The Prometheus Taxonomic Model: a practical approach to representing multiple classifications

Martin R. Pullan[1], Mark F. Watson[1], Jessie B. Kennedy[2], Cédric Raguenaud[2] & Roger Hyam[1]
[1] Royal Botanic Garden Edinburgh EH3 5LR, U.K.
[2] School of Computing, Napier University, Edinburgh EH14 1D5, U.K.

Summary

Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C. & Hyam, R.: The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. - Taxon 49: 55-75. 2000. - ISSN 0040-0262.

A model for representing taxonomic data in a flexible and dynamic system capable of handling and comparing multiple simultaneous classifications is presented. The Prometheus Model takes as its basis the idea that a taxon can be circumscribed by the specimens or taxa of lower rank which are said to belong to it. In this model alternative taxon concepts are therefore represented in terms of differing circumscriptions. This provides a more objective way of expressing taxonomic concepts than purely descriptive circumscriptions, and is more explicit than merely providing pointers to where circumscriptions have been published. Using specimens as the fundamental elements of taxon circumscription also allows for the automatic naming of taxa based upon the distribution and priority of types within each circumscription, and by application of the International Code of Botanical Nomenclature. This approach effectively separates the process of naming taxa (nomenclature) from that of classification, and therefore enables the system to store multiple classifications. The derivation of the model, how it compares with other models, and the implications for the construction of global data sets and taxonomic working practice are discussed.

Introduction

A biological classification provides a means of identifying, categorising and referring to organisms. However, the complexity of the living world, and the wide variety of techniques for surveying it (phenetics, cladistics, etc.), mean that one cannot simply assume a single, common reference classification categorising all organisms. The same organism may at times be classified according to different taxonomic opinions and subsequently have several alternative names. Modern classifications are usually improvements on previous ones, but sometimes the existence of alternative or variant classifications reflects the fact that there is disagreement as to how to interpret the data on which the classification is based. This will become increasingly true with more extensive use of molecular data leading to new generic alignments. As alternative classifications multiply, biologists will commonly be faced with the need to compare and contrast them in order to identify how they differ in their organisation.

The use of computers in taxonomy has grown rapidly over the last decade. During this period a number of specialist databases have been implemented specifically for handling taxonomic data. As can be seen in Table 1, almost all of these systems are designed to handle only a single taxonomic view. This is because these systems take an over-simplified view of the relationship between nomenclature and classification (see also comments by Zhong & al. 1996, and Berendsohn 1995). The usual approach to handling taxonomic data has been to use names as identifiers of taxon concepts, with statements regarding the taxonomic status of a taxon assigned to the name. This unrealistically forces the adoption of a single consensus classification. Considering the increasing use of databases in botanical research and international policy making (e.g. the development of conservation strategies), we feel that these

Table 1. A selection of taxonomic database system

Database systems/models using single classifications	**References**
ALICE (ILDIS)	Allkin (1988), Allkin & Winfield (1989), http://158.43.192.14/town/square/fd95
ASC (*model only*)	Anonymous (1993), http://www.ascoll.org/
BG-BASE	Walter & O'Neal (1993), http://bgbase.rbge.org.uk/
BioCISE	Berendsohn & al. (1999)
BRAHMS	Filer (1994), http://www.brahms.co.uk/
CDEFD (*model only*)	Berendsohn & al. (1996), http://www.bgbm.fu-berlin.de/CDEFD/CollectionModel/cdefd.htm
CRIS	Anonymous (1994), http://www.nmnh.si.edu/cris
FLORIN	Anonymous (1998), http://www.florin.ru/florin
GRIN	Sinnot (1993), http://www.nal.usda.gov/ttic/coagra/grin.htm
HYPERTAXONOMY	Skov (1989)
ITIS	Anonymous (1995)
MUSE	Humphries & al. (1990)
PANDORA	Pankhurst (1991, 1993), http:/www.rbge.org.uk/pandora
PLANTS (USDA)	http://plants.usda.gov/plantproj/plants
PRECIS	Gibbs Russell & Arnold (1989)
SMASCH	Duncan & al. (1995), http://www.calacademy.org/smasch.html
SYSTAX	http://www.biologie.uni-ulm.de/systax
TAXON OBJECT	Saarenmaa & al. (1995)
TROPICOS	Crosby & Magill (1988), http://mobot.mobot.org/
ZOE	http://www.keil.ukans.edu/~neodat/muse.html
Database systems/models incorporating multiple classifications	**References**
IOPI ('potential taxon' concept)	Berendsohn (1995, 1997)
HICLAS ('taxon view' concept)	Zhong & al. (1996), http://aims.cps.msu.edu/hiclas/home.html

limitations are in fact driving decision-making concerning the standardisation of taxonomic treatments and creating a false impression of the state of taxonomic knowledge. This compromises the scientific integrity of many data sets currently under construction, and is an area which requires serious and immediate consideration.

The solution of course, is to produce a system that will support all views of taxonomic classifications without forcing a judgement as to which are 'correct'. Such a system must be able to handle multiple classifications arising from the combination of historical data, newly described taxa, new revisions and conflicting opinions in an unbiased manner.

Both Zhong & al. (1996) and Berendsohn (1995, 1997) have proposed models for handling multiple classifications, although they have tackled the problem from somewhat different perspectives and with different objectives in mind. The HICLAS model proposed by Zhong & al. appears to have been constructed as a tool for the working taxonomist, allowing them to represent and compare various different classifications in terms of the operations performed on existing concepts. However, this is carried out without a specific representation of the underlying taxonomic concept and without considering how data relating to names (and not taxon concepts) can be stored. This limits its usefulness in the broader context of storing taxonomic information.

The IOPI model proposed by Berendsohn (1997) takes a broader view and is intended to provide a framework for general taxonomic information systems. However, it is designed only to be able to represent existing classifications, and does not allow for comparison or manipulation of taxon concepts. The IOPI model recognises the importance of circumscriptions in differentiating classifications, however, comparisons between taxon concepts cannot be made as there is no explicit representation of these circumscriptions.

The Prometheus model provides a mechanism for both representation and manipulation of taxon concepts. Taxonomists will be able to undertake new revisions using detailed circumscription data, whilst using the same system non-specialists can search for botanical information (e.g. distributions, descriptions, images, DNA sequences, etc.) simply using plant names. When making queries using names, users will be made aware of alternative classifications associated with that name, and can elect to view the results using one or more of these. In doing this we avoid creating a false impression of the state of taxonomic knowledge, and yet to a large extent shield the non-specialist from the underlying taxonomic detail.

