OMICS A Journal of Integrative Biology
Volume 10, Number 2, 2006

Standard Data Model Representation for Taxonomic Information

J. KENNEDY,[1] R. HYAM,[2] R. KUKLA,[1] and T. PATERSON[3]

ABSTRACT

The names used by biologists to label the observations they make are imprecise. This is an issue as workers increasingly seek to exploit data gathered from multiple, unrelated sources on line. Even when the international codes of nomenclature are followed strictly the resulting names (Taxon Names) do not uniquely identify the taxa (Taxon Concepts) that have been described by taxonomists but merely groups of type specimens. A standard data model for exchange of taxonomic information is described. It addresses this issue by facilitating explicit communication of information about Taxon Concepts and their associated names. A representation of this model as a XML Schema is introduced and the implications of the use of Globally Unique Identifiers discussed.

This paper is part of the special issue of OMICS on data standards.

INTRODUCTION

TAXONOMIC IDENTIFICATION is emerging as a significant problem for the integration and comparison of diverse datasets for analysis across all fields of biology from genomics to ecology. Therefore, the availability, exchange, and interpretation of taxonomic information (e.g., species check lists, distribution data, and identification data) is of critical importance to a wide range of biologists. This information is provided by a number of global and local taxonomic database services. These databases hold records, often based on valid scientific names for species, according to their own model of what constitutes a taxonomic "entity" or concept (i.e., a species or higher level taxon). They typically model a single view of taxonomy, whilst making some attempt to relate their concepts to synonymous names or concepts. For example, annotations of Genbank, DNA sequences, (Genbank, 2006), typically label the source species according to the NCBI Taxonomy, (NCBI, 2006). Whilst specifically disclaiming any "taxonomic authority," NCBI attempts to provide a single consensus view on taxonomy, and represent name alterations and "corrections" by encoding synonym relationships for use by their search engines (e.g., example the genus *Fugu* has recently been "renamed" *Takifugu*), which does not deal with complexity of semantics in taxonomic data.

Taxonomy is an investigative science that seeks to categorize and classify biological organisms into a hierarchy of taxonomic groups, typically reflecting the evolutionary relationships between these groups (or

[1]School of Computing, Napier University, Edinburgh, Scotland, United Kingdom.
[2]TDWG, Royal Botanic Garden, Edinburgh, Scotland, United Kingdom.
[3]Roslin Institute, Roslin, Scotland, United Kingdom.

taxa, singular taxon). Taxonomic classifications represent alternative and evolving hypotheses rather than static descriptions of absolute taxa and represent an opinion, according to one authority, at a given time. Consequently there is no single universally correct or consensus Taxonomy. Classifications, and the definition of constituent taxa (including species), vary over time and between investigators, and reflect many variables—including the geographical range of a study, interpretation of collected specimens, the fossil record, morphology, genetics and molecular phylogeny. New classifications (taxonomic revisions) may arise following more detailed study of specimens, the discovery of new taxonomic information, or indeed following the description of new species and groupings. The increasing use of DNA sequence comparison as a tool to analyse phylogenetic relationships is accelerating the rate of taxonomic revision and is unlikely to stabilize in the foreseeable future.

However, biologists and other users of taxonomic information require a system that provides unambiguous taxonomic identifications for organisms (i.e., as individuals of some named taxonomic group). On the one hand, it is essential for life scientists to accurately record their specimens of study or observation, in order to allow meaningful interpretation and reuse of data sets. On the other hand, outside the realm of academic study, unambiguous taxonomic identifiers are also essential for a wide range of regulatory bodies, and local and international organizations that require to legislate about, record, enumerate, protect, or otherwise accurately reference taxa—typically, at species level.

These needs are becoming more self-evident with the ever-increasing volume of electronically curated biological data from both laboratory and field study, and with increasing concern over the measurement and protection of global biodiversity. Consequently, a model of taxonomic information must be able to represent the precise meaning of taxonomic entities created and analyzed by taxonomists, whilst also allowing other users of taxonomic information to unambiguously reference these taxonomic concepts as identifications of the organisms that they are describing. The development of a consensual and inclusive model of taxonomic information will facilitate the capture, comparison and resolution of taxonomic data across disparate sources for these differing user communities.

The International Working Group on Taxonomic Databases (TDWG, 2006) was formed to establish international collaboration among biological database projects so as to promote the wider and more effective dissemination of information about the world's heritage of biological organisms for the benefit of the world at large. Part of its work is the development, adoption, and promotion of standards and guidelines for the recording and exchange of data about organisms. It has several subgroups involved in the development of standards for taxonomic information including Taxonomic Names/Concepts (TCS, 2006), Biological Collections Data (ABCD, 2006), and Structure of Descriptive Data (SDD, 2006). At the TDWG annual meeting in 2003, the importance of the development of a common mechanism for the providers of taxonomic information to exchange information with other providers and users of varying expertise in taxonomy was recognized. This paper presents the work of the Taxonomic Names/Concepts subgroup in the development of a model for taxonomic information, the Taxonomic Concept Schema (TCS) that seeks to accommodate the realities of taxonomic practice in such a way that the taxonomic information can be used not only for the purposes of taxonomic study but to provide a reference mechanism allowing life scientists and others to unambiguously record the taxonomic identifications necessary in their own field of study. The schema has followed an iterative design, with changes to the model being incorporated following presentation to and feedback from the community. The current version, TCS version 1.01, was accepted as a TDWG standard following the 2005 TDWG meeting.

