The concept of "potential taxa" in databases

Walter G. Berendsohn[1]

Summary

Berendsohn, W. G.: The concept of "potential taxa" in databases. – Taxon 44: 207-212. 1995. – ISSN 0040-0262.

The concept of a "potential taxon" as a name- and literature-related data area in botanical databases is introduced. A potential taxon is a name with taxon circumscription information attached to it by means of one or more literature references. As a compromise solution between linking information in database systems entirely to specimen data or only to accepted names, using potential taxa can effectively preserve information links without hindering rapid database input and information processing. It is suggested that a potential taxon be cited by its name, followed by the abbreviation "sec." (for secundum = according to) and at least one of the literature references used to define it.

Traditional taxon concepts in databases

Databases used in fields as diverse as archaeology, biodiversity and environmental impact studies, gene sequencing, collection management, and pharmacognosy, to name but a few, make use of scientific plant names. With scant exceptions (namely, in taxonomic and nomenclature-oriented databases) these names are thought to stand for taxonomic groups (taxa).

Biological information is, in a general way, linked to taxa which in turn are designated by name. However, the notion of the essence of a taxon varies greatly, depending on context.

- From a purely nomenclatural point of view, taxa are containers for nomenclatural types which in turn are tags governing the application of the corresponding names. Whereas the definition (circumscription) of a taxon – as well as its rank and its position within the generic or specific classification – bear on the choice of the name that one must (or ought to) apply to it, the name itself provides no definition of taxon circumscription (although it defines the rank and position of the named taxon). In particular, the originial definition of a taxon (description or diagnosis, whether full or only rudimentary), while required for purposes of valid publication of a name, is largely irrelevant for the delimitation of the taxon (or taxa) to which the name is later on applied.
- The practising taxonomist will often use a rather vague taxon concept. Taxonomists can recognize a taxon in the field and in the herbarium, can convey this notion to others and tend to think of it as of a collection of objects, the sum of herbarium specimens and living plants they, and others before, have seen and memorized.
- Botanists working on an inventory or a Flora (and, in fact, all users of botanical names) will demand an operational definition, i.e. a set of criteria defining the

[1] Botanischer Garten und Botanisches Museum Berlin-Dahlem, Königin-Luise-Str. 6-8, D-14191 Berlin, Germany.

taxon, such as an identification key that enables to assign new material to an extant taxon.

- Finally, from a mathematical point of view, a taxon is a point in a multidimensional space formed by character states (Anderson, 1990) or, especially for higher taxa, character state combinations.

No specimens, no taxon?

The taxonomist correctly understands a taxon as a hypothesis, as a set of biological objects within a classification unit supposedly linked by phylogenetic descent, or as a set of criteria applying to such objects. In contrast, the user of taxonomic information tends to understand and use a taxon as if it represented a foregone conclusion. This is a misunderstanding because a scientific hypothesis must be testable and, depending on the result of testing, is bound to change. Specimens are the basic operational taxonomic units (OTU, see Dunn & Everitt, 1982). Consequently, the definition of a taxon should ideally include reference to all specimens used to form its concept and thus allow for re-examination of the taxonomist's conclusions. This is not realistic, if only for the sheer amount of such material. Floras and taxonomic monographs make a step in that direction by providing lists of exsiccata. However, often only selected specimens are cited, such lists being chiefly intended as vouching for a taxon's geographical distribution.

Moreover, even if all specimens used for the treatment were cited, the criteria employed to group specimens into taxa might still be debatable. In particular, a given specimen might be assigned to different taxa by different taxonomists; i.e., alternative taxonomies may exist.

Specimen based databases to the rescue!

If the botanical specimen is the basic unit of taxonomic study, "shouldn't we be databasing data rather than concepts derived from that data?" (White, 1994). Why not join efforts and build the ultimate database and expert system? It would store coded descriptive information on all objects (specimens and observations) used to form the idea of a specified taxon. Identification keys and the description of the taxon can then be derived from the data stored for individual specimens (see Maxted & al., 1993, for current methods). This will permit automatic placement of additional specimens in the appropriate taxon or, if they do not fit within the circumscription of any one existing, it will draw attention to a problem area and enable taxonomists to change the classification system accordingly.

Another strong argument in favour of a specimen-based system is that information exists that refers directly to specimens rather than to a taxon, e.g. data on herbarium labels or properly documented samples for chemical analysis.

As long as we do not insist on a taxon concept based on the intuition of the talented taxonomist (thereby effectively disqualifying taxonomy as a natural science), the data necessary for such a system could be obtained routinely. Also, there is no fundamental technical obstacle with which modern database theory could not cope. In fact, systems already exist which to a certain extent allow for such data processing, e.g. PANDORA (Pankhurst, 1991), which uses the DELTA format (Dallwitz, 1980, 1993; see also White, 1994) to store descriptive information.

While this may be a feasible approach for well-defined taxa worked upon by a single specialist, it does not solve the problem of alternative taxonomies. In many cases an unequivocal definition of taxonomic units cannot be achieved (Raven, 1974). Anyhow, the creation of a system as outlined would demand full global cooperation and absolute funding priority, which is an unrealistic prospect. Indeed, we do not have a name for many plant taxa, let alone a complete inventory of specimens already existing in herbaria all over the world. It is not even generally agreed that the effort to create a global specimen inventory would be justified, if it could be funded. Although some descriptive data may be obtained directly by automated specimen processing (cf. Molvray & al., 1993), most will have to be observed and registered by trained personnel. Last but not least, the taxonomic community has yet to agree on a minimum set of characters to describe higher plants.

Descriptive specimen-based data will therefore, for the time being, remain within the realm of the individual taxonomist investigating a given taxonomic group with its specific set of diagnostic characters. Even so, modern plant database systems must at least consider the option of including specimen information, and their structural design should provide linkage points between specimens and other information.

Taxon- or name-based databases – a real-world problem

Most of the databases involving plant names are not specimen-based. Information on plants is used and processed by many people outside the taxonomic community, who often will not even know what a specimen looks like. Plant information systems may eventually become the taxonomic "vehicle of communication", manageable also by the non-taxonomist, the lack of which Heywood lamented as early as 1984. Databases, in preference to books, are increasingly used by non-taxonomists to link their own data – often rather uncritically – to a particular taxon. Alternative nomenclature, synonymy, etc. make the value of this approach questionable: the nomenclatural principles alluded to above lead to ambiguity as to the limits of a taxon referred to only by name. The sole circumscription-relevant information provided by a name is that its type belongs to the named taxon. Beyond that common baseline, taxonomists may have grossly divergent views on what taxon a given name refers to. For example, *Ficus aurea* Nutt. is used by Adams (1972) for a species restricted to the Antilles and Florida. *Ficus aurea* Nutt., however, may turn out to include what has up to now been considered a group of "good" Central American species (C. C. Berg, 1989, in litt.). This is not yet too disturbing: information filed under what is now considered a synonym may be transferred to *Ficus aurea* sensu lato. But how about the reverse situation, when – perhaps based on clear new evidence of phylogenetic divergence and/or morphological distinctness – a previously recognized taxon is split? Or when divergent views persist? One of the narrowly defined taxa will again carry the same name as the entire group. Data linked to that name – e.g. about uses, resistance genes, chemical components, or floral biology – must then either be discarded as ambiguous, or re-investigated, or allocated on the basis of circumstantial evidence. Traditional databases provide no way to tackle this problem except by massive editing of the relevant information, often resulting in its effective loss.

Another important problem involves the creation and updating of databases using information from other databases or published inventories (cf. Bisby, 1993). More and more plant data are becoming available in electronic form: a recent inventory of databases in Latin America includes 109 such systems (P. Davila, in litt.) and the IOPI Common Directory Database of Databases currently holds information on 343 databases (R. Pankhurst, pers. comm., March 1995). New databases often rely on data imported from existing ones. If two different, overlapping data sets are imported, every name occurring in both has to be investigated as to possible conflicts in the taxon concept before the data can be incorporated. If the decision is to import the information unchecked and postpone such revision, a share of junk data is accepted that may greatly frustrate future users. The alternative is manual editing of the whole input, resulting in considerable effort and, often, a disappointingly small data quantity.

"Potential taxa", a workable compromise

Due to the inherent limitations of nomenclature a name may correctly designate several perhaps equally well-founded concepts of a taxon. For the purpose of information handling, a way has to be found to differentiate between different taxa bearing the same name. In an information system, this can be achieved by introducing a data element or data area which impartially mirrors alternative taxonomies, and allows for the inclusion of all information-bearing individual taxonomic concepts, including misnomers. The "potential taxon" is such an element. It is obtained by adding a circumscription reference to a name. The proposed notation for potential taxa extends that for misapplied names, for which the term "sensu" is in common use, and might make use of the designation "secundum" (according to; abbreviated "sec.") followed by an appropriate reference.

A database system using potential taxa is able to treat an unrestricted number of different concepts related to a specified name. In a relational database system, this does not pose a technical problem, because the entity type "potential taxon name" must have but three attributes: a pointer to a name, a second one to a circumscription and status intersection which handles the name status and the connection to the circumscription reference, and finally a pointer to an entity-type handling classification.

This solves the problem of the import of overlapping data sets. Incoming data records related to a taxon that differs in any important aspect from those already present are simply attributed to a new potential taxon, with the name of the source added as circumscription reference. Taxonomic editing may result in merging several such potential taxa into a new potential taxon, which is then credited to the person who effected the merge.

As real names, names of potential taxa may be treated as taxonomic synonyms if one defines "synonymy" in a wider than the traditional sense. Any name that has a circumscription reference may be a potential taxon, and the same name may stand for several potential taxa. Thus, "*Ficus aurea* Nutt. sec. C. D. Adams (1972)" may be treated as a pro-parte synonym of "*Ficus aurea* Nutt. sec. C. C. Berg (in litt.)" in the database system.

The problem of alternative taxonomic hierarchies is also solved. By default, potential taxa are classified in accordance with their circumscription reference. The genus "*Cecropia* Loefl. sec. W. C. Burger (1983)" is assigned to the "*Cecropiaceae* sec.

W. C. Burger (1983)" while "*Cecropia* Loefl. sec. Croat (1978)", for example, would be assigned to the family "*Moraceae* sec. Croat (1978)". However, in the introduction to *Flora of Barro Colorado Island,* Croat states explicitly that he follows the system of Dalla Torre & Harms (1900-1907). This being the case, "*Cecropia* Loefl. sec. Croat (1978)" can be assigned to "*Moraceae* sec. Dalla Torre & Harms (1900-1907)".

Practical implications and possible problems

A database system based on potential taxa is open to an inflation of records: any name referred to in a publication may form a new potential taxon. The potential taxon concept does not in itself provide a means of automatically or manually editing additional information. This causes no technical problem but information retrieval from the database may become difficult. Taxonomic monitoring of data input is needed to avoid this problem. A possible alternative to the potential taxon concept, to allow a single "alternative acceptable name" for every taxon (Bisby, 1993), would however impose an unbearable restriction on scientific data for non-cogent reasons of technical data management and is not acceptable for long-term information system planning.

The risk certainly exists that users of the information system be swamped by information they are not seeking. Access filters must therefore be designed, to shield the user from uncalled-for complexities and make the database appear like a traditional one-taxon-one-name database. Such a "preferred taxon view" of the data can be achieved by means of a user-definable hierarchy of preferred references.

A further problem to be solved is the merging of information from different sources. The potential taxon concept avoids the pollution of existing good data by newly introduced data by keeping them separate. For taxonomists, this feature provides the possibility of permanently storing a much larger portion of the result of, e.g., their literature searches in a database, thus making it accessible to others (see introduction in Maxted & al., 1993). As a fringe benefit, the problem of giving proper credit to information sources is solved, reference to them being permanently stored in the database.

A data model using the potential taxon concept does not inherently depend on a consensus view. At least, an agreed or consensus taxonomy (Bisby, 1993) is not a prerequisite for data entry. When agreement exists, users who enquire for a given potential taxon, defined e.g. by reference to a particular identification key, will be led by the system to the consensual taxon. The same system can thus service non-professional users as well as taxonomists who will normally request all information available.

Publicly available systems, such as the projected Global Plant Checklist of the International Organization for Plant Information (IOPI), should provide not only a consensus view at the generic level but a consensus classification. Again, this can be presented by means of a hierarchy of preferred references. Once achieved, the consensus is simply treated as a separate reference, to take priority over other views. Different classification may still be obtained at a switch, a feature particularly important at higher ranks, where alternative taxonomies tend to be prevalent and are widely used, e.g. for arranging botanical collections.

Does it work?

It does. A detailed data model incorporating the potential taxon concept has been developed in the process of drawing up the IOPI project plan for a Global Plant Checklist (K. Wilson & al., unpublished). The model was based on database projects at the botanical gardens of Berlin-Dahlem and La Laguna (El Salvador) and was developed in parallel with these projects over the last 3 years. The data structure was shown to be working using a prototype programmed in Microsoft Visual Basic based on the Microsoft Access database engine.

Acknowledgements

I should like to thank F. Bisby for focusing the discussion on the term and concept of "taxon" in the frame of the project plan for the IOPI Global Plant Checklist, and P. L. Nimis and B. Valdés for helpful comments on the taxon concept during meetings of the European Union concerted action "A Common Datastructure for European Floristic Databases". Comments by two reviewers, P. F. Stevens and especially J. McNeill, helped to clarify several points that would otherwise have remained obscure. I am indebted to W. Greuter for clarifying the nomenclaturist's point of view.

Literature cited

Adams, C. D. 1972. *Flowering plants of Jamaica.* Mona.

Anderson, L. 1990. The driving force: species concepts and ecology. *Taxon* 39: 375-382.

Bisby, F. A. 1993 Botanical strategies for compiling a global plant checklist. Pp. 145-157 *in:* Bisby, F. A., Russell, G. F. & Pankhurst, R. (ed.) *Designs for a Global Plant Species Information System*. Oxford.

Burger, W. C. 1983. Flora costaricensis. Families 54-70. *Fieldiana, Bot.,* ser. 2, 13: 1-254.

Croat, T. B. 1978. *Flora of Barro Colorado Island.* Stanford.

Dalla Torre, C. G. de & Harms, H. A. T. 1900-1907. *Genera siphonogamarum ad systema Englerianum conscripta.* Leipzig.

Dallwitz, M. J. 1980. User's guide to the DELTA system: a general system for coding taxonomic descriptions. *C.S.I.R.O. Austral. Div. Entomol. Rep.* 13.

– 1993. Progress report on the CSIRO DELTA programs. *Delta Newslett.* 8: 5.

Dunn, G. & Everitt, B. S. 1982. *An introduction to mathematical taxonomy.* Cambridge.

Heywood, V. H. 1984. The current scene in plant taxonomy. Pp. 3-15 *in:* Heywood, V. H. & Moore, D. M. (ed.), *Current concepts in plant taxonomy*. London.

Maxted, N., White, R. J. & Allkin, R. 1993. The automatic synthesis of descriptive data using the taxonomic hierarchy. *Taxon* 42: 51-62.

Molvray, M., Kores, P. J. & Darwin, S. P. 1993. Inexpensive digital data acquisition for morphometric study. *Taxon* 42: 393-397.

Pankhurst, R. 1991. *Practical taxonomic computing*. Cambridge.

Raven, P. H. 1974. Plant systematics 1947-1972. *Ann. Missouri Bot. Gard.* 61: 66-178.

White, H. 1994. Data or concepts – what should we be coding. *Delta Newslett.* 10: 3-14.