Returning to first principles we considered the taxonomic process in detail and modelled taxon concepts in terms of the actual data on which they are based (often groups of herbarium specimens). We believe that this approach has more effectively separated the nomenclatural process from that of classification, and therefore more closely models taxonomic working practice than any other published model. Furthermore, the separation of the processes of nomenclature and classification, and implementation of the automatic naming of taxa, allows the model to be used as an experimental tool with which a taxonomist can manipulate taxon concepts without regard to the names of the concepts, therefore avoiding unintentional bias. The automatic naming of taxa also provides a mechanism for verifying the nomenclature previously applied to existing taxonomic concepts rather than merely echoing the nomenclatural assertions of the author of the classification, which is the case in the HICLAS and IOPI models.

In the following sections we explain how names and taxa are represented in the Prometheus model, how the relationships between taxa are represented, and contrast our model to those already published. We start by considering the processes involved in a traditional taxonomic revision.

Taxonomic Revision Process

The processes involved in the production of taxonomic treatments are well established and detailed accounts of them have already been published (e.g. Watson 1997). Here we present a distillation of these accounts and include only those elements of the process that are relevant to our argument. These are:

1. The ‘taxonomic process’ at the level of species and below is specimen based (also including other elements e.g. illustrations, all hereafter referred to as 'specimens').
2. The ‘taxonomic process’ above the level of species is taxon based.
3. The result of the ‘taxonomic process’ is a hierarchical set of nested groups of specimens and/or taxa. These nested groups are the *only* explicit, testable representation of the circumscription of the taxa they represent.
4. The ‘taxonomic process’ usually manipulates and refines existing taxonomic concepts, both as a starting point for the delimitation of individual taxa, and as a means of delimiting the bounds of the study group. The results of a revision of a group can therefore only be

studied within the context in which they were created (see later comments on limiting the scope of classifications).

5. Taxa can only be named after the groups have been formed and the distribution and priority of the nomenclatural types have been examined: the processes of naming and classification are independent. Indeed this concept is the basis for Principle II of the International Code for Botanical Nomenclature (the *Code*; Greuter & al., 1994).
6. Relationships between a taxon and other taxa in terms of synonymy can only be determined after the classification process and are a consequence of that process. Except in the case of simple synonyms where one taxonomic concept is completely subsumed into another, a complete set of taxonomic (heterotypic) and nomenclatural (homotypic) synonymic relationships cannot be determined solely through examination of the distribution of types; pro parte synonyms can only be detected through comparison of the entire specimen content of alternative taxon concepts.
7. Descriptions can only be generated after the groups have been formed. In this way the descriptions do not represent the circumscription of the taxon but are rather a product of it. Without supporting lists of specimens, descriptions only represent generalisations of the taxonomist's taxon concepts and are subject to unintentional bias and misinterpretation. They may be accurate but they will always be imprecise.
8. Identification of specimens does not contribute to the overall classification process unless it is performed as part of a taxonomic revision and can be viewed in the context of the other specimens with which it is grouped. This means that publications such as checklists, and Floras that do not cite specimens, do not contribute to classifications. A distinction should be made between data obtained from such sources and data that makes explicit statements about the delimitation of and relationships between taxa (e.g. monographs, revisions and monographic Floras).

How do existing models relate to the taxonomic process as described above

PANDORA (Pankhurst 1993) was the first taxonomic database to truly recognise the hierarchical nature of taxon concepts within the underlying taxonomic model. However, this system made no distinction between the processes of naming and classification and hence, like all the systems before, could only represent one taxonomic view. It was also the first taxonomic database (as opposed to a collections management system, such as BRAHMS or BG-BASE) that recognised the importance of specimens in the 'taxonomic process'. Mechanisms were provided for grouping specimens according to taxon and generating descriptions of the taxa on the basis of the constituent specimens (Pankhurst & Pullan 1996). It is important to note that in the PANDORA model the specimens were not considered as *defining* the taxon rather as being *attributes* of the taxon and so could only be viewed in the light of a single taxonomic framework.

The 'potential taxon' concept of Berendsohn (1995) was the first recognition of the need to separate the processes of naming and classification in order to represent multiple classifications in a database. This, coupled with the idea of linking taxon concepts in a hierarchical structure, formed the basis of the taxonomic side of the IOPI data model (Berendsohn 1997). Prior to publication of the IOPI model, Berendsohn (1995) recognised that the definition of a taxon should ideally include reference to all specimens used to form its concept. He considered the use of specimens as a mediator for taxonomic data in this way as being impractical. In the light of this conclusion, the 'potential taxon' was proposed as a "compromise" and consists simply of a link to a taxon name, and one or more links to references where the taxon is circumscribed and/or assigned a taxonomic status. This allows instances of the use of the same name in differing contexts to be distinguished and so provides the basis for storing multiple classifications. There are, however, a number of limitations to this approach. Firstly, as names are directly linked to taxon concepts this means that the IOPI model does not fully separate the processes of naming and classification. Secondly, no representation of the circumscription is stored: the system is concept-based (not specimen-based), and therefore not capable of comparing taxon circumscriptions. Thirdly, no definition is provided as to what constitutes a circumscription, therefore any reference to a name may and probably will become a new ‘potential taxon’. In cases where no objective circumscription information is given, and hence where no real distinction between taxon concepts can be made, ‘potential taxa’ would proliferate to no good purpose. Berendsohn’s (1997) solution is to use taxonomic experts to decide when a reference to a taxon name warrants the creation of a new potential taxon. However, by requiring this level of intervention, the model ceases to be able to provide a totally impartial view of the data. For this reason we feel that it is important to distinguish between data that contribute to classification and data that do not. We therefore conclude that the ‘potential taxon’ concept provides a good basis for the representation for multiple classifications, but needs refinement.

The HICLAS system of Zhong & al. (1996) takes a completely different approach to the representation and storage of multiple classifications, although the basic unit of the system, the ‘taxon view’, is conceptually similar to the ‘potential taxon’ concept of Berendsohn (1995). A ‘taxon view’ consists of a taxon name plus an indication of where, when and by whom it was published. Based on the premise that “new classifications are usually built by sharing, changing and tuning taxonomic concepts of existing classifications”, the model allows the management of lineage relationships between taxon views. In the HICLAS model it was recognised that only certain types of taxonomic information contribute to classification. This contrasts with the 'potential taxon' idea in the IOPI model where almost every recorded use of a taxon name would require the creation of a new 'potential taxon'. Hence, the HICLAS model does not suffer from the problem of proliferation of ‘potential taxa', as in essence it only deals with 'real taxa'. Zhong & al. (1996) have not, however, explored how data that do not contribute to classification should be related to the various classifications they store. Therefore the HICLAS model is of limited use as a general taxonomic information database. The HICLAS model like the IOPI model does not attempt to store information regarding the circumscription of taxa. Although the HICLAS system is capable of tracking the operations involved in the taxonomic process, insufficient information is stored to allow the consequences of those operations (i.e. cross-classification comparison) to be properly explored. Furthermore, as most authors do not make explicit statements regarding these operations, the information required by the HICLAS system can only be obtained by later interpretation of the data source. By and large the apparent operations will be entered into the HICLAS system by a third party and so will be subject to misinterpretation. These factors limit its usefulness as a tool for the working taxonomist.