The three central components of our model are Taxon Concepts (the taxonomic groupings as defined and described by taxonomists in scientific publications, which may increasingly be provided by authoritative sources and integrators of taxonomic information on the Internet and elsewhere); Taxon Names (the formal scientific names applied to Taxon Concepts, according to the rules of Nomenclature) and Taxon References (in which the intention is to refer to some existing Taxon Concept, without asserting that this reference to the concept constitutes a new definition, in other words, a published Taxon Concept in itself). Of course much biological data may only include a scientific name as reference, in which case it cannot automatically and unambiguously be resolved to an actual Taxonomic Concept, but can be considered to resolve to a Nominal Concept, in other words, a concept that consists of that Scientific Name but with no explicit definition.

An implementation of this model as an XML schema seeks to provide a structured format for the exchange, integration, and comparison of taxonomic information according to an XML data structure between the providers or authors of taxonomic concepts and between less specialist users of these taxon definitions. The XML schema aims to be inclusive of all alternative representations of taxonomic concepts and has been accepted as a TDWG standard for data exchange. This will facilitate the development of taxonomic name/concept resolution services and allow the meaningful interpretation, integration, and comparison of biological datasets across disciplines. The model may be implemented in other technologies in the future.

TAXONOMIC INFORMATION

Biologists use scientific names to label the taxa described in their data; however, these names are not unique identifiers for the taxon concepts that are produced by the classification process. Precise rules of taxonomic nomenclature control the application of formal scientific names to taxa by taxonomists (Greuter et al., 2000; ICSP, 1990; ICTV, 2000; Ride et al., 1999). These names are derived from the first published name, which has its primary "type specimen" included within a particular taxon circumscription (the nomenclatural rule of priority ICZN Article 23 & ICBN Article 11). As a direct consequence of these rules, the exact same scientific name will be used in alternative classifications to represent potentially contradictory views or circumscriptions of taxonomic groupings (Fig. 1); furthermore, very similar taxonomic concepts can hold different valid scientific names when subtle differences in taxonomic opinion alter the placement of type specimens (Fig. 2) or places them in different higher taxa. So that whilst scientific names *per se* remain stable over time, being attached to their "type specimens"—these names are not unique identifiers for taxon concepts—and the "meaning" of any name is context dependent and can only truly be interpreted according to which taxonomic classification is being referenced.

The issues of ambiguity surrounding taxonomic classification and naming are well understood by expert taxonomists who in addition to defining new Concepts as part of taxonomies may record asserted relationships between named concepts in alternative published taxonomies. For example a taxonomist might assert that their named concept shares a name with a separate concept in another taxonomy, or is equivalent to some other concept, which might have the same or different name, or is included within some other defined concept. In other words, the actual Taxon Concepts may have relationships that are separate from, and often not reflected in, formal nomenclatural relationships which only track the placement of type specimens.

In order to distinguish between concepts in alternative taxonomies it is necessary to capture a unique identifier for the concept, which generally will be resolvable to the citation where the definition appeared. This is generally expressed as a record of the full scientific name applied (often including the original authorship of that name or combination), together with the authorship of the source classification, typically denoted as "*sec*. citation" (*secundum*, "according to," from Latin: after, following) or *sensu* (from Latin: perception or opinion). Examples of two separate Taxon Concepts sharing the same scientific name would be *Carya ovata* Gleason *sec.* Gleason 1952 versus *Carya ovata* Gleason *sec.* Stone 1997. The actual definition of the referenced concept will be resolvable if the citation provides a detailed description of the taxon.

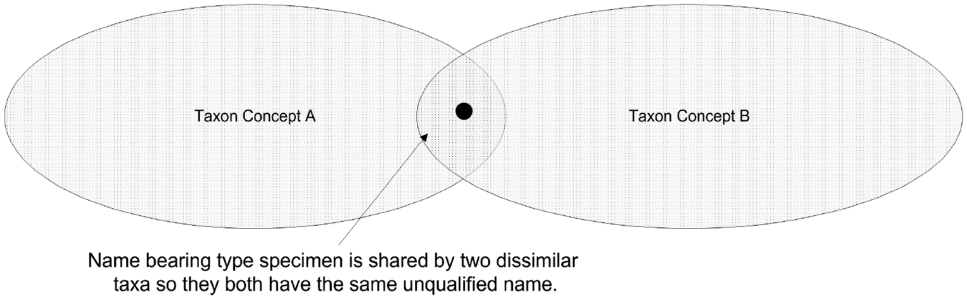

FIG. 1. Dissimilar taxon concepts can share a single unqualified name.

