in some information systems, large amounts of information stored within may not

be interpreted with respect to specific classifications and cannot be aggregated according to suspected relationships of shared content. This is not a criticism of the Prometheus approach, which is logically correct. Nonetheless we believe these restrictions may be excessive for many information systems.

3 The PeroBase Model

Our motivation for developing a new information model is to provide the taxonomic framework for PeroBase, an encyclopedic database that manages information on the biology of peromyscine mice for users ranging from the interested layperson to specialists in *Peromyscus biology*. Peromyscine taxonomy has undergone a few major revisions and is still in flux, so we need a model that can represent incompatible classifications. In studying the four information models discussed above, we concluded that the ASC and HICLAS models were not sufficiently complete to handle our nomenclatural information. We are most impressed by the Prometheus model, but do not have the resources to compile the needed specimen lists. Furthermore, we desire the ability to organize information according to a fully-connected taxonomic hierarchy that integrates numerous lower level classifications. This is prohibited in the Prometheus model. We therefore have derived a model that we believe to be more appropriate for our purposes by relaxing the restrictions of the Prometheus model and incorporating ideas from the other models.

3.1 Nomenclature

Scientific taxon names are represented by the NOMENCLATURAL TAXON (NT) entity type of the Prometheus model with minor modifications primarily to accommodate differences between the International Code of Zoological Nomenclature (ICZN) [13] and the International Code of Botanical Nomenclature [14](Fig. 2). As in Prometheus, an NT is the combination of a name element, a taxonomic rank, a taxonomic placement for species ranked NTs, a publication, and a name-bearing type. Our NT differs primarily in the way nomenclatural status is assigned. These assignments are made through the NOMENCLATURAL STATUS ASSIGNMENT entity type analogous to the way nomenclatural status is assigned to potential taxa in the IOPI model. NOMENCLATURAL STATUS ASSIGNMENTs are used to record formal published acts that affect the application of the principle of priority in determining valid names. This usually involves suppression of senior synonyms or homonyms in favor of junior names in prevailing use. In these cases, the preferred NT is also associated with the NOMENCLATURAL STATUS ASSIGNMENT. The publications in which these assignments are made are recorded as associations of NOMENCLATURAL STATUS ASSIGNMENTs with PUBLICATIONS. When status assignments affect the priority of groups of secondary homonyms or heterotypic synonyms, the suppression is only effected while the types of the NTs are considered to fall within the same circumscription. When this is not the

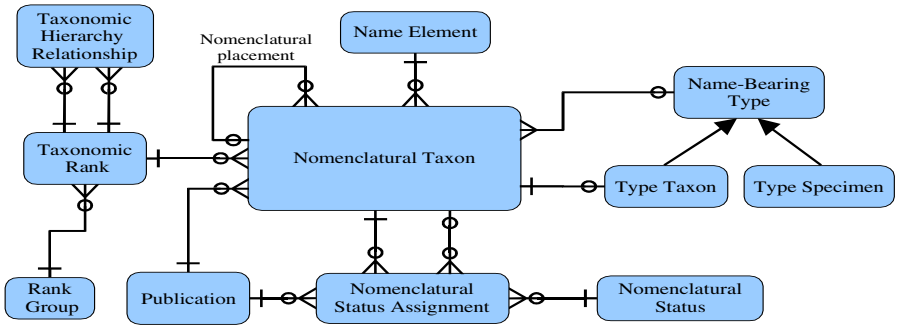

Fig. 2. The NOMENCLATURAL TAXON and associated entity types

case, these status assignments are to be ignored by the system for the purpose of determining valid names.