From the above paragraphs it is clear that neither the HICLAS nor IOPI taxonomic models store taxonomic data in a completely objective manner. We now describe a model that incorporates and combines many of the aspects of the models described above and yet addresses the shortcomings.

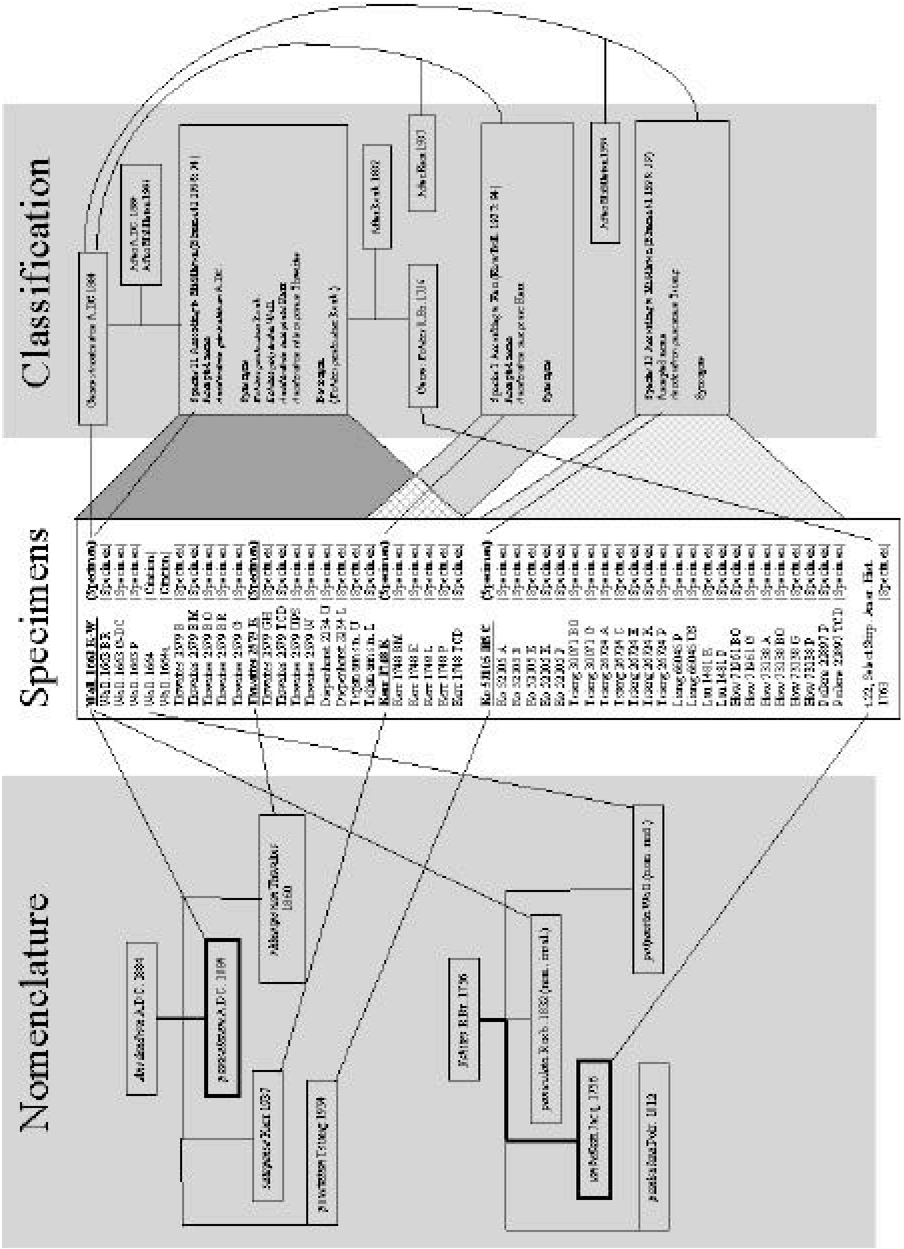


Fig. 1. An illustration of how taxonomic concepts can be represented and compared through examination of the specimens included in taxa.

The example here, taken from Middleton 1996, shows some of the results of a revision of the genus *Anodendron*. Three species rank taxa are shown on the right hand side of the diagram. Two species remain after the revision and one (species 1 according to Kerr) is subsumed into species 11 according to Middleton 1996. The taxa on the right hand side of the diagram can be related to the names shown on the left hand side of the diagram by examining the type specimens (underlined and bold in the centre of the diagram). The type species of the genera *Anodendron* and *Echites* are shown by the thick lines on the left hand side of the diagram.

The Prometheus Model

Using specimens to circumscribe taxa. - We have discussed the limitation of taxonomic database models that omit the circumscription of taxa, and we have indicated that it would be possible to circumscribe taxa in terms of the specimens and subordinate taxa that have been explicitly included in a published account of a taxon. We must, however, justify this assertion. It is a widely held belief that the circumscription of a taxon can be encapsulated in the description of a taxon. Traditionally this has taken the form of a written account of the 'relevant' features of the taxon, although formats for encoding these descriptions for computational purposes also exist (e.g. DELTA, Dallwitz & al. 1993). Descriptions are not fundamental to the taxonomic process and regardless of the manner in which they are stored or presented, they suffer from the following weakness. Unless a list of specimens from which the descriptions have been generated is published along with the description, then the assertions made in the description are not testable, and the characters used in the description are open to misinterpretation if not precisely defined (e.g. broad statements such as 'leaves hairy'). It would also have to be assumed that only this set of specimens was used to generate the description, and that the description did not also include elements derived from a taxonomist's mental taxon concept. This is often not the case and therefore descriptions are not guaranteed to be objective. Moreover, the character sets used to classify taxa vary from classification to classification, thus preventing direct comparison of classifications based on descriptions alone.

