Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration

Jessie B. Kennedy, Robert Kukla and Trevor Paterson.
School of Computing, Napier University, Edinburgh, EH10 5DT, U.K.
{j.kennedy, r.kukla, t.paterson}@napier.ac.uk

Abstract. Biologists use scientific names to label the organisms described in their data; however, these names are not unique identifiers for taxonomic entities. Alternative taxonomic classifications may apply the same name, associated with alternative definition or circumscription. Consequently, labelling data with scientific names alone does not unambiguously distinguish between taxon concepts. Accurate integration and comparison of biological data is required on taxon concepts, as defined in alternative taxonomic classifications. We have derived an abstract, inclusive model for the diverse representations of taxonomic concepts used by taxonomists and in taxonomic databases. This model has been implemented as a proposed standard XML schema for the exchange and comparison of taxonomic concepts between data providers and users. The representation and exchange of taxon definitions conformant with this schema will facilitate the development of taxonomic name/concept resolution services, allowing the meaningful integration and comparison of biological datasets, with greater accuracy than on the basis of name alone.

1 Introduction

Scientific names are inherently poor identifiers for organisms, because although names are formalized and validated according to strict codes of nomenclature, the same name can be applied by taxonomists to alternative taxonomic views of the extent or definition of a taxon (e.g. a species, genus *etc*.). Biologists (i.e. the 'users' of taxonomic classifications) identify and label their data with scientific names, by identifying their organisms according to a particular taxonomic classification, as found for example in field guides, but without recognizing and recording that taxonomic context. As a consequence datasets cannot be reliably integrated on the basis of the scientific names because the context or meaning of the name is not captured.

Taxonomic identification is emerging as a significant problem for the integration and comparison of diverse datasets across all fields of biology from genomics to ecology. For example, annotations of Genbank DNA sequences typically label the source species according to the NCBI Taxonomy (www.ncbi.nlm.nih.gov/Taxonomy). 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 (for example the

genus *Fugu* has recently been 'renamed' *Takifugu*). Such an approach cannot deal with complex, changing and unrecorded relationships between names as used according to alternative taxonomic views. For example, the alternate classification and reclassification of Orangutans into separate species or subspecies means that sequence data might be labelled according to a variety of alternative classifications. (Currently over 50,000 nucleotide sequences are ascribed to *Pongo pygmaeus*, with fewer than 100 for each 'subspecies' *abelii* and *pygmaeus*). It is not clear how the NCBI Taxonomy might handle the alternative reclassification of these sub-species as species or whether the 50, 000 *P. pygmaeus* sequences include data that some taxonomists would ascribe to *abelii* (species or subspecies). These problems impact on other areas of biology and beyond. For example, the increase between 1996 and 2000 in the number of officially endangered primate species is partly attributable to the decision in 2000 to accept the reclassification of some subspecies (including Orangutan) at the species level [1]. Clearly consideration of species names in isolation, without the appropriate classificatory context, makes it difficult to interpret biodiversity data such as the distribution of Orangutans, when collected at different times, and labelled according to different (unrecorded) classification contexts.

1.1 Taxonomy and Nomenclature

Taxonomists classify organisms into hierarchically ranked taxa according to their evolutionary relatedness, based on any of a variety of types of biological evidence (morphology, genetics, palaeontology *etc.*). Alternative classifications (taxonomic revisions) arise over time reflecting new or alternative taxonomic opinion following more detailed study, the discovery of new taxonomic information such as evidence about relationships between taxa, description of new species, and increasingly molecular phylogenies based on DNA sequence comparison. Therefore taxonomy is itself an investigative science, and taxonomic classifications represent partial and evolving hypotheses rather than static identifications of absolute taxa. Any recorded taxonomic classification represents an opinion, according to one authority, at a given time. Relationships may be expressed or inferred between successive or alternative taxonomies, relating the concepts (taxa) in one classification to concepts in another, but without knowing the total genetic history of all life on earth it is not possible to derive a final, 'true' classification of existing (and extinct) organisms.

Taxonomists use scientific names in order to label and communicate about the taxonomic concepts that they create. Names are applied to the taxa in a given classification according to the codified rules of nomenclature, based on 'typification' (i.e. by reference to archived 'type' specimens) and following the principle of 'priority' where names are dependent on the oldest type specimen included in the circumscription of a taxon. This system provides stability to scientific names over time, as they are preserved in relation to their original use and type specimen. However, as a direct consequence of the application of these rules the same valid scientific name will apply to different views of a taxon according to different postulated taxonomic classifications. Indeed it is also true that very similar taxonomic concepts may have different names according to different classifications.

Names therefore are a *part* of a 'taxon concept', and cannot be used to unambiguously identify a concept. The identifiers used by experimental biologists to label organisms as a member or instance of a particular taxonomic concept should unambiguously refer to the taxon concept itself: true integration therefore requires unique identifiers for taxon concepts. We propose these concept identifiers should minimally include the *scientific name applied* and the *classification context*. This context is represented by the authorship of the concept, i.e. an 'According To' or *secundum* (*sec.*) citation. Assigning identifiers for concepts allows simple resolution of taxon concepts based on identity, particularly if GUIDs were to be adopted for concepts.