Every NT has a taxonomic rank represented by an associated TAXONOMIC RANK entity. The ICZN defines three groups of taxonomic ranks over which it claims authority: the species-group, the genus-group, and the family-group. Nomenclatural rules apply differently to nomenclatural taxa depending on the group to which they belong. These taxonomic rank groups are represented in the PeroBase model with TAXONOMIC RANK GROUP entities. Taxonomic ranks are ordered into taxonomic hierarchies to which classifications adhere. The names and order of many ranks are considered obligatory by convention; however, most ranks are optional and taxonomists are free to insert any number of additional ones. This network of superior-subordinate relationships between taxonomic ranks is represented through the TAXONOMIC HIERARCHY RELATIONSHIP entity type. The particular associations represented constrain the taxonomic hierarchies that are permissible in the database.

An NT is associated with a single NAME ELEMENT, but is intended to represent a full scientific name. For taxa above the species rank, the two are the same. However by the principle of binominal nomenclature, the name element of taxa of species and subspecies ranks must be prefixed by the full name of the taxon under which they are placed. A recursive relationship implements the nomenclatural placement of species group NTs permitting the composition of their full names.

Every nomenclatural taxon of family-group rank or below has an actual or potential name-bearing type according to the ICZN. Correspondingly in PeroBase, every NT of family-group rank or below must be associated with a NAME-BEARING TYPE (NB-TYPE). A NB-TYPE may represent either a type specimen for species group taxa, a type species for genus group taxa, or a type genus for family group taxa. To model this, two subtypes of NB-TYPE are used: TYPE SPECIMEN and TYPE TAXON. A TYPE SPECIMEN need not have any attributes in the database, but a TYPE TAXON must have an additional relationship with

an NT that specifies the taxon that is the type species or type genus. Note that NB-TYPE represents only the real or potential existence of a type. While information about real type specimens could be attached to NB-TYPEs, this is not necessary. This "virtual" type concept serves primarily as a reference tool for the determination of synonym and homonym relationships among taxa. This corresponds to the use of "dummy" types in the Prometheus model.

3.2 Classification

The association of a name with a particular circumscription is represented by a TAXONOMIC TAXON (TT) entity (Fig. 3). Both classifications and circumscriptions in the PeroBase model are implemented as trees of TTs whose edges are TAXONOMIC RELATIONSHIP entities. TAXONOMIC RELATIONSHIPs may represent either the classification of a TT under a TT of superior rank (placement), or the inclusion of a monotypic species-level TT in the circumscription of a polytypic species-level TT (inclusion).

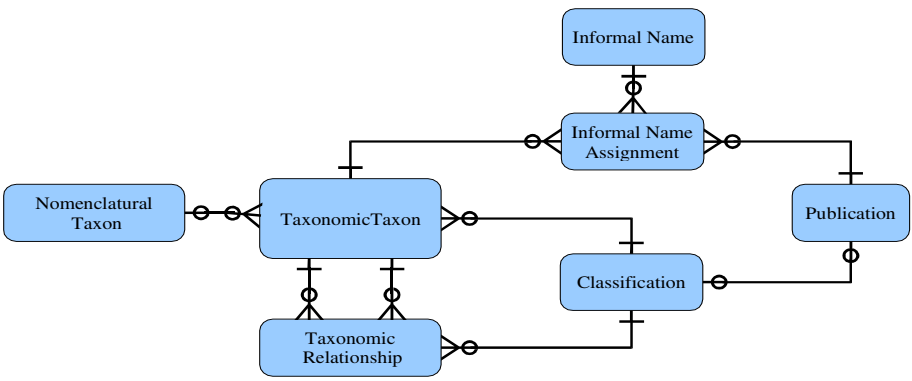

Fig. 3. The TAXONOMIC TAXON and related entity-types

Every TT and every TAXONOMIC RELATIONSHIP is created in the context of a particular classification and must be associated with a CLASSIFICATION entity. The CLASSIFICATION entity is merely a label for use in grouping TTs and TAXONOMIC RELATIONSHIPS to form classification systems. The classification system itself is represented by a tree of TTs whose edges are all placement TAXONOMIC RELATIONSHIPs linked to the CLASSIFICATION.