We conclude that the only objective and testable mechanism for defining taxa is the list of specimens or subordinate taxa that was given when a taxon concept was published (see Fig 1). Berendsohn (1995) also recognised this, but considered it impractical to use specimens in this way. He did not expand on his reasoning, except to state that the sheer quantity of information required would be too great to incorporate in any large-scale database. We disagree with this and believe that by taking a pragmatic approach to segregating data which do contribute to a classification from those which do not, then no more information, above that required for the production of a standard taxonomic publication, would be required to represent a taxon. Following this line of reasoning we would argue that the list of specimens examined during a revision and then published in a taxonomic revision should be as complete as possible.

Separating nomenclature from classification. - Our first aim in designing this model was to adequately separate the concepts of nomenclature and classification. In order to achieve this we have elaborated two basic entities: the Nomenclatural Taxon (NT) and the Circumscribed Taxon (CT).

The Nomenclatural Taxon. - The 'Nomenclatural Taxon' (NT) is the basic building block of the model. It is a container for the minimum amount of information required to represent a scientific name according to the rules of the *Code*. The required elements of an NT are as follows:

- Rank of the NT
- Name element
- Nomenclatural placement of the name element
- Type definition
- Author, place and date of first publication of the name
- Nomenclatural status
- Nomenclatural conservation or rejection status

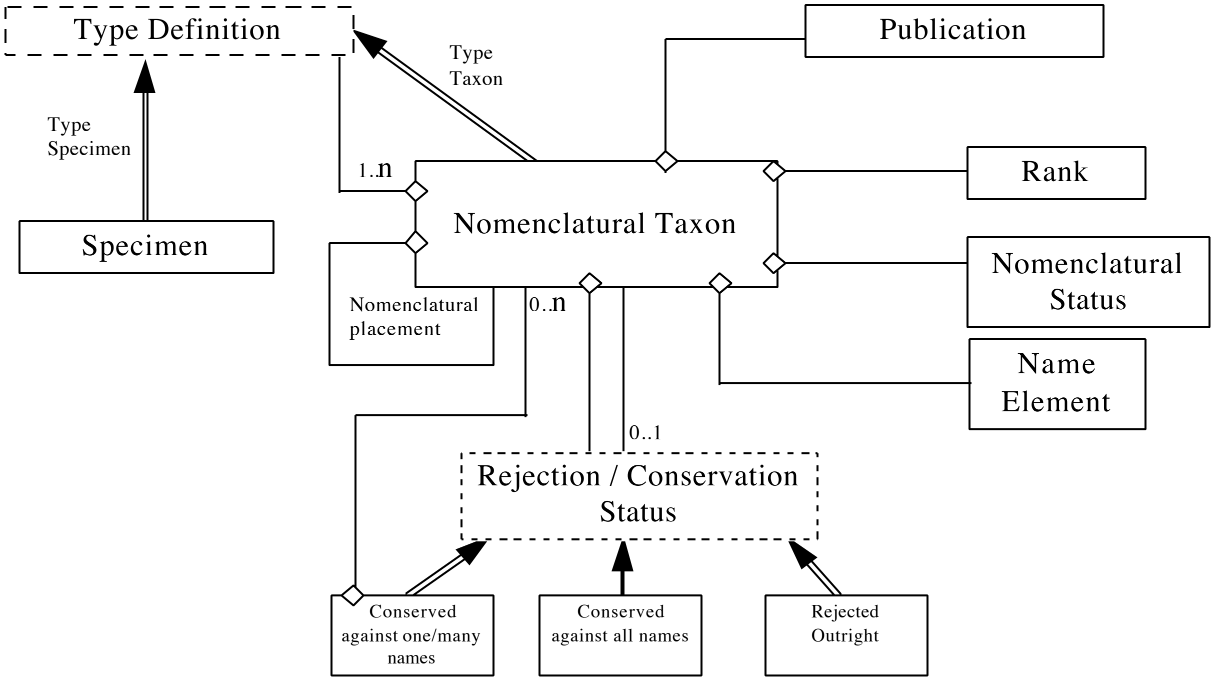
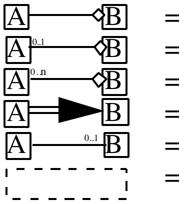


Fig. 2. An object model* illustrating the relationships between the elements of Nomenclatural Taxa.



B *must* have one and only one A
B *may* have one but only one A
B *may* have one or more A's
A is a kind of B
A may have a single association with B
Abstract concepts

* The object model follows a UML (Rumbaugh *et al.* 1998) style of notation. It should not be interpreted in the same way as an

entity relationship model.

The relationships between the elements of an NT are summarised in Fig. 2.

The ***rank*** assigned to an NT determines the allowed behaviour of the NT. The rank definitions are therefore fundamental to the operation of the model. The rank controls whether or not the certain types of link can be made to or from an NT and determines what kind of type information is required to define the name.

The ***name element*** is the part of a name that applies at the rank assigned to the NT, e.g. for a species this will be the specific epithet, for a genus the generic name.

The ***nomenclatural placement of the name element*** (the linkage to another NT of higher rank) is given only when required for nomenclatural completeness, and this requirement is governed by the rank of the name. For example, an NT of rank species should be linked to an NT of rank genus in order to be able to build the correct binomial. It should be noted that this placement link does not represent a taxonomic opinion it is merely a record of the use of that particular combination of genus and specific epithet. Where no extra information is required for nomenclatural completeness, such as with generic or familial names there would be no indication of the placement of the name.

The nature of the ***type definition*** will depend upon the rank of the NT. At or below species rank the type will be a reference to one or more specimens (including all materials that can be used as type material: herbarium specimens, illustrations, etc.). The kind of type (holotype, lectotype, syntype, etc.) is stored as an attribute of the type material.

The types of all names are specimens or corresponding elements. Generally, above the rank of species, a name of a taxon will serve to indicate that type. In the Prometheus model the type of these names is indicated by linking the NT that represents the name to a subordinate NT: thereby forming a chain of NTs. Following this chain of NTs down to a name at species rank, the real type (specimen or illustration) can be found. In exceptional circumstances the *Code* (Art. 10.4) makes special provision for the conservation of the type of the name of a genus by a specimen or illustration. In our model we handle this by allowing an NT of generic rank to link directly to its type specimen.

The ***author, place and date of publication*** are required in order to be able to uniquely refer to a name and to obtain the correct name for a circumscribed taxon (see below) by application of the *Code*.

An explicit statement of the ***nomenclatural status*** of a name is required as it would be impractical, if not virtually impossible, to derive this information within the system. Names flagged as invalid or illegitimate will not be included in the automatic assignment of names to taxa.