Taxonomic concepts are created and defined (or revised) in taxonomic publications. These publications may include various levels of detail defining each taxon, which might include: character descriptions (i.e. a list of structure, attribute, value triples), lists of archived specimens which are included in the taxon (specimen circumscription), relationships to other concepts in the same classification (including parent-child relationships between a taxon and its subordinate taxa), relationships with concepts in earlier alternative classifications, assignment of rank (family, genus, species *etc.*) and application of a scientific name for this taxon. Individual taxonomists have different perceptions or models for what constitutes and defines a taxon. This makes comparison of alternative taxon concepts problematic, even if the full rationale for the classification is available. However, comparing components of concept definitions might allow experienced Taxonomists to establish and record relationships between concepts with different GUIDs (e.g. two concepts can be considered equivalent for some particular purpose).

1.2 The users of taxonomic classifications

The complex issues of ambiguity surrounding taxonomic classification and naming are well understood by expert taxonomists, but their importance and consequences are probably not considered relevant by experimental biologists who wish to use the names as static identifiers for the organisms described in their data. The explosion in biological data makes the accurate identification of source organisms critical. For example a researcher will frequently wish to identify which available datasets contain information on a particular organism of interest. Typically datasets are annotated by scientific name. However, correct identification of these datasets requires matching the taxonomic concepts as used in the source datasets, with the taxonomic concept of interest to the researcher (as defined by their reference classification). This requires either the use of identifiers for concepts, or comparison of the actual definitions of the concepts of interest with the definitions used by the authors of each dataset. A corollary of this is that datasets should be marked up with unambiguous taxonomic concept identifiers, for example they should reference the identification guide or classification system used by the researcher: identification by scientific name alone is insufficient.

By way of example a researcher wishing to access data on a fictitious species *Aus bus* from globally distributed databases might minimally want to recover data about any species that had ever been known as species *Aus bus*, or they might want to extend this query to recover information about all named species asserted to be synony-

mous with *Aus bus* at some level. Alternatively, if they have precise knowledge of the underlying concept described as species *Aus bus* they may only want to retrieve information about concepts closely related to their own concept of *Aus bus*, regardless of their identifying names. Such detailed exploration of all species that overlap or are equivalent with *Aus bus* is only possible if 'names' are resolved according to the concepts to which they have been attached, so that data is retrieved on the basis of concept comparison, regardless of nomenclatural issues. Firstly however we require a common exchange schema to facilitate the representation, exchange and query of concepts.

In the following section we describe the current use of biological nomenclature and present an example to illustrate the problems associated with relying upon scientific names as identifiers for organisms. In section 3 we discuss the variety of approaches taken by biologists when describing taxonomic concepts and in section 4 argue the case for a standard schema to allow the exchange of this data to permit potential comparison and resolution of taxonomic concepts. In section 5 we present our work in defining the Taxonomic Concept Schema, an XML exchange standard for taxonomic concepts and names and compare this to other models in section 6. Finally some conclusions are drawn in section 7.

2. Using Names As Identifiers Of Concepts

The formulation and application of valid scientific names for taxonomic groups is governed by separate codes of nomenclature for botany, zoology, bacteria and viruses (ICBN [2]; ICZN [3]; ICSP [4], ICTV [5]). According to these rules the name of a taxon is usually determined by the oldest type specimen included in its circumscription. The history of the fictitious genus *Aus* detailed in Figure 1 (and described more fully online [6]) illustrates how the rules of nomenclature provide stability for names throughout the history of taxonomic revisions, but automatically mean that names cannot be used as unique, non-ambiguous identifiers of taxon concepts. In fact the use of species names can never be truly separated from a taxonomic classification because the rules of binomial nomenclature obscure the boundary between classification and nomenclature for taxon names below the level of 'genus' (see for example [7]).

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 outwith 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. Published and electronically deposited data might therefore be labelled with a variety of names, of varying precision and specificity. For example data about a particular species of 'daisy' can be found labelled as: lawn daisy, English lawn daisy, european lawn daisy, USDA code BEPE2, APNI code 163507-3, ITIS TSN 36862, *Bellis perennis*, *Bellis perennis* L., *Bellis perennis* L. Sp.Pl. 886, *Bellis perennis* L. Species Plantarium 2 1753, *Bellis perennnis* L. Species Plantarium (1753): 886, *Erigeron perennis* (L.) Sessé & Moc., *Conyzopsis bellis* EHL Krause. Integration and

resolution between such diverse and semantically distinct names is clearly non-trivial, where even a 'single' name might be recorded with minor variations due to errors and corrections in spelling, or there may be variation in the abbreviations used.

A growing number of taxonomic resources and databases are available online, which seek to provide an integrated record of the names and taxonomic relationships for a particular narrow or wide taxonomic range (e.g. FishBase, www.fishbase.org; ITIS, www.itis.usda.gov). These taxonomic databases require quite complex models of taxonomic names in order to represent their data and to account for the needs of their users. Historically such databases only represented single, aggregated views of taxonomy, but it is now recognized that the issue of multiple classifications should be addressed. This requires consideration both of the synonymies between names as used in alternative classifications, and the application of the same name to different concepts in alternative classifications. Current representations of synonymy between names fail to capture the full complexity of these relationships which imply differences between concept definitions not simply between names.

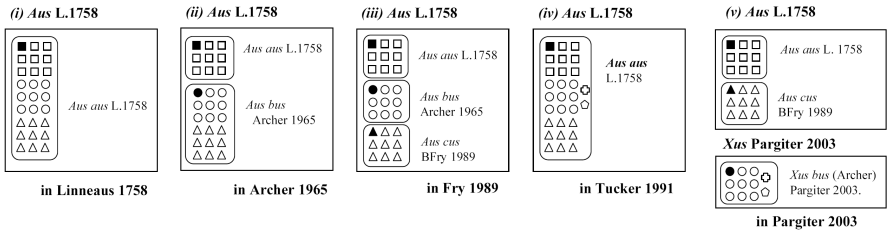

Figure 1. Taxonomic history of the imaginary genus *Aus* L. 1758 (i) through four subsequent revisions (ii – v). Individual specimen organisms are represented by the symbols ○, □, △ *etc.*, with nomenclatural type specimens infilled: ▲, ■, ●. In 1965 Archer split *Aus bus* Archer 1965 from *Aus aus* L.1758 (ii), which was in turn 'split' creating *Aus cus* Fry 1989 (iii). Discovery of new specimens in 1991 caused Tucker to re-'lump' taxa in a single species *Aus aus* L.1758 (iv), but according to Pargiter these new specimens indicated that *bus* (Archer) in fact belonged in a separate new genus as *Xus bus* (Archer) Pargiter 2003 (v). Comparing the specimen circumscription of the various views on the taxa it is clear that the underlying concepts referred to by the various names change over time. For example compare *Aus aus* L.1758 in (i) versus (ii); or *Aus bus* Archer 1965 in (ii) and (iii); or the relationship of *Aus bus* with *Xus bus*.

3. Defining Taxonomic Concepts

A taxonomic concept is one view of what constitutes a taxonomic entity, be it a species, genus or taxon of higher rank. Typically this would be represented as a published opinion or hypothesis according to a given author team, and include a valid scientific name as controlled by the rules of nomenclature. Care should be taken to distinguish between published taxonomic concepts, representing taxonomists' classifi-

cation hypotheses, and the publication of data by biologists who are only identifying organisms according to some preexisting taxonomic concept, i.e. name usage [8].

A minimal representation of a taxon concept is therefore a scientific name plus citation of definition (i.e. an attribution). In this respect any first usage of a scientific name represents an original taxon concept, as published by the author of the name. As the rules of nomenclature require the original author to be included as part of the name, e.g. *Aus aus* L. 1758, this combination does not uniquely distinguish the original concept in a taxonomic database, as the same name might be valid for subsequent revision concepts, which should be distinguished by recording the originator of the concept, in addition to the author of the name (as part of the full scientific name), e.g. *Aus aus* L. 1758 *sec*. Fry 1989. Recording the originating (*sec*.) authorship for a concept therefore distinguishes between concepts, but does not provide any information with which to *compare* different concepts. The meaningful comparison of defined concepts would require the user to consult and interpret the original citations, where available. Any computer-assisted automatic comparison and resolution of concepts will require that the elements of the concept definition are stored as part of the electronic representation of the concept in the taxonomic database sources.

We have modelled how taxon concepts can be represented with varying complexity by a range of creators and users of concepts (including taxonomists, database providers and experimental biologists). Detailed analysis of the components that are used by taxonomic databases or found in taxonomic publications to define their taxon concepts includes (i) specimen and taxon circumscriptions, (ii) character descriptions or circumscriptions and (iii) relationships with other taxon concepts.

There are a wide variety of relationships that might be expressed between taxon concepts, which have been considered in detail by others (e.g. [9]; see online documentation, section 2.3 [10]). These relationships may implicitly or explicitly represent set-based relationships defining the extent of overlap with or inclusion of other concepts, or they may capture 'nomenclatural' relationships. However, the description of types of relationships is complicated by the interdependence of nomenclature and classification. A strict interpretation of terms such as synonymy, homonymy *etc*. implies relationships between the definitions of names, and it is questionable whether a relationship between names can be asserted in the absence of the context or usage of those names. Any relationship between taxon names at least minimally considers relationships between the type specimens determining the names. In the Taxonomic Concept Schema (TCS) model presented in this paper a 'nomenclatural' relationship is expressed as a relationship between two concepts, implying between the names of the concepts.

4. The Requirement for Data Exchange Standards

Given that there are an increasing number of important database providers of taxonomic information, and a large potential user base amongst biologists and non-scientists, it is necessary to facilitate data exchange between the providers and the users, so that data can be integrated from multiple sources, without losing or misrepresenting the semantics of the data according to the providers' information models.