A circumscription is represented as a set of all NB-TYPES included within the boundaries of a TT. This set is obtained for a TT by finding the species-level TTs that are leaves in any classification tree rooted at that TT and adding for each polytypic leaf TT all monotypic species-level TTs that are related through inclusion type TAXONOMIC RELATIONSHIPs.

A classification may include TTs established in previous classifications as long as neither their associated NTs nor their full set of included types are changed. For example, a classification that rearranges existing species into new subgenera without changing the contents or generic placements of those species can use existing TTs for those species. If an existing taxon is moved from its current position to another, the contents of both its former and new parents have changed, and new TTs are required to represent the new circumscriptions. However, if the former and new parent taxa are themselves both placed under the same TT, no new TT is required for that common parent since its circumscription has not changed. A change in the rank of a taxon requires a new NT and therefore a new TT. A change in the placement of a taxon of species level rank results in a mandatory name change, requiring a new NT and therefore a new TT as well.

The TT representing a newly described species or subspecies may be placed under a previously existing TT even though the circumscription of the parent has expanded to include the new type. Without this exception, the addition of a new type would require new TTs for all superior taxa on all paths to the roots of all classifications that include the new taxon [4]. The different circumscriptions for the same TT are still separable since the original and new classificatory relationships are associated with different CLASSIFICATIONs. The full circumscription for a TT, including all subsequent additions, can be obtained for any point of view by adding the relevant classificatory relationships of subsequent classifications to those of the original definition.

3.3 Synonyms

Homotypic, heterotypic, and most *pro-parte* synonyms for a particular taxonomic taxon can all be found algorithmically as in the Prometheus model. Homotypic synonyms are simply different nomenclatural taxa of the same rank with the same name-bearing type. Heterotypic synonyms for a particular taxonomic taxon are the names of all of the TYPE TAXON OR TYPE SPECIMEN entities included in the taxon's circumscription. The correct name for a particular taxonomic taxon can be found from among its heterotypic synonyms by identifying the one that was published first after eliminating those specified as invalid in associated nomenclatural status assignments.

Most *pro-parte* synonyms for a particular taxonomic taxon can be found by finding all taxonomic taxa in the same rank group that contain any of that taxon's included type specimens. As Pullan, et. al. [12] point out, this will not identify all *pro-parte* synonyms because some taxa may share specimens without sharing any types. We suspect that this level of resolution will rarely be necessary for taxon-based information systems; however, such instances of overlap could be indicated with a new type of taxonomic relationship if so desired.

3.4 Determinations: Assigning Data to Taxa

Descriptive information can be applied to both NTs and TTs. Very often in the biological literature descriptive information is attributed to a taxon identified

by name without specification of the classification assumed. This information is therefore name-based only and can be assigned with full confidence only to NTs [12]. The name under which descriptive information is originally published is recorded in the PeroBase model through an association with an NT via a NOMENCLATURAL TAXON DETERMINATION. We agree with Berendsohn [5] that in many cases information not derived from identified specimens may still be attributable to particular taxon concepts with reasonable confidence. In the PeroBase model, data is associated with a TT through a TAXONOMIC TAXON DETERMINATION entity. The person responsible for the determination and the date on which it was made are both recorded with the determination so that corrected assignments can be made without erasing the history of previous assignments.

4 Conclusion

Accurate information models of biological taxonomy are difficult to design due to the inherent complexity of taxonomic data and nomenclature. As a result, most biological databases have been developed from overly simplistic representations of taxonomy. This is unfortunate because over time the information in these databases will no longer reflect current taxonomic opinion. Keeping these databases up-to-date will require periodic large-scale overhauls, work that could have been largely avoided through the use of a more flexible taxonomic data model. Databases intended to manage descriptive biological information for non-taxonomists tend to have the most simplistic models of taxonomy, yet they may have the greatest need for taxonomic flexibility.

Four models have previously been proposed to permit the simultaneous management of multiple biological classifications, a primary requisite for adaptable biological databases. While sharing many similarities, each of the models has taken a unique approach to solving the problem. These differences reflect slightly different priorities and intended uses and the models have succeeded to various degrees. In our opinion, the most accurate and powerful representation of taxonomy to date is the Prometheus model [12], but large amounts of specimen-level data must be compiled to realize the full strengths of that model. For many kinds of biological information systems this may not be practical. A new model is needed to approach that level of taxonomic flexibility in databases that deal with information above the specimen level of resolution. We have presented a new model of taxonomy derived from the Prometheus model for this purpose. We believe the new model offers the best compromise so far proposed between accuracy and flexibility on one hand and practical applicability on the other.

Our model has been developed to serve as the taxonomic framework for PeroBase, a multi-disciplinary descriptive database of information pertaining to peromyscine mice. Current work involves the implementation of the model with taxonomic information drawn from the literature for this group.

Acknowledgement

This work was funded by NSF grants DBI-9723223 and DBI-9807881.

References

1. Allkin, R., Bisby, F.A. Databases in Systematics. Academic Press, New York (1984)
2. Allkin, R., White, R.J.: Data management models for biological classification. In: Bock, H.H. (ed.): Classification and Related Methods of Data Analysis. Elsevier Science Publishers B.V., North-Holland. (1988) 653–660
3. Allkin, R., White, R.J., Winfield, P.J.: Handling the taxonomic structure of biological data. Mathematical and Computer Modelling 16:6/7 (1992) 1–9
4. Association of Systematics Collections: An Information Model for Biological Collections (Draft) March 1993 version: Report of the Biological Collections Data Standards Workshop August 18–24, 1992. Association of Systematics Collections. Available from: gopher://kaw.keil.ukans.edu/11/standards/asc (1993)
5. Beach, J.H., Pramanik, S., Beaman, J.H.: Hierarchic Taxonomic Databases. In: Fortuner, R. (ed.): Advances in Computer Methods for Systematic Biology. Johns Hopkins University Press, Baltimore (1993) 241–256
6. Berendsohn, W.G.: The concept of "potential taxa" in databases. Taxon **44** (1995) 207–212
7. Berendsohn, W.G.: A taxonomic information model for botanical databases: the IOPI model. Taxon **46** (1997) 283–309
8. Blum, S.D. (ed.): Guidelines and Standards for Fossil Vertebrate Databases: Results of the Society of Vertebrate Paleontology Workshop on Computerization. November 1–4, 1989; Austin, Texas (1991)
9. Jung, S., Perkins, S., Zhong, Y., Pramanik, S., Beaman, J.: A new data model for biological classification. CABIOS **11**:3 (1995) 237–246
10. Krebs, J., Kaesler, R., Chang, Y-M, Miller, D., Brosius, E.: PaleoBank: a Relational Database for Invertebrate Paleontology: Data Model. Paleontological Institute, U. Kansas. http://history.cc.ukans.edu/~paleo. (1996)
11. Pankhurst, R.J.: Taxonomic databases: the PANDORA system. In: Fortuner, R. (ed.): Advances in Computer Methods for Systematic Biology. Johns Hopkins University Press, Baltimore (1993) 230–240
12. Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C., Hyam, R.: The Prometheus taxonomic model: a practical approach to representing multiple classifications. Taxon **49** (2000) 55–75
13. International Commission on Zoological Nomenclature: International Code of Zoological Nomenclature (4th ed.). U.California Press, Berkeley (1985)
14. Greuter, W., Barrie, R.R., Burdet, H.M., Chaloner, W.G., Demoulin, V., Hawksworth, D.L., Jorgensen, P.M., Nicholson, D.H., Silva, P.C., Trehane, P., MacNeill, J. (eds.): International Code of Botanical Nomenclature (Tokyo Code). Regnum Veg (1994) 1–389
15. Zhong, Y., Jung, S., Pramanik, S., Beaman, J.H.: Data model and comparison and query methods for interacting classifications in a taxonomic database. Taxon **45** (1996) 223–241