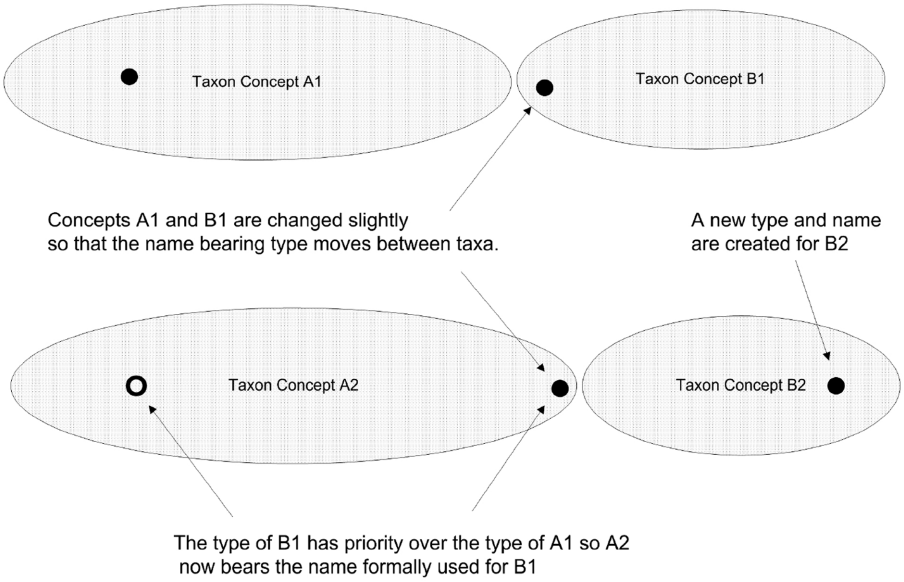

FIG. 2. Slight changes in taxon concepts may result in name changes that do not reflect concepts.

Although, the taxonomic community has its own practices and requirements for advancement of taxonomic theory, however, the fundamental output of taxonomic research, the classification of organisms (i.e., as taxa: species, genera, families), provides the reference framework for all biological disciplines, and must therefore be accessible to these non-expert users.

TAXONOMIC NAMES AS NON-UNIQUE IDENTIFIERS OF TAXON CONCEPTS

To reiterate, although a valid and accepted name must be unique within a nomenclatural code, it is necessarily reused across published taxonomic views with different definitions, and therefore cannot represent a unique identifier of a Taxon Concept but only a collection of type specimens. Furthermore names can be corrected or changed independently of the taxonomic process, so that the original published name in a taxonomic revision may be altered if later deemed to be misapplied according to the rules. A common example of this is when a species epithet is discovered not to be of the same gender as the genus in which it is placed (e.g., ICBN article 60.1).

A detailed model of Taxonomic Names as represented across the biological kingdoms according to the formal rules has been produced for TCS with the help of nomenclatural experts. The model of Scientific Names can represent the separate components of the Name including the atomic components of the name and authorship, details of nomenclatural history and typification. Minimally, a Scientific Name can just be a string representation of the "label," in other words, a normalized version of a full name, including nomenclatural citation (authors/year) in the form mandated by the relevant code.

Where a full scientific name is used with attribution to the authors of the name *and* of the taxonomic revision, this represents a clear identifier for a concept. However, this level of detail is rare out with specialist taxonomy. Most users and creators of biological data are not expert in taxonomy, and the names or labels that they use to refer to specimens and organisms include *ad-hoc* labels, common names or the (sometimes approximate or inaccurate) scientific name for a species or higher taxonomic group. A "name" might be represented in a variety of forms, each of which might have different types of explicit or implicit definitions, and might be considered to be more or less accurate or precise in terms of the definition of that name.

Currently there is no widely adopted system which allows accurate reference to taxonomic concepts to be recorded in biological data, and many biologists and others are naïve as to the implications of relying on identification by scientific name alone. Nomenclatural precision is often confused with taxonomic pre-

cision. Without a system that allows explicit resolution of taxonomic identifications against verifiable taxonomic concepts (according to published classifications), the unambiguous integration and comparison of biological data is not possible, even where a perfectly correct, unqualified scientific name is recorded. Even when preferred taxonomies or indexes of names are enforced in biological databases (such as the DNA sequence databases) there is little support for true taxonomic comparison between alternate taxonomic views, other than by simple name matching.

TAXON REFERENCES TO DESCRIBED TAXON CONCEPTS

In a wide sense any recorded use of a taxonomic name might represent a potential taxon concept, that is, represent a particular author's "concept" or opinion of what that taxon is (Berendsohn, 1995). Such an interpretation would logically lead to every "name usage" representing a potential taxon concept. Such an explosive interpretation of Taxon Concept would lead to an unmanageable "inflation" of records. A process of "taxonomic monitoring" (Berendsohn, 1995) to limit concept inflation could be introduced to limit acceptance of valid taxon concepts to those created by taxonomists (and documented in refereed taxonomic publications, monographs, identification guides); however, imposition of central control of who is competent to define Taxon Concepts—something that could potentially restrict scientific freedom—would be problematic. Ultimately concept inflation is unlikely because most workers do not wish to create new versions of taxa but simply to reference existing taxa. They actively seek to use common, stable terms so that the data they are labelling can be related to data from other sources. The emergence of online taxonomic resources, which provide taxonomic names according to the provider's own aggregated view of taxonomy will contribute to this damping effect. Indeed practising biologists are increasingly likely to use such databases, rather than the taxonomic literature, as the source of the taxonomic references for their data. The providers who supply well defined, easily referenced and stable sources of concept IDs will be used. It appears likely that the IDs provided by a handful of such services will become the *lingua franca* of the biological community in time. The 'taxonomic monitoring' suggested by Berendsohn will be produced by default.