The *Code* makes provision for the ***conservation and rejection of names***. This includes names that are conserved against all other names (App. IIB of the *Code*), names that have been conserved specifically against one or more other names (App. IIA and IIIA), and names that have been rejected outright (App. IV and V). All conserved names at the rank of family and genus are conserved against all homotypic names at the same rank, and all conserved names are subject to priority when competing with other conserved names. The conservation/rejection status of a name is indicated by a flag, and in the cases where the code makes an explicit statement regarding the relationship between a rejected name and a conserved name this shown by a link between the appropriate NTs.

Circumscribed Taxon. - A 'Circumscribed Taxon' (CT) is conceptually different to an NT even though they contain some similar information. A CT contains the representation of taxonomic opinion, i.e. the circumscription of the taxon. The circumscription is expressed in terms of either groups of specimens or groups of subordinate CTs. The required elements of a CT are as follows (see Fig. 3):

- Rank of the taxon
- Circumscription details
- Ascribed name
- Author and date (and publication details if published)

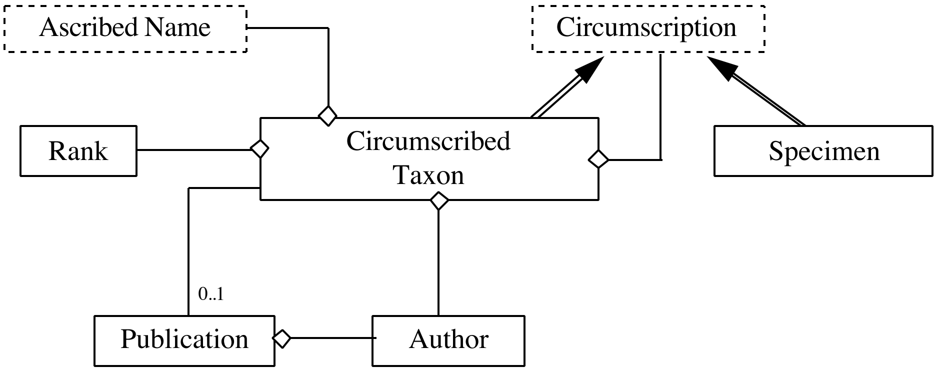


Fig. 3 An object model illustrating the relationships between the elements of Circumscribed Taxa. See Fig. 2 for details of notation.

The relationships between the elements of a CT are summarised in Fig. 3.

The ***rank*** assigned to a CT determines the allowed behaviour of the CT. The rank controls whether or not certain types of link can be made to or from a CT and which rules should be applied when determining the correct name (referred to by us as the 'calculated name').

The ***circumscription*** of a CT is a list of either specimens or subordinate CTs that delimit the CT. The 'calculated name' of a CT is obtained by examination of the circumscription list. Naming is therefore dynamic and depends upon the presence of types in the circumscription and their relative nomenclatural priorities.

The ***ascribed name*** is the name given to the taxon by the person whose view is represented. For published classifications this will be the scientific name used by the author, but for unpublished work in progress this may be an informal name invented by the worker. A nomenclaturally correct name is automatically obtained by the system based on the types included in the circumscription. In certain cases this 'calculated name' will differ from the ascribed name, e.g. through error of the taxonomist.

The ***author*** is the person whose taxonomic view is being represented. In the case of working (unpublished) groups this will be just a name and date, but when representing published classifications the literature citation details must also be included (as for an NT). One individual may have published several circumscriptions of the same taxon at different times, so precision is important.

The relationships between NTs, CTs and specimens are illustrated Fig 4.

How classifications are represented. - A classification is represented by the relationships between CTs, i.e. the fact that a taxon is a member of another taxon of higher rank is indicated by a link between the appropriate CTs. The nested nature of the hierarchy is achieved by allowing multiple CTs to be subordinate to another, e.g. a genus rank CT may have multiple species rank CTs subordinate to it. Each separate classification that is represented in the system is represented by a separate hierarchy of CTs, and as explained below, unless explicitly stated in the data source there should be no links between CT hierarchies from different data sources (see 'limiting the scope of classifications' below).

Automatically naming CTs. - The automated process of applying the correct name to taxa can only be undertaken after the CTs have been formed and grouped into a nested hierarchy. Once this has been achieved the CTs can be automatically named through the application of the relevant Articles of the *Code*. We have reduced the process down to the following simple algorithm.

1. Find all the type specimens included in all subordinate CTs.
2. Find the NTs associated with these type specimens that are directly or indirectly types of NTs at the appropriate rank.
3. From these select the NT with the earliest validly published name element. Due regard should be given to names that are not to be used (e.g. names that have other names conserved against them, or those that are formally rejected in the *Code*). This is the name element for the 'calculated name' of the CT: any name elements of lower priority, but included in the same taxon, become synonyms of the 'calculated name'.
4. For name elements that involve concatenation with other names in the construction of the full scientific name, a check is then made on the NT side for previous publication of that full name. If, for example, a specific epithet is placed in a genus for which no previous publication of the name exists, then the need for formal publication of the new combination is highlighted
5. In the case where a CT has been defined, and there are no declared holotypes or lectotypes included in the definition, then either a lectotypification is required or a new taxon has been created (a name must be given to the new taxon through publication and type declaration).

Synonyms. - Unfortunately this algorithm does not fully capture the process of identifying and categorising synonyms. The above process is purely nomenclatural whereas the concept of synonymy has elements of classification as well as nomenclature. In reality synonymic taxa are taxa whose circumscriptions overlap. These can be divided into two categories, full synonyms where one taxon concept is fully subsumed within another and pro parte synonyms where only part of a concept has been included in another. In the latter case the part of the concept included in another may not include the type, in which case the synonymic relationship will not be detected through nomenclatural processes alone.

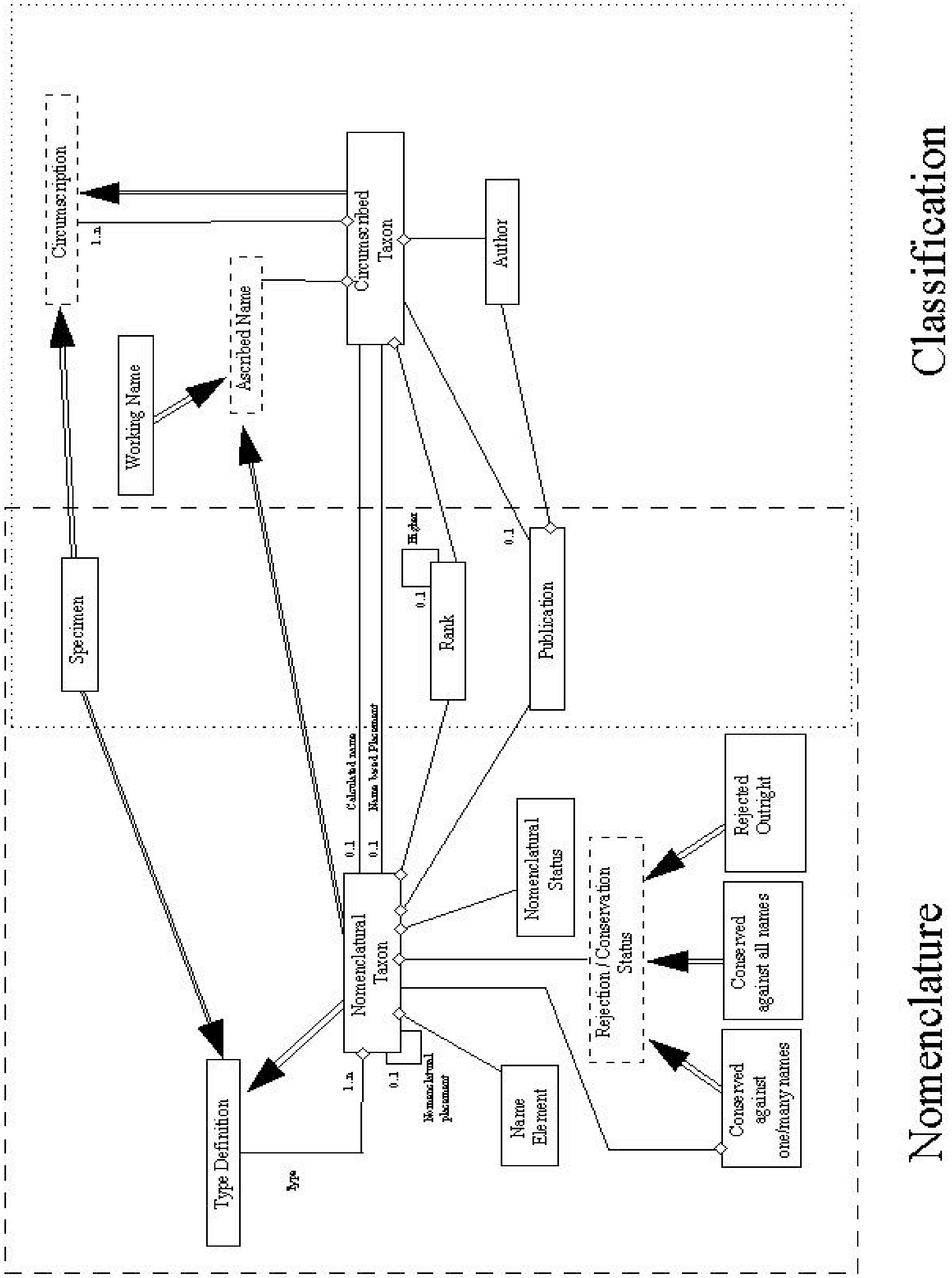


Fig. 4. An object model illustrating the relationships between Nomenclatural Taxa and Circumscribed Taxa. (See Fig 2 for details of notation)

Within the framework of our model we can see two possible mechanisms for handling synonyms. The first is by inspection. That is, for every specimen and/or taxon that is included in the circumscription another a check must be made as to whether or not that specimen/taxon has been included in the circumscription of another taxon. In this way a full list of synonymous taxa can be obtained. The circumscription of each synonym must then be compared with that of the current taxon in order to ascertain whether it is a full or a *pro parte* synonym.

The second approach is to record the operations performed on taxa during the revision of a group (as in the HICLAS model, Zhong & al., 1996). For example if a taxon is split to form two new taxa, the splitting operation would be recorded and used to relate the three taxa. In contrast to the HICLAS system, circumscriptions would also be split thus conveying not only a description of the operation but also the consequences of that operation.

The second approach has its advantages in that it is far less computationally demanding. However, it would be limiting, as it would result in the loss of the experimental elements that we were trying to incorporate in the model. For example, we would lose the ability to discover hitherto unrealised relationships between taxa. For this reason we prefer the first approach and current research is directed at developing a database system that is able to efficiently accommodate this model.

Basionyms and replaced synonyms. - There is a special class of relationships between homotypic names not dealt with in the previous section, and that is the relationship between a basionym and its subsequent combinations, or a replaced synonym and its avowed substitute. The IOPI model treats basionyms as a class of synonym. We feel that this is inappropriate because the taxonomic status of the basionym relative to its subsequent combinations depends upon the context in which the information is being viewed (i.e. the basionym could be accepted in one classification and a synonym in another). Furthermore, in contrast to normal synonyms, the relationships between basionyms and their combinations are purely nomenclatural and do not convey any information on classification. For this reason the relationship between a basionym and its combinations should be treated separately (on the NT side) and can and should be calculated automatically. Because Prometheus is the only model that truly separates nomenclature from classification this is a unique way of handling basionyms.

A basionym is the earliest legitimate, validly published association between a scientific name and a type. Subsequent use of the same type but in different contexts (e.g. moving a species from one genus to another), results in the generation of a new combination, i.e. an appropriately modified scientific name. In order to track these relationships using the NT model it is merely a matter of comparing the dates, validity and legitimacy of all the NTs based upon the same type.

The same arguments apply to the relationship between replaced synonyms and the corresponding avowed substitute (*nomen novum*).

Limiting the scope of classifications. - The hierarchical approach to the representation of classifications is not new and was first implemented in the PANDORA database system (Pankhurst 1993). As PANDORA is only capable of storing a single classification it avoids the problem of having to define the scope of a classification, and it is quite acceptable to link taxa as far up the hierarchy as desired. However, when dealing with multiple classifications it should be possible to link up the hierarchy only as far as the available information will allow. This can best be illustrated by considering the typical contents of the revision of a genus that relate to classification:

1. A brief history of the genus, giving a review of the placement of the genus in higher taxa and the type of the generic name.
2. A key to the subordinate taxa within the genus.
3. A list of types included within the circumscription of each of the subordinate taxa.
4. A list of all the specimens studied under the taxa to which they belong.

From this account we can see that individual taxa are well circumscribed in terms of their specimens. The genus itself is well circumscribed by the list of subordinate taxa. Although an indication of the higher taxon is given, there is no further information presented regarding the circumscription of the higher taxon. In the IOPI model (Berendsohn 1997) a new ‘potential taxon’ would be constructed for the higher taxon indicating that there is some circumscription information to be gained from this reference. Although the IOPI model will allow incomplete hierarchies of 'potential taxa', a 'potential taxon' can be linked with a 'potential taxon' of higher rank even if they were the result of different taxonomic works. We believe this extrapolates the information available far beyond what can be supported by the original sources and in the Prometheus model only CTs from the same taxonomic work can be linked to form a taxonomic hierarchy. In the example above the reference to the taxon in which the genus is placed is merely a name-based reference (see previous section), and in the Prometheus model the placement linkage should run back to the NT side. In effect this means that in the Prometheus model, not only will there be no complete classification tree on the NT side, there will also be no complete classification tree on the CT side. Imposing these limits will prevent the proliferation of informationless CTs, giving a clear and concise representation of the classification information available at any given level in the hierarchy.

Mechanisms for generating NTs and CTs. - Within the model there are two mechanisms for generating NTs and CTs we have called these the *representational approach* and the *experimental approach*.

- *Representational Approach:* taken when it is required to make a representation of a published classification. Under these circumstances it will be necessary to represent all the names in the classification. This is achieved by first generating the appropriate set of NTs, then creating a set of CTs, one CT for each taxonomic opinion contained within the publication. Each CT is linked to an NT in order to record the ascribed name used for the taxon that the CT is intended to represent. The circumscription will initially be restricted to the defining type specimen/taxon, but this must be expanded to include all the specimen information included in the original data source.

Revisions that build on past classifications can be undertaken using the representational approach. A representation of the past classification is entered, a copy of that classification made, and the copy modified to reflect the changes in taxonomic opinion between the two classifications. The operations involved in making the modifications do not have to be explicitly recorded (cf. HICLAS, Zhong et al. 1996), but can be elucidated from the changes in taxon circumscriptions. We feel that this is a better way of recording these operations than in the HICLAS model where assertions regarding these operations are made without underlying evidence to support them (see earlier comments).

- *Experimental Approach:* adopted when a rearrangement of specimens and taxa is to be undertaken without reference to the botanical names of the taxa being constructed. Each taxon defined by this method is given a working name, and the current botanical name can be obtained at any time by examining the distribution and priority of types included in the circumscription. The working name merely provides a means of identifying the CT being worked on, whilst this may be a scientific name at this point the name has no nomenclatural or taxonomic significance. The act of publishing the circumscription creates a permanent association between it and a scientific name. To reflect this, the working name is abandoned and an ascribed name is recorded by a link between the CT storing the circumscription and the appropriate NT.

Partial circumscriptions and higher level classifications. - Certain kinds of taxonomic publications contribute to classification but do not include complete circumscriptions for all the taxonomic elements that they refer to. Floras by and large do not contribute to the global picture of taxonomic understanding because they are, in the main, intended only to be representations of a classification as applied to a particular geographical area. As a result of this geographical restriction they usually only present incomplete representations of current taxonomic thinking, and will not necessarily cite the type specimen if it is from outside the area. The same can also be said for other similar publications that sometimes cite specimens e.g. enumerations and checklists. The taxon circumscription that these publications provide is therefore usually limited to a geographic area. Nevertheless, these partial circumscriptions are represented in our model as CTs with the ascribed name linking to an NT. If the type of that NT is not explicitly included in the CT, then the type of the NT is implicitly included. Where this is the case, if the implicit type takes precedence over any other types that may have been included in the CT, then the 'calculated name' is qualified with '*sec*.' (*secundus*), the author of the circumscription. If this CT is then seen to be subsumed into another taxon then when the classification is viewed from this new perspective then '*sec*.' is changed to '*sensu*' indicating that the author of the new CT considers the name to have been misapplied. In these instances any implicit types are disregarded when obtaining the 'calculated name' of the new taxon. This treatment of misapplied names closely follows recommendation 50D in the *Code*.

Higher level classifications usually refer to exemplar taxa, e.g. molecular phylogenetic reconstructions at the family level based on sample species. Essentially these are name-based studies even though they are used in the context of a classification. Such classifications are represented using 'minimal' CTs whose circumscription lists merely contain the implicit type, and voucher specimens at the appropriate level. The presence of these 'minimal' CTs would alert other taxonomic workers that these classifications are based upon non-explicit taxon concepts.

Name-based data. - Not all taxonomic data contribute to a classification. Information on classification can only be obtained from explicit statements of circumscription such as those found in monographic treatments. There is a substantial body of taxonomic work that is effectively name-based, and therefore makes no direct contribution to the understanding or delimitation of classifications. Name-based information can be subdivided into that which is vouchered and that which is not. Unvouchered information can only be linked to an NT; with vouchered information, however, there is a link to the herbarium specimen from which the data (e.g. DNA sequences) was derived or to which the information has been ascribed. Specimens cited as vouchers for such information do not contribute to the delimitation of classifications, as there is no information that can be used to place the identified specimen in the context of other specimens in a taxon circumscription.

It is sometimes thought that determinations on herbarium specimens can be used to build taxon concepts. Indeed it would be possible to include collections annotated by a specialist, as well as those cited through publication, into a representation of their taxon concept. In some well-collected groups this would be a daunting task and moreover would draw conclusions from the data beyond that which they can support. Determinations on specimens only reflect the taxon concept of the identifier at the time that the identification was made. Determinations that post-date the publication of a classification cannot have been included in the published taxon concept, and early determinations may well have been superceded by the time it is finally published. Annotated specimens undoubtedly were used in formulating a specialist's idea of a taxon. Unless these specimens are cited in the published classification it is impossible to know whether or not they were still included in the concept at the time of publication. This reaffirms our view that the only objective representation of a taxon concept is held in the list of specimens cited when the concept was published.

Non-taxonomic works that make reference to taxonomic concepts (e.g. vegetation surveys) will sometimes cite the taxonomic reference used to make the identifications. It would be tempting to relate the information contained in the non-taxonomic work to the taxon concepts contained in the taxonomic reference used. In the context of the Prometheus model this would mean that the data would be attached to a CT. Storing such relationships in this way is sometimes referred to as 'concept mapping' (Berendsohn *Pers. Comm*, and see 'concept synonyms' in Berendsohn 1997). We feel that this would be inappropriate because there is no guarantee that the taxon concept of the identifier is the same as that of the author of the classification; it cannot be assumed that the original author would agree with the identification. Instead it is more appropriate to store the taxonomic reference used as an attribute of the identification. We have therefore explicitly excluded 'concept mapping', and relegate this kind of name-based information to the status of a determination.

The taxonomic model must make sure that a clear separation is made between name-based data from that which contributes to a CT. In the Prometheus model, name-based information is linked directly to an NT, and in the case of vouchered data an additional link to the voucher specimen is provided. In this way auxiliary data is stored without regard to the taxonomic status of the name. This is in contrast to the IOPI model in which auxiliary data can only be associated with accepted 'potential taxon' names (Berendsohn 1997).

Implications for taxonomic working practices

We anticipate that using a database system based on the Prometheus model will increase the ease with which taxonomists can experiment with classifications and compare taxonomic concepts. It will then also be possible to trace the evolution of these concepts during the process of a revision. The system should make it easier for taxonomists to work in an unbiased manner as once an appropriate collection of specimens have been gathered they can be divided into taxa without reference to earlier opinions. These innovations will increase the objectivity and testability of conclusions, and help counter the criticism that taxonomy is purely subjective.

Employing such a database system in the production of a revision, facilitates the full documentation of the materials used for that revision, and provides a permanent record that can be used as an adjunct to a published revision. We would not necessarily advocate publishing in print extensive lists of specimens in every case, although for maximum accessibility of the information this may be desirable (for examples of this see Middleton 1996, and Phipps & Muniyamma, 1980). However, any newly published taxon should as a minimum have some indication of the location where the full specimen list can be found (e.g. a website).

Implications for global taxonomic datasets and general users

There is on-going debate within the taxonomic community on the practicality of registering new taxon names. Should this community decide to implement a central register of names then some of the concepts put forward in the Prometheus model would useful when designing such a system. This is because the nomenclature and classification sides of the model are logically separated, and so the implementation of the two halves of the model could be physically separated. That is the NT side could form the basis of a global type registration scheme to which individual CT side data sets make reference. This would give the appropriate standard nomenclatural framework without having to impose a single taxonomic viewpoint.

There is growing pressure on the taxonomic community to provide a stable list of scientific names for use in legislation, conservation, etc. In contrast to the IOPI model we keep the various classifications fragmented, by doing this it is possible to provide a more reasonable mechanism for selecting a particular taxonomic view, and providing a mechanism of indicating the 'preferred' view. In the IOPI model the preferred view can be given by the arrangement of a user-defined preferred reference list. This would be global in action, i.e. applying equally to all parts of the classification hierarchy established in the IOPI model. Therefore it would be difficult for a user to decide what the appropriate order for their preferred reference list should be. Furthermore, it is difficult to envisage what knock on effects changing the order of the reference list would have on the whole dataset. Under the Prometheus model the general user would navigate through the system simply using names; they would interface with the system as if it were name based. The results from each name-based query would indicate to the user when alternative taxonomic opinions regarding a name are available. The user would select their preferred classification and the result of their query would be displayed within that context.

Conclusions

Extant taxonomic databases are not capable of dynamically handling multiple, contradictory classifications, nor do they differentiate completely between nomenclature and classification. We have addressed these problems and have developed the Prometheus model, which will allow users to switch between classifications, and to compare and contrast them in an even-handed manner.

Berendsohn (1997) pointed out that there is a frequent error of over-simplification of taxonomic data by non-taxonomist database designers. This has lead to the development of databases which inaccurately or incompletely store taxonomic data. Models that accurately handle taxonomic data are thus inherently complex and rather difficult to readily assimilate. It is important to note that in the production of our taxonomic data model we returned to first principles and looked at the processes that taxonomists use. The resultant model shows close similarities with the IOPI model (Berendsohn 1997), even though it was built up from different perspectives. The fact that they appear similar is encouraging support for the accuracy of both. Prometheus differs in the following significant areas:

- Nomenclature is clearly separated from classification.
- Circumscriptions are represented by specimens or subordinate taxa.
- Specimens form the main link between names and taxa.
- There is no complete tree-structured hierarchy of names nor taxa.
- Concept-mapping is explicitly excluded.

In the Prometheus model we recognise the true relationship between a name and a taxon, and only deal with the actual data given at source. This follows Berendsohn's (1995) profound statement that "as long as we do not insist on a taxon concept based on the intuition of a talented taxonomist (thereby effectively disqualifying taxonomy as a natural science) the data necessary for such a [database] system could be obtained routinely". In that paper Berendsohn went on to dismiss a specimen-based approach as unworkable due to the sheer scale of data entry that would be required. We think that the we have developed a workable system incorporating these ideas as a solution to handling alternative classifications. Indeed, for the Prometheus model to work no more data than that required to produce a traditional taxonomic treatment must be collected. A World inventory of all herbaria is not necessary, nor desirable as only a small fraction of specimens are cited in classifications. That is, the Prometheus system would not include more information than a working taxonomist would consider, but it can analyse it more accurately!

The Prometheus taxonomic data model is the first stage in the production of a working database system for the handling of multiple classifications. Current research is directed at finding the most appropriate database model in which to implement the taxonomic data model, determining the appropriate user interface and refining the rule sets required to maintain data integrity. These will form the subjects of future publications. A prototype system incorporating the fundamental aspects of the model has been completed. This is being refined to include the complex rule sets required by using existing data on European *Apium* (Apiaceae), and performing a revision of *Globba* (Zingiberaceae) using the prototype software. The latest information on the project and downloadable products are available via the Prometheus website [www.dcs.napier.ac.uk/~prometheus]).

Acknowledgments

We wish to acknowledge the valuable input of Peter J Barclay and Mark F. Newman in the discussions on the formulation of the model, and Walter G. Berendsohn, David G. Long, John McNeill and Richard J. Pankhurst for their constructive comments on the manuscript. The Prometheus project is funded by BBSRC/ESPRC grants 95/BIF10516 & 754/BIF10518.

*Literature cited (including *electronic publications)*

Allkin, R. 1988. Taxonomically intelligent database programs. Pp 315-331. *in*: Hawksworth, D.L. (ed.), *Prospects in systematics*. [Syst. Assoc. Vol. 36] Oxford.

- & Winfield, P.J. 1989. *ALICE user manual*, *version 2*. Kew.

*Anonymous, 1993. An information model for biological collections. Report of the biological collections data standards workshop August 18-24, 1992. Association of Systematic Collections committee on computerization and networking. Draft version March 1993 [gopher://muse.bio.cornell.edu:70/11/standards/asc, see also http://www.ascoll.org]

*-, 1994. *Logical data model for museum collection transactions management*, version 1.0 20 Jun 1994. A project of the collections and research information system (CRIS) development program. National Museum of Natural History, Smithsonian Institution, Washington. [gopher://nmnhgoph.si.edu: 70/11/.compute/.cris/.logicaldm].

-, 1995. *Interagency taxonomy information system (ITIS), requirements and design statement*, version February 28, 1995. Washington.

*-, 1998. FLORIN Information System: an integrated system to handle botanical data. Moscow. [http://www.florin.ru/florin/gen/first.htm].

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

- 1997. A taxonomic information model for botanical databases: the IOPI Model. *Taxon* 46: 283-309.

*-, Anagnostopoulos, A., Hagedorn, G., