Taxa, Taxon Names and Globally Unique Identifiers in Perspective.

Roger Hyam - March 2009
PESI WP4 Project Officer, Botany Natural History Museum London.
Research Associate Royal Botanic Garden Edinburgh.
email: roger@hyam.net

Introduction

There has been an explosion in the amount of biodiversity data available on the internet. This has lead to the possibility of mass collaboration between individuals and projects through on-line sharing of data. Currently the only way of combining the taxonomic part of data from multiple sources is by matching the scientific names used. This paper explores the repercussions of this approach and attempts to formalise taxonomic data exchange.

Nomenclatural Codes

The nomenclatural codes arose as a codification of the accepted best practise. Major advances were made around the turn of the twentieth century when the International Commission on Zoological Nomenclature was founded (1895) and the International Botanical Congress published the "International rules of Botanical Nomenclature" (1905). Today there are a number of nomenclatural codes and proposed replacement codes including the ICBN[1] (McNeill et al. 2006), ICZN[2] (Ride and International Commission on Zoological Nomenclature. 1999), ICNCP[3] (Brickell and International Society for Horticultural Science.;International Commission for the Nomenclature of Cultivated Plants. 2004), ICNB[4] (Lapage S and International Union of Microbiological Societies.;International Union of Microbiological Societies. 1992), PhyloCode[5] (Queiroz 2006). Of these by far the most significant for the majority of scientists are the ICBN and the ICZN. Discussion here is restricted to these two codes because of their significance but it should be noted that these codes probably govern the names of a minority of organism. The majority of the organisms on earth are bacteria whose names would be governed by the ICNB (May 1988).

Every scientific name that is validly published at the rank of species and below is directly bound to a preserved type specimen that is stored in a reference collection. Names above the rank of species are indirectly bound to type specimens via nominated type species. The type specimens bind the names used in the literature into the biological reality found in the reference collections.

The correct name for any circumscribed taxon is calculated from the first published name who's type specimen is included within the circumscription of that taxon, taking into account the placement of the taxon within a particular classification and the names in use for other taxa in that classification (the nomenclatural rule of priority ICZN Article 23 & ICBN Article 11). If there are no suitable name bearing type specimens within the circumscription then a new name is published and a type specimen assigned to it.

---

1 http://ibot.sav.sk/icbn/main.htm
2 http://www.iczn.org/
3 http://www.ishs.org/sci/icracpco.htm
4 http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=icnb.section.185
5 http://www.ohio.edu/phylocode/

There are many complex rules concerning matters such as binomial names (ICZN Article 5 and ICBN Article 23.1), missing type specimens (ICZN Article 74 and ICBN Article 9.9) and publication of duplicate names (ICZN Article Article 52 and ICBN Article 53.3) but these are not relevant to the discussion presented here. Only the fact that names are determined via an algorithm based on name bearing types is important to this discussion (ICZN Article 61 and ICBN Article 7).

The Taxon Concept Model

For over one hundred years the codes have been used to determine the names of taxa and scientists have cited names when they wanted to reference the taxa used in their experiments. Unfortunately this system is flawed. It is possible for radically different taxa to have the same name. Two taxa need only share a single specimen – the type – to have the same name (Figure 1). The converse is also possible – two taxa could share all specimens but the name bearing types. Minor changes in taxon circumscription between classifications can lead to disproportionate changes in nomenclature (Figure 2). It is also possible for a taxon to change its name without changing its circumscription – if it is discovered to include an older name's type for example. Scientists who cite names in their work therefore do not precisely cite the taxa to which their observations pertain. To precisely cite taxa they should include details of the taxonomic treatment they are employing as well as the name.

This separation between the notions of a name and the taxon to which that name refers is now widely accepted (Berendsohn 1995; Franz *et al* 2004; Kennedy *et al*. 2006; Pullan *et al.* 2000 and works cited therein). The problems that can arise can be illustrated with a simple example.