The taxonomic identifications in datasets carry an inherent quality score; that is, they might be very reliable if competently performed and recorded according to recognized field guides, or they may be of lower quality, perhaps identifying only to a general taxonomic group not a specific species or subspecies, or they may use common or incorrect names, and they may not detail the identification context or field guide followed. No information model can improve data quality in these circumstances—but it can provide a mechanism to make the problems of imprecise identification more explicit. For example, if only a scientific name is recorded in the absence of a taxonomic context (i.e., the name lacks a *sensu*), in our model this would only point to a Nominal Taxon Concept, which consists solely of a Taxon Name, and would be of little use for taxonomic comparison and be recognized as of low data quality although it could lead to discovery of a set of Taxon Concepts that could be used, with caution, in further analysis.

USER COMMUNITY FOR TAXONOMIC INFORMATION

As discussed in the introduction, the "non-expert" user community for taxonomic information is virtually unlimited. From molecular biologists who must annotate DNA sequences with "species" of origin; to field ecologists who report on species diversity and abundance; through to inter-governmental bodies who compose lists of endangered and protected species. The importance of meaningful taxonomic annotation becomes apparent once users want to reinterpret, re-use or aggregate earlier data; for example to compile a number of field studies performed across a geographical range and timespan. Without full documentation of what the taxonomic identifications in the original datasets *mean* (i.e., without record of the actual taxon concept), there is no way to ensure that data on a species identified as "X" in Hungary collected in 1952 relates to data on a species called "X" collected in Peru in 2005.

Clearly, the non-expert user needs some simple label or identifier for the Taxon Concepts that they are identifying against, and we have explained above how the use of Scientific Name alone cannot provide this unambiguous identifier. Ideally, a user would record the identification of a specimen to a Taxon Concept in a recent monograph of the group. Failing this, the user may determine the specimen to a Taxon Concept in a field guide such as Stace (1999) or a checklist such as those provided by the International Taxonomic Information System (ITIS, 2006), which in turn references monographic treatments. It could become a minimum requirement of Scientific Journals, Database Repositories, and Research Funders that unambiguous taxonomic identification and attribution is included in published biological data. This will be facilitated by the implementation of the TCS based data exchange standards.

This is particularly important for legislative authorities. If the accepted taxonomy of a group has changed between passing of a law and its application how does the entity the law was passed against relate to the currently accepted entity? Statutory authorities are increasingly having to track the evolution of their national species taxonomies through time. This is only possible with the notion of Taxon Concepts and can not be achieved with taxon names alone.

DEVELOPMENT OF GLOBALLY UNIQUE IDENTIFIERS

The correct representation of a scientific name can be non-trivial—and in any case does not represent a unique identifier of a Taxon Concept, therefore any system that simplifies the reference of a valid Scientific name or a published Taxon Concept will improve the quality of Taxonomic Information referenced in Biological datasets (improving reliability, accuracy, and interpretability as well as ease of use).

It is an aim of TDWG to promote the reuse of valid Taxon Names and Taxon Concepts through the TCS data exchange standard. Clearly, the most efficient means to achieve this is through the reuse of previously defined taxonomic information. At present, each taxonomic database has their own internal and sometimes external identifiers for their taxon names or concepts (e.g., the TSN numbers used by ITIS). These are not represented in the TCS transfer schema, as there is no guarantee that any given database ID would map uniquely to a TCS concept nor remain stable over time. The TCS schema was devised to allow exchange of concepts together with their definitions, and could be used to represent concepts stored in any global repository or local cache. To provide a stable and resolvable identifier for these concepts it would be highly desirable if GUIDs for taxon concepts were adopted. These could be assigned and maintained locally (by data owners) or globally according to agreed international policies, and would provide a stable reference to a taxon concept as represented according to TCS. Once in place this mechanism would simplify the mark-up of biological data, according to available concept definitions, and could assist data retrieval based on concept comparisons.

Availability of GUIDs would also help reduce the redundancy and overlap between different data providers who currently reproduce alternative representations of the "same" concept. Under the auspices of TDWG, prototype GUIDs servers are being developed in conjunction with SEEK, GBIF and the wider biological community to determine the feasibility of providing GUIDs using LSIDs for taxon concepts, and other stable concepts such as Taxon Names, Publications and Specimen Identifiers. The first concepts that will be available for representation and caching as TCS defined concepts, which will be assigned LSIDs, will be those exported from existing taxonomic database providers such as ITIS.

THE TDWG TAXON CONCEPT SCHEMA (TCS)

The need for data exchange standards across the domains of biology, particularly in the context of biodiversity studies, has been identified by the Global Biodiversity Information Facility (GBIF, 2006) and the Scientific Environment for Ecological Knowledge (SEEK, 2006) amongst others. The common approach being taken to provide these standards is the development of XML Schemas that define the data transfer structure as an XML document, including the structure of the metadata associated with the actual data. This approach mirrors that already taken to provide Data Description, or "Mark-up" Languages such as EML

(EML, 2006), CML (CML, 2006), and GML (GML, 2006). The necessary information exchange standards for taxonomic information might include those for taxon concepts, Specimen Records, Collection Details, Publications, Observation Data, Geographical Location and People (i.e., Authors). Standards and protocols for some of these facets are already available or under development, including DIGIR (DIGIR, 2006) and ABCD (ABCD, 2006) for detailing and exchanging information regarding biological specimens; TaxMLit, allowing the complete mark-up of the content of taxonomic work (Weitzman and Lyal, 2004); and a number of standards for publication information (MODS, 2006; XMLMARC, 2006; XOBIS, 2006).

In order to achieve global data exchange standards it is necessary that the standards process should be open and inclusive, and it is desirable that proposed standards should be consistent, and well documented. TDWG has taken a lead in providing an international forum for the development of standards for biological data exchange. Current standards being developed (as XML schema) include the ABCD Task Group On Access to Biological Data (providing standards for transfer and discovery of biological collection data sets); the SDD Task Group on Structure of Descriptive Data (developing a standard for storing and transferring detailed, character-based, descriptions of specimens or taxa) and the Taxonomic Names Task Group on Taxonomic Concept Standards (developing a standard for storing and transferring information about taxon concepts and names, the work we present here). Because of the overlap between these three proposed schemas (for example in their use of taxonomic names and concepts and their referral to specimens and collections) it is proposed to modularize their implementation to allow reuse of each other's data structures. Furthermore, because each type of document will need to provide similar metadata elements describing the data transferred in a document (e.g., the source, ownership, version) it is proposed that documents conforming to each of these three schemata are wrapped in a common format descriptor document.

The TCS schema and underlying model aims to be inclusive of all other models of taxonomy, and allow data from any data source to be accurately represented. A strength of the TCS schema is that it supports many recent innovative models and implementations of taxonomic information as well as dealing with legacy data. Several of these models have been developed specifically to allow the representation of multiple, alternative taxonomic views (HICLAS [Zhong et al., 1996, 1999]; PROMETHEUS [Pullan et al., 2000]; BERLIN/IOPI [Berendsohn, 1995, 1997; Berendsohn et al., 2003]; TAXONOMER [Pyle, 2004]; NOMENCURATOR [Ytow et al., 2001]; uBIO [UBIO, 2006], rather than the standardized single view represented by many global taxonomic checklists (e.g., ITIS, Species2000 [SPECIES 2000, 2006]). Although the requirements for simple data discovery and exchange between database providers has favoured the development and implementation of simple generic data query and retrieval protocols, which use simple models for the underlying data structure (e.g., the successful DIGIR [DIGIR, 2006] protocol with the underlying DARWIN CORE [Darwin Core, 2006] data representation), such flat, unstructured representations of taxonomic information may not be adequate for representing semantically complex information. Various service providers, such as uBio and Species2000 are providing rich mechanisms for resolving names across distributed taxonomic databases. However, resolution services based on taxon concepts as represented by the TCS should provide more meaningful comparison of taxonomic identifiers.

The TCS schema was derived by composing an abstract model of taxonomic concepts as discussed above, which seeks to account for all the facets that different data providers and users might wish to include in their definition of a taxon concept. This was facilitated by detailed consultation with representatives of several taxonomic databases and researchers with an active interest in modelling and implementing taxonomic information systems. The abstract model has been represented as an XML schema that defines the structure of XML documents for the exchange of information about taxonomic concepts. This exchange schema aims to capture data as understood by the data owners without distortion, and facilitate the query of different data resources according to the common schema model. The full schema and documentation can be found at ⟨http://tdwg.napier.ac.uk/TCS_1.01/v101.xsd⟩. An overview detailing some of the elements of the transfer schema is shown in Figure 3.

Each document would carry *MetaData* recording source and creation details of the *DataSet,* together with the details of the taxonomic concept information represented. To allow cross-referencing within the document *Specimens*, *Publications*, *TaxonNames*, and *TaxonConcepts* are given local identifiers (ids), which can

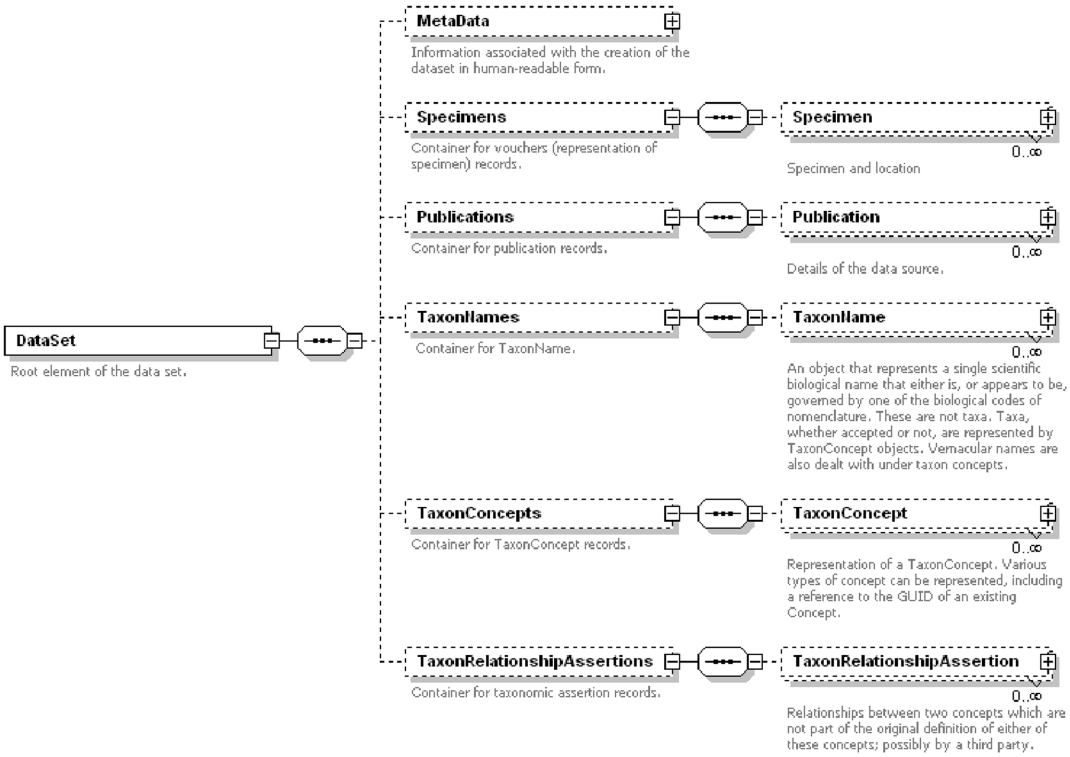

FIG. 3. Overview of the proposed TDWG TCS XML schema.

be substituted with global IDs (GUIDs) if these are available. As well as recording *TaxonRelationships* with other *TaxonConcepts* as part of their definition, the transfer document may also be used to detail third party *RelationshipAssertions* between existing *TaxonConcepts*.

Any combination of the optional component elements shown in Figure 4 would be used to detail *TaxonConcept* definitions according to the data model of the data provider, but typically at least *Name* and *AccordingTo* would be required ("Nomenclatural Concepts" may only provide *Name*). For these two components the detail recorded in different data sources will vary, so a simple string representation will always be provided, whether or not detailed decomposition is possible. The *TaxonRelationship* element allows the *TaxonConcept* to be defined in relation to existing *TaxonConcepts*. This can include hierarchical relationships to parent or child taxa in the same classification, or synonymy and set based relationships with *TaxonConcepts* defined in alternative classifications, based on the extent to which two concepts are congruent or overlap. *SpecimenCircumscriptions* list the specimen details that the *TaxonConcept* is *CircumscribedBy*. *CharacterCircumscriptions* refers to the descriptions as defined in SDD.

CONCLUSION

The computerized systems and databases used by biologists and the bioinformatics community are largely blind to the problems inherent in the (ambiguous) identification of organisms by scientific name alone. As we have discussed, accurate integration of biological data sets is problematic due to many reasons, including the lack of standards for capturing and exchanging taxonomic concepts, and the lack of a global system for taxonomic concept resolution with GUIDs, which can be used to refer to and aid matching concepts for data annotation and integration. Solutions to these problems require ensuring that references to biological taxa in data sets cite the scientific name in the context of a particular classification, in other

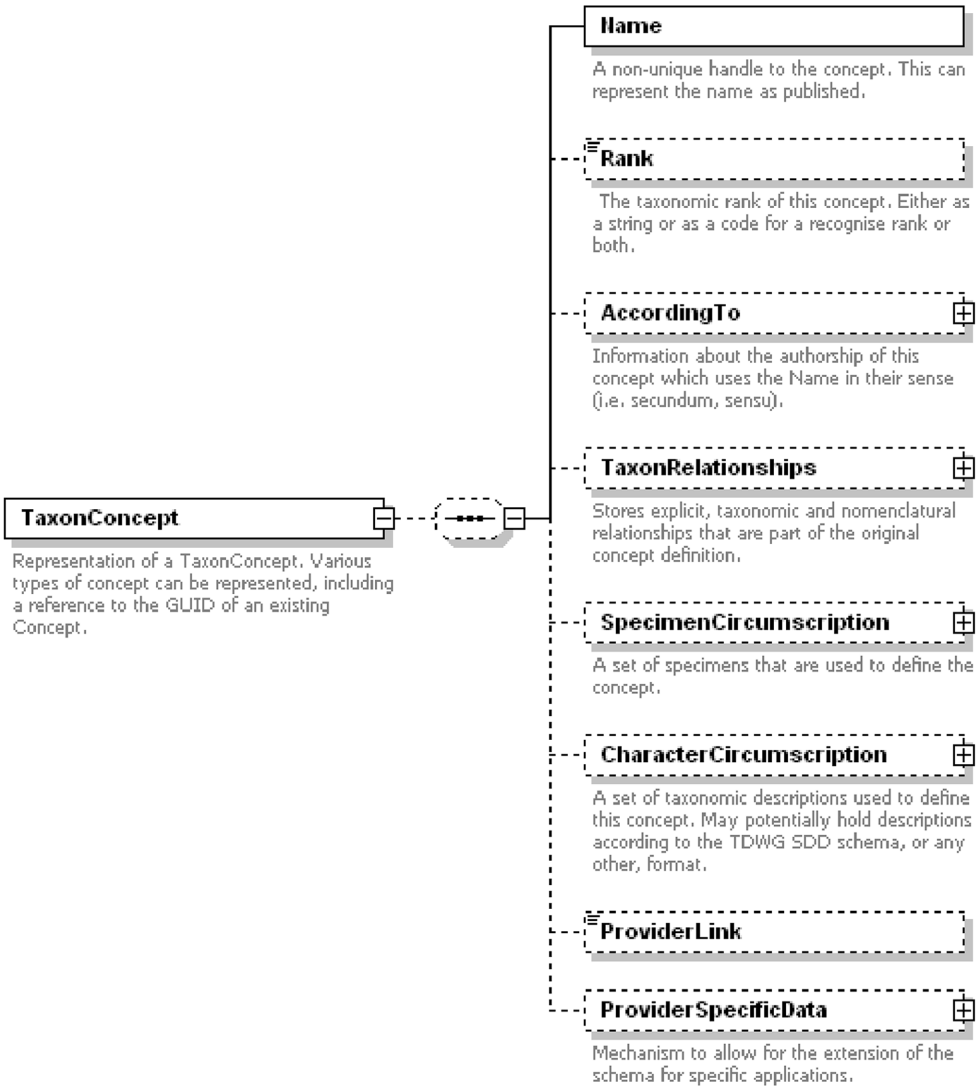

FIG. 4. The TaxonConcept component of the TDWG TCS XML schema.

words, Taxon Concepts, as used in Koperski et al. (2000). Where it is not possible to ascribe defined concepts to datasets (such as with legacy data) poorly defined nominal concepts can be used (i.e., concepts with a name but no definition), thus making explicit the deficient quality of the taxon identification. The TCS has been used to map data from a variety of sources and is currently being used as the basis for a taxonomic name/concept resolution service in the SEEK project, a concept browser in GBIF and as the basis for exchanging names by the name data providers such as IPNI (IPNI, 2006) and Index Fungorum (Index Fungorum, 2006) and its incorporation into EML is proceeding.

The Taxonomic Databases Working Group, supported by the Gordon and Betty Moore Foundation, is currently reviewing their standards processes and developing an architecture for integrating their standards. As a result of this, other implementations of TCS are being evaluated instead of XML. In particular, TDWG is evaluating the feasibility of using LSIDs as globally unique identifiers for taxonomic information and the use of RDF to expose the associated metadata. Therefore, we will be investigating the use of such technologies to represent the TCS model over the coming months.

ACKNOWLEDGMENTS

We are most grateful for detailed and helpful discussions on aspects of individual taxonomic models from Walter Berendsohn (Berlin Model), Donald Hobern (GBIF), Sally Hinchcliffe (IPNI), Paula Huddleston (ITIS), Paul Kirk (Index Fungorum), James Ytow and Dave Roberts (Nomencurator), Frank Bisby, Andrew Jones, and Richard White (Species2000), Richard Pyle (Taxonomer), Bob Peet (Vegbank), James Beach, Aimee Stewart, Rob Gales, Dave Vieglas, Matt Jones, SEEK, and other colleagues within TDWG, including Gregor Hagerdorn (SDD), Chuck Miller (Missouri Botanical Gardens), Jerry Cooper (Landcare Research), and Stan Blum (Darwin Core/DIGIR2). We also thank Dawn Field and Tatiana Tatusov for organizing the "eGenomics: Cataloguing Our Complete Genome Collection" workshop, at which much of this material was presented. This work was carried out under the auspices of TDWG and jointly funded by GBIF and SEEK, supported by the U.S. National Science Foundation.

REFERENCES

ABCD. (2006). Biological collections data sub-group, access to biological collection data. Available at: ⟨www.bgbm.org/TDWG/CODATA.⟩ Accessed 14th March, 2006.

BERENDSOHN, W.G. (1995). The concept of "potential taxa" in databases. Taxon **22,** 207–212.

BERENDSOHN, W.G. (1997). A taxonomic information model for botanical databases: the IOPI model. Taxon **46,** 283–309.

BERENDSOHN, W.G., DÖRING, M., GEOFFROY, M., et al. (2003). *MoReTax: Handling Factual Information Linked to Taxonomic Concepts in Biology* (Bundesamt für Naturshutz, Bonn).

CML. (2006). Chemical markup language. Available at: ⟨http://wwmm.ch.cam.ac.uk/wikis/wwmm/index. php/Main_Page⟩. Accessed 14th March, 2006.

DARWIN CORE. (2006). Darwin core. Available at: ⟨http://darwincore.calacademy.org⟩. Accessed 14th March, 2006.

DIGIR. (2006). Distributed generic information retrieval. Available at: ⟨http://digir.net⟩. Accessed 11th January, 2006.

EML. (2006). Ecological metadata language. Available at: ⟨http://knb.ecoinformatics.org/software/eml⟩. Accessed 14th March, 2006.

GBIF. (2006). Global Biodiversity Information Facility. Available at: ⟨www.gbif.org⟩. Accessed 14th March, 2006.

GENBANK. (2006). GenBank. ⟨www.ncbi.nlm.nih.gov/Genbank/index.html⟩. Accessed 14th March, 2006.

GML. (2006). Geography markup language. Available at: ⟨http://opengis.net/gml⟩. Accessed 14th March, 2006.

GREUTER, W., McNEILL, J., BARRIE, F.R., et al. (2000). *International Code of Botanical Nomenclature. Adopted by the 16th International Botanical Congress St. Louis, Missouri, 1999* (Koeltz Scientific Books, Königstein).

ICSP. (1990). *International Committee on Systematics of Prokaryotes International Code of Nomenclature of Bacteria* (American Society for Microbiology Press, Washington, DC).

ICTV. (2000). *International Code of Virus Classification and Nomenclature*. Available at: ⟨www.ncbi.nlm.nih.gov/ICTVdb⟩. Accessed 14th March, 2006.

INDEX FUNGORUM. (2006). *Index Fungorum*. Available at: ⟨www.indexfungorum.org⟩. Accessed 14th March, 2006.

IPNI. (2006). *International Plant Names Index*. Availabl at: ⟨www.ipni.org⟩. Accessed 14th March, 2006.

ITIS. (2006). *International Taxonomic Information System*. Available at: ⟨www.itis.usda.gov/⟩. Accessed 14th March, 2006.

KOPERSKI, M., SAUER, M., BRAUN, W., et al. (2000). *Referenzliste der Moose Deutschlands* (LV Druck im Landwirtschaftsverlag GmbH, Münster-Hiltrup).

MODS. (2006). Metadata object description schema. Available at: ⟨www.loc.gov/standards/mods⟩. Accessed 14th March, 2006.

NCBI. (2006). NCBI taxonomy. Available at: ⟨www.ncbi.nlm.nih.gov/Taxonomy⟩. Accessed 14th March, 2006.

PULLAN, M.R., WATSON, M.F., KENNEDY, J.B., et al. (2000). The Prometheus taxonomic model. Taxon **49,** 55–75.

PYLE, R.L. (2004). Taxonomer: a relational data model for managing information relevant to taxonomic research. Phyloinformatics **1.** Available at: ⟨www.phyloinformatics.org/pdf/1.pdf⟩. Accessed 14th March, 2006.

RIDE, W.D.L., COGGER, H.G., DUPUIS, C., et al. (1999). *International Code of Zoological Nomenclature*, 4th edn. (International Trust for Zoological Nomenclature, London).

SDD. (2006). TDWG structure for descriptive data sub-group. Available at: ⟨http://160.45.63.11/Projects/TDWG-SDD/⟩. Accessed 14th March, 2006.

SEEK. (2006). Science environment for ecological knowledge. Available at: ⟨http://seek.ecoinformatics.org⟩. Accessed 15th January, 2006.

SPECIES 2000. (2006). Species2000. ⟨www.sp2000.org⟩. Accessed 14th March, 2006.

STACE, C.A. (1999). *Field Flora of the British Isles* (Cambridge University Press, Cambridge).

TCS. (2006). Taxonomic names/concepts sub-group, taxonomic concept schema. Available at: ⟨http://tdwg.napier.ac.uk⟩. Accessed 14th March, 2006.

TDWG. (2006). International Working Group on Taxonomic Databases. Availalbe at: ⟨www.tdwg.org⟩. Accessed 14th March, 2006.

UBIO. (2006). Universal Biological Indexer and Organizer. Available at: ⟨www.ubio.org⟩. Accessed 14th March, 2006.

WEITZMAN, A.L., and LYAL, C.H.C. (2004). An XML schema for taxonomic literature—taXMLit. Available at: ⟨www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf⟩. Accessed 14th March, 2006.

XMLMARC. (2006). XML Machine Readable Cataloging. Available at: ⟨http://laneweb.stanford.edu:2380/wiki/medlane/xmlmarc⟩. Accessed 14th March, 2006.

XOBIS. (2006). XML Organic Bibliographic Information Schema. Available at: ⟨http://xobis.stanford.edu/⟩. Accessed 14th March, 2006.

YTOW, N., MORSE, D.R., and ROBERTS, D.M. (2001). Nomencurator: a nomenclatural history model to handle multiple taxonomic views. Biol J Linnean Soc **73,** 81–98.

ZHONG, Y., JUNG, S., PRAMANIK, S., et al. (1996). Data model and comparison and query methods for interacting classifications in a taxonomic database. Taxon **45,** 223–241.

ZHONG, Y., LUO, Y., PRAMANIK, S., et al. (1999). HICLAS: a taxonomic database system for displaying and comparing biological classification and phylogenetic trees. Bioinformatics **15,** 149–156.

Address reprint requests to:
Dr. J. Kennedy
School of Computing
Napier University
Merchiston Campus
Edinburgh, Scotland, EH10 5DT, United Kingdom

E-mail: J.Kennedy@napier.ac.uk