This is necessary both from the perspective of database providers who wish to aggregate information from multiple data sources into a single representation of taxonomy without duplication of concepts, as well as for taxonomically naive users who wish to integrate data from multiple database providers. If no exchange standard is globally adopted, it will be necessary for any application or service that seeks to query multiple taxonomic databases to implement bespoke query and exchange protocols for each provider. It would then be impossible to develop standard mechanisms to match or resolve concepts between different sources, and no guarantee of any protocol's stability or longevity.

The need for data exchange standards across the domains of biology, particularly in the context of biodiversity studies, has been identified by GBIF [11] and SEEK [12] 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 (EcologicalML [13]), CML (ChemicalML [14]) and GML (GeographyML [15]). The necessary information exchange standards for taxonomy might include those for taxon concepts, Specimen Records, Collection Details, Publications, Observation Data, Geographical Location and People (i.e. Authors *etc.*). Standards and protocols for some of these facets are already available or under development, including: DIGIR [16] and ABCD [17] for detailing and exchanging information regarding biological specimens; TaxMLit allowing the complete mark-up of the content of taxonomic work [18], and a number of standards for publication information (MODS [19]; XOBIS [20]; XMLMARC [21]; *etc.*).

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 (International Taxonomic Databases Working Group, www.tdwg.org) 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 in section 5). 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 (for example the source, ownership, version *etc.*) it is proposed that documents conforming to each of these three schemata are wrapped in a common format descriptor document.

5. The TDWG Taxon Concept Schema (TCS)

Considered in abstraction, models for both a taxon name and a taxon concept consist of a *label* plus *definition* plus *author*. Therefore, as demonstrated by Pyle [22], a taxon concept can be represented as a taxon name (protonym) plus definition plus author. Taxonomic definitions of names include the type specimen for that name and application of the rules of nomenclature, whereas the taxonomic definition of a concept might take several explicit (or implicit) forms. A model for names that includes relationships between names might be considered as incorporating elements of a concept model as the relationships between names actually refers to both the usage context and typification of that name.

Because of the structural similarity between elements of names and concepts, and to encourage a more rigorous representation of taxonomic identifiers (as defined concepts rather than somewhat ambiguous names), an XML schema is proposed for the representation and exchange of information regarding taxon concepts. Because the schema includes a representation of names it will be possible to use this schema to represent names as being concepts that lack definitions (i.e. as nominal concepts).

By making explicit the differences in composition between various types of taxon concept definition, the schema will allow users to be aware of the variable accuracy or quality of resolution, whether based solely upon names or upon more richly defined taxon concepts. Various service providers, such as uBio (www.ubio.org) and Species2000 (www.sp2000.org), 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 (see acknowledgments). 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 tdwg.napier.ac.uk. The TDWG review process is open and inclusive, giving the opportunity to any interested party to comment and suggest amendments to the proposal.

An overview detailing some of the elements of the transfer schema is shown in Figure 2. Each Dataset will carry MetaData detailing the source of the transferred document. To allow cross-referencing within the document, Vouchers (Specimen records), Publications and TaxonConcepts are given local identifiers (IDs) that could be substituted with global IDs (GUIDs) if these are available (see below). As well as recording the details of TaxonConcepts (which can include Relationships with other TaxonConcepts, see Figure 3), the transfer document may also be used to detail third party RelationshipAssertions between existing TaxonConcepts.

Because the model represented by the schema aims to be inclusive no 'components' of a taxon concept definition are required by the schema, but are optional constituents of a concept as represented by a given provider. However, in order to be useful, a minimal representation would generally include both a Name and details of the concept authorship (i.e. AccordingTo, or *sec.*). The representation of a full scientific name (NameDetailed) that conforms to the requirements of all existing codes of Nomenclature has been developed outside the project (by the Linnean Core interest group [23]) and integrated into the schema.

The various elements of the schema materialize information defining the concept according to the original authors of the concept. This might include details of the concept's Relationships to other pre-existing concepts, including its circumscription by (inclusion of) other (lower rank) taxon concepts, or its membership of higher rank concepts. Further Relationships may detail similarity or overlap with concepts created by other authors. These latter relationships can be considered 'horizontal' in the sense that they can relate concepts defined according to different taxonomic classifications, whilst the hierarchical relationships between concepts within a classification are 'vertical'. A full list of the types of relationships that may be expressed between two concepts is provided online [10].

The manner in which a concept may be circumscribed by 'Character' data is as yet undefined in the schema, and would require a formal model for representing character descriptions. Various structured models for character data have been proposed (see for example [24]), and the SDD working group of TDWG is developing a schema for specimen or taxon descriptions that could be included or referenced within a TCS CharacterCircumscription. The TCS schema does however provide the mechanism for circumscribing concepts by reference to identifiers of specimen records (Vouchers in the schema). Individual specimens that circumscribe a taxon can be labelled according to whether they are accepted holotypes, isotypes, neotypes *etc.* for that taxon, according to the codes of nomenclature.

The structure of the TCS schema allows internal reference and reuse of 'top-level' elements (i.e. TaxonConcepts, RelationshipAssertions, Voucher and Publication records). Indeed it is hoped to standardize the representation of Publications and Vouchers (including Specimens) across the TDWG schemas (see above). Where any of these reusable elements are globally defined and resolvable via Globally Unique Identifiers (GUIDs) it will be possible to represent them in transfer documents simply by reference to this GUID (see below).

Some taxonomic work is concerned with re-using existing taxonomic concepts. For example a taxonomist creating a revision of a large taxon may accept various included taxa according to the work of various other published taxonomists, but wish to record opinions about the relationships between these concepts. Where these relationships are not created as part of a new concept definition they are treated as 'third party' in the schema, and stored as RelationshipAssertions with an AccordingTo authority.

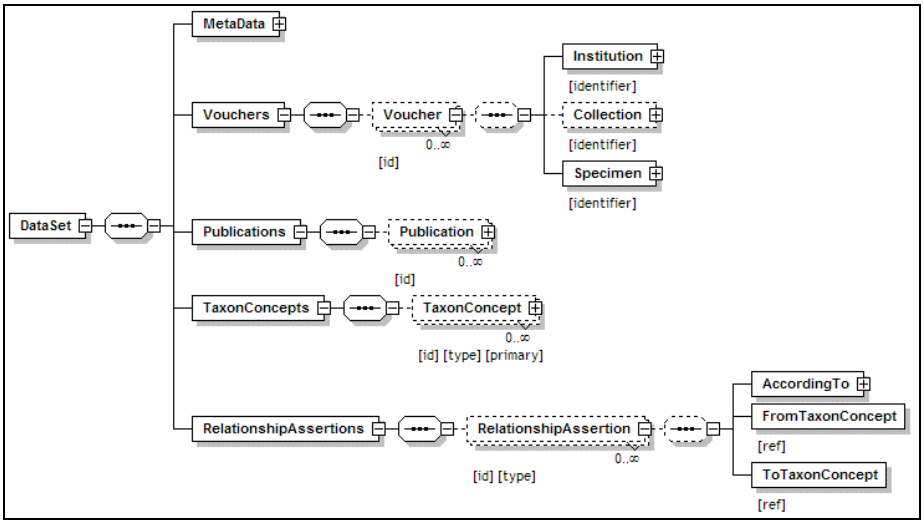

Figure 2. (legend overleaf).

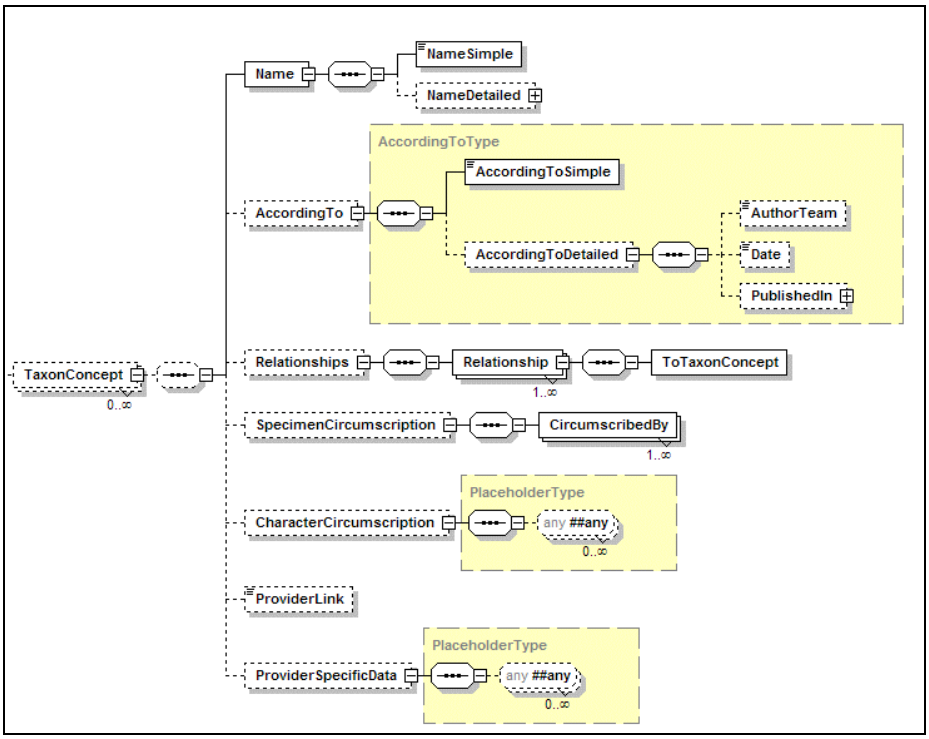

Figure 3. (legend overleaf).