Take a hypothetical case where an ecologist wishes to combine two studies S1 and S2 in a meta analysis. The studies include taxon names that occur in two taxonomic treatments, T1 and T2. T1 recognises two taxa called A and B whilst T2 recognises only one taxon called A and places B as a synonym of it. The ecologist has to make a decision as to whether it is safe to combine data from S1 and S2. If S1 uses the name A and S2 the name B then it probably isn't safe without further investigation. The authors of S2 may have deliberately exclude material that the authors of S1 would have included. If both studies use the name A then it is still potentially dangerous to combine the data because if either study were based on a different treatment one would have specifically excluded material that the other included. The only safe case is where both studies use the name B as this indicates they were both using T1 and therefore the same concept of B. It is therefore difficult to automatically combine data on the basis of names alone without a taxonomic judgement being made – probably by a human. This example only considers two possible taxonomic treatments. If there are other treatments that recognise B as the accepted name for different taxa then either S1 or S2 could have used one of those other treatments. In such a case it wouldn't be safe to combine the studies even if they had used the **same** name. Indeed as it is impossible to know that all possible treatments have been found and accounted for (proving a negative) it is never totally safe to combine studies on the basis of taxon names even if those studies use the same name! This would appear to be a *reductio ad absurdum* and we should therefore examine what happens in real world data.

Names that are synonyms indicate that taxon concepts have changed – they once represented an accepted taxon that is no longer recognised. It may have been split into several parts and the type specimen now resides in another taxon with an older name that takes priority. Names that have synonyms also indicate changes in concepts. They were derived from combining multiple other taxa or at least the parts of those taxa containing type specimens. Names that are of accepted taxa and do not have synonyms can be considered 'pure' as they represent a single taxon concept for the

purposes of estimation (although even here it is possible to have multiple nested concepts that do not affect the names as mentioned above). The Catalogue of Life 2008 Annual Checklist[6] (Bisby et al. 2008) contains 1,192,015 accepted names and 720,040 synonyms, a total of 1,912,055 names. Of the accepted names 262,293 have synonyms whilst 929,722 (48.6% of the accepted taxa) lack synonyms and could be considered to represent single taxon concepts. This is a crude estimate as the Catalogue of Life is a synoptic work and may provide a simplified view of the global nomenclature. Many groups are only partially studied and need further work. Those that are included may not have full synonyms. Geoffroy and Berendsohn 2003 carried out a more detailed analysis on a smaller group, the "Reference List of German Mosses" Koperski et al. 2000, in which detailed examination of the relationships between twelve different classifications is presented. Geoffroy and Berendsohn conclude: "In terms of databasing this means that only for 13% of the taxa the name can serve as a direct index to other data (and this can only be said for the set of treatments scrutinised by the authors)." These two observations point towards the effect of taxon concepts being large and probably involving the majority of taxa. Despite this the use of a taxon concept based approach remains very rare. Researchers continue to report their results using taxon names alone, not indicating the taxonomic treatment they were following. It appears that the taxon concept effect although real has little detectable effect on the way the results of research are used.

In relatively recently surveyed areas of high endemism, for example New Zealand, where many taxa have only been described once and there haven't been multiple rounds of taxonomic revision the issues associated with taxon concepts as described above are likely to be be less significant. Unfortunately these areas are frequently prone to large numbers of misapplied names. This occurs when specimens are consistently identified to the wrong taxon, typically a similar looking European species. When this situation is detected the name of the European species effectively has two taxa associated with it, the 'right' one and the 'wrong' one. An analogy is a word with two meanings in two geographical locations.

Names as tags and 'Taxonomic Intelligence'

The taxon concept conundrum (that fact that the world gets by happily using names alone and not qualified taxon concepts) would remain an interesting sideline if it were not for the impact of the internet and the possibility of automatically or semi-automatically combining data from multiple sources. Large amounts of information is being tagged with taxon names and published on the internet. GBIF contains 140 million occurrence records linked to names, GenBank contains around one hundred million sequences on over 260,000 named organisms. Searching for a commonly occurring scientific name using generic search engines such as Google typically results in thousands of hits and hundreds of images. The iSpecies.org[7] website carries out this kind of federated search over multiple data suppliers automatically.

Some of the resources returned when searching for a name may provide synonymic names that can be fed back into the search process to retrieve more resources tagged with other related and presumably substitutable names. Projects, such as uBio[8], offer query expansion services to support this process. uBio also exploits published taxonomies such as Catalogue of Life to expand queries up or down a taxonomic hierarchy. The combined process has been termed 'Taxonomic Intelligence' (Patterson et al. 2006).

---

6   http://www.catalogueoflife.org/annual-checklist/2008/search.php
7   http://www.ispecies.org/
8   http://www.ubio.org/

Progress has been both rapid and remarkable. The extant systems are proving useful in discovery of research materials. Unfortunately these methods can only lead to query **expansion**. Using the example above a search for A will return all resources tagged with either A or B. Likewise searching for B will return all resources tagged with B or A. The fact that one of the studies involved may have deliberately excluded measurements of B is not detected. The process results in the broadest possible interpretation of any taxon. It takes the *sensu lato* approach for everything – the ultimate "lumpers" approach to taxonomy. This would occur even if the query expansion process permitted selection of a preferred classification over which the query were expanded because each synonym relationship has to be treated as a wholesale movement of a taxon not the movement of a type specimen alone. Splitting of taxa is by and large invisible.

The process of expanding from one tagged resource to other resources that bear the same or related names can be conceptualised as creating a graph of resources where each resource is a node and the edges are formed by matching the name strings in the resources. For some names this graph of linked resources may be small. For others it could be enormous. These graphs are effectively a new form of taxon circumscription. Through this mechanism the *de facto* species concept is becoming the "Google Species Concept" where a species is defined as anything that is returned by a search for its name.

The two most used statistical measures in information retrieval are precision and recall. Precision is the proportion of the resources returned that are relevant to the search terms used. Recall is a measure of the number or resources returned as a proportion of the number that should have been returned. If a search has 100% recall and 100% precision it will find all the relevant and only the relevant resources. The trouble with taking a taxonomic intelligence approach to query expansion is that relevance is not clearly defined. If study S1 has deliberately excluded taxon B but another resource (T2) says B is a synonym of A there is no way of excluding S1 from the search results for search term 'B'. Without knowing the classifications used by S1 not even a human can know to exclude it. Query expansion increases the likelihood that such results will be inappropriately included before a human can intervene.

Through the use of information retrieval techniques scientific names are effectively being used to build a 'Folksonomy'- a bottom up classification emerging from social tagging of data. An observation is tagged with a name and is published on the web. There is no differentiation between whether this is defining the taxon or referencing a predefined taxon. The boundary between identification and classification is not explicit. The observation tagged with a name may be a DNA sequence that will subsequently be used to confirm the identity of another 'unknown' sequence. No taxonomist needs to be involved in the process. Counter intuitively taxonomic revisions (whether monographic or flauristic/faunistic) are likely to make query results less reliable because they introduce new name to name relationships through synonymy and so increase the results returned but decrease the precision of information retrieval.

This "names as tags" approach provides compelling results in a short time span. It supplies usable systems for finding research materials. Its weakness is that it is likely to create misleading results if automated analyse requires other than a *sensu lato* view of every taxon. This includes even apparently trivial analyses such as plotting distribution maps. The danger is that precisely because these systems work so well for information retrieval people will assume that they are also suitable for research and further analysis. An analogy is using a high magnification but low resolution microscope. It is adequate for many tasks but may not reveal the details necessary for research. Unless individual taxa are tagged with Globally Unique Identifiers fine grained taxonomy will

never be accessible to detailed or automated analysis.

Globally Unique Identifies (GUIDs)

The importance of GUIDs can be best illustrated with a mapping example. Imagine a data supplier who publishes a million occurrence records. Each record contains a longitude, latitude and a taxon name. A consumer of the data is interested in a subset of records that form dots within a particular polygon on a map. When the supplier updates the data the consumer sees the dots on the map change. What has changed? If a dot has disappeared has it been removed from the data set or has its location changes? If a dot appears on a map has it actually been moved from elsewhere or is it new data? If the taxon name for a dot changes is it a new record or a re-determination of an old record? None of these questions can be answered unless the dots bear identifiers that are separate from their content. The data consumer is probably most interested in change - monitoring the effects of environmental impacts. Although information can be gleaned from the two maps it is only in the form of coarse measures and does not separate experimental noise from biological signal. An example might be "Data supplier X reports fewer occurrences of species Y from area Z this year than last". It is difficult to speculate as to why this effect has occurred without being able to "bore down" into the data and see why the changes have arisen.

If every record in the data set has an immutable identifier then it would be immediately apparent to the consumer which records had changed, which were additions and whether the changes may be for biological reasons. Furthermore if each record had an identifier that was unique within a global scope (was guaranteed not to clash with identifiers from any other source) then the consumer could mix records from different data sources and still be sure to pick out biological signal from data noise. The consumer could also individually credit the suppliers for use of their data.

For these reasons tagging data with GUIDs is the single most important step in enabling the sharing and integration of biological data across the internet.

From the point of view of the hypothetical ecological study introduced above if the taxa (not the names) in T1 and T2 had been tagged with GUIDs (as well as names) and the studies S1 and S2 had used the GUIDs to refer to the taxa (rather than just names) none of the ensuing complexity would have occurred. The ecologist could easily tell whether it was appropriate to combine results from S1 and S2. It would be reasonable for a machine to make this decision automatically and only bring matters to the attention of the ecologist if there was a conflict.

GUID Technologies

The term Globally Unique Identifier (GUID) is used in two slightly different ways. In computer science GUIDs are values that are complex strings of characters that are extremely likely to be unique in any context. In the biodiversity informatics community the term is used in a narrower sense. In this sense GUIDs have three related properties. They are not only globally unique they are also resolvable (or actionable) and identify a typed object .

Uniqueness

There are two principle ways of achieving global uniqueness. One is to generate a long and complex number that is highly unlikely to be generated twice and so is functionally unique. This approach enables distributed systems to uniquely identify data without significant central co-

ordination. The most common implementation of this approach is the <u>Universally Unique Identifier</u> (UUID) standard[9]. UUIDs are widely used in lower level computing applications such as distributed file systems. Another way to establish uniqueness is the use of a central issuing authority. An example of this approach is the Domain Name System (DNS). DNS is a hierarchical naming system for resources on the Internet including websites and email servers. The <u>Internet Corporation for Assigned Names and Numbers</u>[10] (ICANN) issues top level domain names such as .com to lower level issuing authorities who issue subdomain names (e.g. example.com) who then have authority over issuing subdomains of these domains. Theoretically this can carry on for up to 127 levels but practically rarely exceeds five or six. The domain names are then used in protocols such as the Hypertext Transfer Protocol (HTTP) which enable the addition of values after the domain name (e.g. http://example.com/123). DNS has been used to define what is termed a namespace. The locally unique identifier '123' is globally unique when it is combined with "http://example.com/".

Resolution

UUID type GUIDs would help solve the problem of the changing dots on the map given above. Each dot would have its own identifier and so any changes in the data could be distinguished from additions and deletions of dots. What UUIDs would not help with is the provenance of the data behind the dots. If the application needs to know more than the information provided to plot the map (such as the licensing terms or whether the data has been modified since issue) then it must be able to do something with the GUID to access the original data source. The GUID needs to be resolvable to some meaningful information. The identifier contains information which refers to data stored elsewhere, as opposed to containing the data itself. Accessing the value referred to by a reference is called dereferencing. An analogy of this process would be the citing of references in published works. The citations act as identifiers that can be dereferenced to the original papers in the library. Resolution of identifiers is only possible with some form of centralised authority with which the identifiers have been registered. Simply searching for identifiers does not provide authoritative data about GUID tagged data as it may result in multiple hits which may contain different data.

Typing

When a researcher fetches a referenced work from the library they are usually certain how to handle it. It will be a paper or book of some form probably in a language they can read or recognise. Likewise when a machine dereferences an identifier the response it receives needs to be understandable both syntactically and semantically so that it can display the response in an appropriate way or carry out further calculations. The GUID resolution therefore needs to be linked to some form of typing mechanism. If a machine is presented with a GUID it should, for example, be able to tell the users that the GUID represents a taxon from a particular taxonomic treatment.

Technologies

There is some debate over the use of GUID technologies in the biodiversity informatics community. This debate is on-going and some of the technologies involved are summarised here. Unfortunately it is not possible to avoid an 'alphabet soup' of acronyms when discussing these technologies.

---

9 <u>http://tools.ietf.org/html/rfc4122</u>

10 <u>http://www.icann.org/</u>

The most widely used identifiers on the internet are HTTP Uniform Resource Identifiers (HTTP URI) these include the HTTP Uniform Resource Locators (HTTP URL) used as addresses for web pages. HTTP URIs on their own provide uniqueness and resolution but not response typing. Following the set of best practices proposed under the banner 'Linked Data' does provide response typing however.

Life Science Identifiers[11] (LSID) were first proposed by Object Modelling Group[12] and IBM. After several workshops TDWG adopted LSID as its preferred GUID technology. They provide uniqueness, resolution and response typing. The default resolution mechanism is based on DNS (as with HTTP URIs) but there are very few clients that exploit it. Most LSIDs are resolved by being appended to the HTTP URL of a proxy program that fetches the associated data and metadata.

The motivation for TDWG choosing LSID over HTTP URI was principally social. HTTP URIs are considered inherently unreliable by many users because of their experience with broken webpage links and their ease of creation. It takes a conscious action on the part of administrators to implement LSIDs and this instils a sense of importance to their maintenance. A similar motivation is given for the adoption of Digital Object Identifiers[13] by the publishing community. DOIs have a similar resolution mechanisms to LSIDs using either an HTTP URI proxy or the Handle System[14].

Standards

Adoption of GUIDs for taxa has effectively stalled. The major providers of nomenclatural data (IPNI[15], Index Fungorum[16], ZooBank[17]) have issued LSIDs for names and return data using a version of the Taxon Concept Schema[18] (TCS) rendered in RDF[19] as an OWL[20] ontology. Work is currently underway to formalise this as a TDWG[21] standard but in the meanwhile GBIF[22] is using a checklist format based on the Darwin Core standard to exchange taxonomic data. There are proposals for a Global Names Architecture that will also track GUIDs for taxa but implementation of this has yet to start.

Without standard ways of identifying taxa and standard ways of encoding taxonomic assertions there is little impetus for the development of client applications which could better exploit the homogenous environment. This leads to little impetus for further labelling of taxa and the continued use of names strings.

Summary

Scientific, type based names are not adequate identifiers of taxa. Because of this the majority of names are ambiguous to some extent but there appears to be little movement to change working practice and refer to taxa by referencing specific treatments or using GUIDs. Names are widely

---

11 http://lsids.sourceforge.net/
12 http://www.omg.org/
13 http://www.doi.org/
14 http://www.handle.net/
15 http://www.ipni.org/
16 http://www.indexfungorum.org/Names/Names.asp
17 http://www.zoobank.org/
18 http://www.tdwg.org/standards/117/
19 http://www.w3.org/TR/rdf-primer/
20 http://www.w3.org/TR/owl-features/
21 http://www.tdwg.org/
22 http://www.gbif.org/

used as if they were social tags or key words. This leads to effective information retrieval applications but results in adoption *sensu lato* view of all taxa. Machine interpretation of taxonomic data will be imprecise until taxa (not names) are labelled with GUIDs. There is still debate around the best GUID technologies and response formats for names and taxa. This creates a chicken-and-egg situation it is difficult to see a way around.

A tacit assumption has been made here that taxa can be defined i.e. that there is some mechanism by which one could circumscribe a taxon so that another person could ascertain beyond reasonable doubt that a specimen was a member of that taxon. This mechanism is termed the description or circumscription of the taxon. It is a protocol to test the hypothesis "Specimen X is a member of taxon Y". Every taxon has a separate protocol. The protocols may use different morphological or molecular characteristics. Although we will be able to automate (through the use of taxon concepts and GUIDs) the tracking of taxa and specimens that have been identified to those taxa it appears unlikely we will be able to automate comparison of taxa between classifications because they are defined in unrelated ways. i.e. it will be possible to automatically determine that problems exists but there can be no automatic mechanism for resolving these problems.

References

Berendsohn, Walter G. 1995. "The concept of "potential taxa" in databases.." *Taxon* 44:207-212.

Bisby, FA et al. 2008. "Species 2000 & ITIS Catalogue of Life: 2008 Annual Checklist. CD-ROM." http://www.catalogueoflife.org/annual-checklist/2008.

Brickell, Christopher, and International Society for Horticultural Science.;International Commission for the Nomenclature of Cultivated Plants. 2004. *International code of nomenclature for cultivated plants : (I.C.N.C.P. or cultivated plant code) : incorporating the rules and recommendations for naming plants in cultivation*. 7th ed. Leuven Belgium: International Society for Horticultural Science.

Franz, Nico, M., Peet, Robert, K. and Weakley, Alan S. 2004. "On the Use of Taxonomic Concepts in Support of Biodiversity Research and Taxonomy" in Wheeler, Quentin D. "The New Taxonomy" CRC Press, Boca Raton, London, New York.

Geoffroy, Marc, and Walter G. Berendsohn. 2003. "The concept problem in taxonomy: importance, components, approaches." *Schriftenreihe für Vegetationskunde* 39:5-14.

Kennedy, J.B., R. Hyam, R. Kukla, and T. Paterson. 2006. "Standard Data Model Representation for Taxonomic Information." *OMICS: A Journal of Integrative Biology* 10:220-230.

Koperski, M., M. Geoffroy, W. Braun, and S.R. Gradstein. 2000. "Referenzliste der Moose Deutschland." *Schriftenreihe Vegetationsk* 34:1-519.

Lapage S, P, and International Union of Microbiological Societies.;International Union of Microbiological Societies. 1992. *International code of nomenclature of bacteria, and Statutes of the International Committee on Systematic Bacteriology, and Statutes of the Bacteriology and Applied Microbiology Section of the*. 1990th ed. Washington D.C.: Published for the International Union of Microbiological Societies by American Society for

Microbiology.

May, Robert. M. M., 1988. How many species are there on earth? Science (New York, N.Y.) 241 (4872), 1441-1449. DOI:10.1126/science.241.4872.1441

McNeill, J et al., eds. 2006. *International Code of Botanical Nomenclature (Vienna Code)*. International Association for Plant Taxonomy.

Patterson, David J., David Remsen, William A. Marino, and Cathy Norton. 2006. "Taxonomic Indexing--Extending the Role of Taxonomy." *Syst Biol* 55:367-373.

Pullan, M.R., M.F. Watson, J.B. Kennedy, C. Raguenaud, and R. Hyam. 2000. "The Prometheus Taxonomic Model: a practical approach to representing multiple classifications." *Taxon* 49:55-75.

Queiroz, K. 2006. "The PhyloCode and the distinction between taxonomy and nomenclature.." *Syst Biol* 55:160-2.

Ride, W, and International Commission on Zoological Nomenclature. 1999. *International code of zoological nomenclature*. 4th ed. London: International Trust for Zoological Nomenclature c/o Natural History Museum.
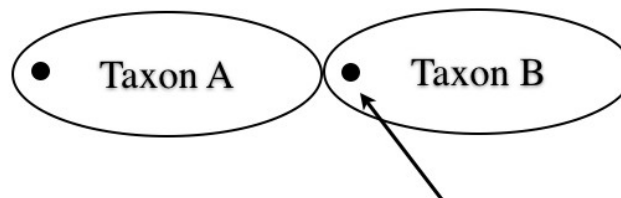
Figures

Figure 1: Dissimilar taxa (from different taxonomic treatments) can share a single unqualified name. (adapted from Kennedy et al ****)
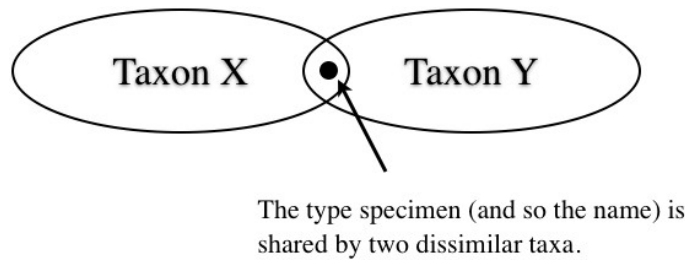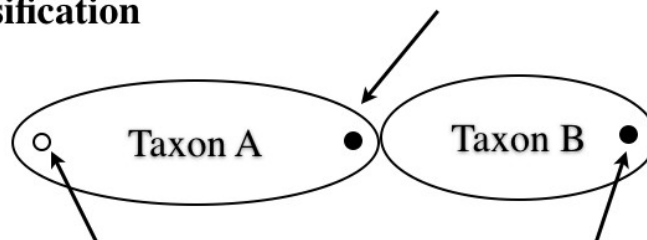

Figure 2: A slight change in circumscription moves the type of Taxon B into Taxon A.

In the revised classification A has B's name and B has a new name. (adapted from Kennedy et al ****)

[vector based version of the diagrams are available – need to know which fonts to use etc..]