Figure 2. Overview of the Proposed TDWG TCS XML Schema. The major components of the schema for transferring taxonomic concepts are shown diagrammatically (XML Elements are shown in boxes, with XML attributes listed [below]; generated with XMLSpy.com software). 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 *Vouchers* (Specimen records), *Publications* and *TaxonConcepts* are given local identifiers (ids), which could be substituted with global IDs (GUIDs) if these are available. As well as recording the details of *TaxonConcepts* (which can include *Relationships* with other *TaxonConcepts*, see Figure 3), the transfer document may also be used to detail third party *RelationshipAssertions* between existing *TaxonConcepts*

Figure 3. XML Schema Diagram for a Taxon Concept. A portion of the proposed TDWG TCS schema for transferring Taxonomic Concepts is shown diagrammatically (generated with XMLSpy.com software). Any combination of the optional component elements 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 *Relationship* 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 (*Vouchers* in Figure 2) that the *TaxonConcept* is *CircumscribedBy*, but the nature of *CharacterCircumscriptions* is as yet undefined. The *PlaceholderType* allows standards developed as other schemas to be incorporated; provision of the *ProviderSpecificData* element allows application specific extensions to the representation of a Taxon Concept.

5.1 Globally Identified Taxonomic Concepts

At present each taxonomic database has its own internal (and sometimes external) identifiers for taxon names or concepts (e.g. TSN numbers used by ITIS *etc.*). These are not represented in the core 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 (i.e. minimally Name plus AccordingTo). Once implemented concept GUIDs would simplify the mark-up of any biological data, according to available defined concepts, and could assist data retrieval based on concept identity. Provision of GUIDs would also help reduce the redundancy and overlap between different data providers who currently reproduce alternative representations of the 'same' concept. Discussion within TDWG, SEEK, GBIF and the wider biological

community is investigating the feasibility of providing GUIDs not only for taxon concepts, but also for other stable concepts such as Publications and Specimens.

The availability of stable GUIDs with which any biologist can annotate their data to unambiguously record the organisms described in their work will greatly facilitate the interpretation, integration and accurate reuse of data across the whole of biology and beyond. Furthermore, eventually it should be possible for a given researcher to chose to recognize and use concepts as provided and defined by a preferred taxonomic resource (e.g. ITIS) or even to capture uncertainty by using less well-defined concepts, or collections of possible concepts were identifications are uncertain.

5.2 Resolving Taxon Names and Concepts

The proposed schema was initially conceived in the context of SEEK's requirement for a taxonomic concept/name resolution service with which to resolve taxonomic names as recorded in ecological data sets, following the realisation that resolution by name alone is insufficient, and in the absence of identification through GUID referenced taxon concepts [12]. Typical scenarios would involve the matching of names as provided by users querying the system with the names as found in the metadata of global data repositories, by resolution through the defined concepts provided by taxonomic name providers and servers.

By capturing the individual components of concept definitions, according to any data model, the schema will allow matching to be performed on any combination of the individual components. The type and accuracy of the comparison performed may vary according to the requirements of the user, i.e. concept matching should be 'fit for the purpose'. For example, a match on the abbreviated scientific name *Aus bus*, will be of lower quality (or precision) than matches specifically to the full, attributed name *Aus bus* L. 1758 *sec.* Fry 1989. For some experimental purposes the loose match to *Aus bus* will be sufficient, but for others greater precision is necessary. A related notion is that comparison matches may be of higher or lower quality, and a 'reliability' score might be provided for different concepts returned by the resolution service.

Where the concepts are fully defined in terms of the components of the TCS model, matching on the actual definitions might be possible. When possible this will allow very high quality matches, for example, where resolution is on the basis of comparing full specimen circumscriptions. Alternatively, resolution only on the basis of name-bearing type specimens would provide a less precise, lower quality resolution, which might still be 'fit for purpose'. Whilst it might be possible to assign 'quality scores' to different components of the concept definition model, in practice it might be necessary to weight these scores to reflect the particular taxon model favoured by a user, or the purposes for which they wish to represent a taxon concept. This would allow users to differentially value the alternative components of a concept definition, and recognize higher value in matches according to their favoured criteria. Implementation of a name/concept resolution service would therefore need to include its own quality model for matching, but allow users flexibility in weighting the comparison algorithms or interpreting the results.

6. TCS in Comparison to Other Models for Taxonomy

As stressed earlier 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 [25,26]; PROMETHEUS [27]; BERLIN/IOPI [7-9]; TAXONOMER [22]; NOMENCURATOR [28]; uBIO www.ubio.org), rather than the standardized single view represented by many global taxonomic checklists (e.g. ITIS www.itis.usda.org; Species2000 www.sp2000.org).

In the TCS model the taxon concept is the core object, which includes name, attribution and definition elements. Whilst many database models also represent a central notion of a taxon object, typically the name is used as an identifier for this object. The Nomencurator database model [28] tracks nomenclatural history using a dual name and publication based model to represent potential taxa by 'name usage'. 'Annotations' are used to record relationships between these name usages, providing an implicit notion of taxon concepts. As such Nomencurator was designed to reflect the manner in which taxonomists work in recording revisions, tracking the development of taxonomic theories by changes in name usage. However, as there is no representation of a taxon concept it is not possible to use the model to define taxa, nor does it readily provide identifiable and exchangeable concepts that can be shared amongst the various users of taxonomy. It should be possible to map each Nomencurator 'name usage' (i.e. name plus publication) to a unique TCS taxon concept, using Name and AccordingTo elements.

The Potential Taxon notion, i.e. the representation of subjective views of a taxon, forms the basis of the Berlin IOPI model for botanical databases [8,9]. In this rich and complex model botanical information can be linked to potential taxa (i.e. name plus circumscription reference) rather than to name alone. Such information can include nomenclatural and systematic relationships as well as linked specimen determinations and character descriptions. Alternate taxonomic classifications are related to potential taxa rather than names, closely corresponding to the TCS model. As with Nomencurator it is envisaged that it will be desirable to present a 'Preferred View' of taxonomy to users, by filtering according to preferred reference authorities. A number of databases implement the Berlin model, including the MoReTax database [29, 9] which defines fundamental, set-based relationships which can be expressed between potential taxa. These relationships are included in the types of relation representable in the TCS [10].

The Taxonomer database model [22] also represents potential taxa, by the intersection of a Name and a Reference, called an Assertion. Assertions of the first usage of that name are treated as a special case, as the name (or Protonym) provides the label for the taxon concept. Protonyms form the name for later revised opinions on a taxon concept as implicitly or explicitly circumscribed in a subsequent publication, represented in the model by an Assertion. Protonyms therefore provide common handle for both the name and any taxon concepts or Potential Taxa that use this name. TCS represents protonyms as the Name components of Original taxon concepts, and TCS Revision taxon concepts may express various synonymy relationships to the Original Concepts sharing a taxonomic name. As with TCS taxon concepts, Assertions may be

linked by concept relationships (such as those defined by Geoffroy and Berendsohn [31]), and can have attached specimen determinations and character descriptions (as text based 'Excerpts'). In the Taxonomer model, however, common names are represented not as individual concepts (or assertions) but as an attribute of an Assertion (which must be or include a Protonym).

The uBio model of taxonomic information underlying their Taxonomic Name Service (www.ubio.org) seeks to separate 'objective' nomenclatural information into a consensual reference model (NameBank), whilst representing classification information in a separate but linked model of subjective opinions (ClassificationBank). uBio assert that this separation whilst providing a rich representation of taxon concepts through classification relationships will allow nomenclaturists to work with bare names and represent relationships between them, without referring to concepts. The justification being that whereas many aspects of nomenclature are not disputed, taxonomic classifications are inherently unstable, disputed hypotheses. On the other hand the TCS does not represent names independently, and relationships must be expressed through a concept that bears a particular name. This reflects our opinion that it is difficult to find any instances where names are used for identification and communication of taxa without at least an implied notion of the concept to which they apply. Datasets containing only name information, are represented by 'nominal concepts' which capture all concepts that share the same name.

As with the Berlin and uBio models, the Prometheus taxonomic database model, which is based on specimen circumscription, clearly distinguishes nomenclatural from classification information [27] and was built to support the working practices of taxonomists performing botanical revisions. In this model naming is an automatic feature of typification in the specimen circumscription. Alternative classification views, based on specimen circumscription, can readily be compared on the basis of set-based relationships (such as those defined in the MoReTax/Berlin model [9]).

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 (for example, the successful DIGIR [16] protocol with the underlying Darwin Core data representation [30]). Whilst such flat, unstructured representations of taxonomic information are certainly simple, they may not be adequate for representing semantically complex information. Species2000 (www.sp2000.org) has developed a Standard Dataset model for exchanging name-based species information according to a single aggregated view of taxonomy, derived from various database sources. Although there is no explicit statement on what 'defines' a named species concept in this model, each species can be recognized as a 'concept' according to the originating source database, or a recorded taxonomic scrutinizer, and could therefore be represented in TCS as a (not well defined) Taxon Concept. The synonymy relationships captured in Species2000 are purely nomenclatural, as the synonyms do not belong to any alternative conceptual hierarchy. Representing such synonymies in TCS would require that each name be represented by a nominal concept.

Whilst the details captured in each of these theoretical and implementation models of taxonomy vary greatly, they tend to converge on a central representation of a potential taxon or taxon concept. TCS can therefore accommodate the salient features of

these models, as well as representing database models that use a more traditional representation of taxonomic names as the identifiers.

7. 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 errors in documenting taxonomic names; the lack of standards for capturing the definition of taxonomic concepts; the inherent ambiguity the taxon definitions associated with taxonomic names; the lack of understanding of this ambiguity by users of biological names; and finally the lack of a global repository for taxonomic concepts 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, which we have modelled as the defining attributes of a Taxon Concept. Data integration can then be achieved either on Concept identity, or on individual components of a defined concept. 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 schema 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.

Acknowledgments

This work was carried out under the auspices of TDWG and jointly funded by GBIF [12] and SEEK [11], supported by the US National Science Foundation. We are most grateful for detailed and helpful discussions on aspects of individual taxonomic models with representatives of the Berlin Model, GBIF, IPNI, ITIS, Nomencurator, Species2000, Taxonomer, Vegbank, and colleagues within SEEK and TDWG.

References

1. International Union for Conservation of Nature. 2004. IUCN Red List of Endangered Species http://www.iucnredlist.org
2. Greuter W, McNeill J, Barrie FR, Burdet HM, Demoulin V, Filgueiras TS, Nicolson DH, Silva PC, Skog JE, Trehane P, Turland NJ, Hawksworth DL. 2000. (Editors & Compilers): International Code of Botanical Nomenclature. 16th International Botanical Congress St. Louis, Missouri, 1999. (Regnum Vegetabile, 138). Königstein: Koeltz Scientific Books.
3. ICZN 1999. (International Commission on Zoological Nomenclature). International Code of Zoological Nomenclature 4th ed. London : ICZN.

4. ICSP. (International Committee on Systematics of Prokaryotes). International Code of Nomenclature of Bacteria, 1990. Washington: American Society for Microbiology Press.
5. ICTV 2000. (International Committee on Taxonomy of Viruses). International Code of Virus Classification and Nomenclature. http://www.ncbi.nlm.nih.gov/ICTV/rules.html
6. http://www.soc.napier.ac.uk/tdwg/index.php?pagename=TCSAndTheLinneanCore
7. Berendsohn WG. 1995. The concept of "potential taxa" in databases. Taxon 22:207-212.
8. Berendsohn WG. 1997. A taxonomic information model for botanical databases: the IOPI model. Taxon 46:283-309.
9. Berendsohn W, Döring M, Geoffroy M, Glück K, Güntsch A, Hahn A, Jahn R, Kusper W-H, Li J, Röpert D, Specht F. 2003. MoReTax: Handling factual information linked to taxonomic concepts in biology. (Schrift. Veget. 39). Bonn: Bundesamt für Naturshutz.
10. Taxonomic Concept Schema Complementary Documentation for Draft Standard http://tdwg.napier.ac.uk/doc/tdwg_tcs.doc section 2.3
11. GBIF.The Global Biodiversity Information Facility, www.gbif.org
12. SEEK 2004. The Science Environment for Ecological Knowledge. http://seek.ecoinformatics.org
13. EML 2004. Ecological Metadata Language. http://knb.ecoinformatics.org/software/eml
14. CML 2004. Chemical Markup Language. http://wwmm.ch.cam.ac.uk/moin/ChemicalMarkupLanguage
15. GML 2004. Geography Markup Language. http://opengis.net/gml
16. DIGIR 2004. Distributed Generic Information Retrieval. http://digir.net
17. ABCD 2004. Access to Biological Collection Data http://www.bgbm.org/TDWG/CODATA
18. Weitzman AL, Lyal CHC (2004) An XML schema for taxonomic literature – taXMLit. available at http://web4.si.edu/sil/bca/status.cfm
19. MODS 2004. Metadata Object Description Schema. http://www.loc.gov/standards/mods
20. XOBIS 2004. XML Organic Bibliographic Information Schema. http://laneweb.stanford.edu:2380/wiki/medlane/schema
21. XMLMARC 2004. XML Machine Readable Cataloging. http://laneweb.stanford.edu:2380/wiki/medlane/xmlmarc
22. Pyle RL (2004) Taxonomer: a relational data model for managing information relevant to taxonomic research. Phyloinformatics 1: 1
23. TDWG Linnean Core Group: http://wiki.cs.umb.edu/twiki/bin/view/UBIF/LinneanCore
24. Paterson T, Kennedy JB, Pullan MR, Cannon A, Armstrong K, Watson MF, Raguenaud C, McDonald SM, Russell G. A universal character model and ontology of defined terms for taxonomic description. Pages 63-78 in Data integration in the life sciences (DILS 2004) Lecture Notes in Bioinformatics 2994 edited by E. Rahm. Berlin: Springer-Verlag.
25. Zhong Y, Jung S, Pramanik S, Beaman JH. 1996. Data model and comparison query methods for interacting classifications in taxonomic databases. Taxon 45: 223-241.
26. Zhong Y, Luo Y, Pramanik S, Beaman JH. 1999. HICLAS: a taxonomic database system for displaying and comparing biological classification and phylogenetic trees. Bioinformatics 15(2):149-156
27. Pullan MR, Watson MF, Kennedy JB, Raguenaud C, Hyam, R. 2000. The Prometheus taxonomic model: a practical approach to representing multiple taxonomies. Taxon 49(1): 55-75.
28. Ytow N, Morse DR, Roberts DMcL. 2001. Nomencurator: a nomenclatural history model to handle multiple taxonomic view. Biological Journal of the Linnaean Society 73: 81-98.
29. Koperski M, SauerM, Braun W, Gradstein SR. 2000. Referenzeliste der Moose Deutschlands. (Schriftenreihe Vegetationskunde 34). Bonn: Bundesamt für Naturshutz.
30. DWC 2004. The Darwin Core. http://speciesanalyst.net/docs/dwc
31. Geoffroy M, Berendsohn W. 2003. The concept problem in taxonomy: importance, components, approaches. Pages 5-14 in Berendsohn et al. 2003 [9].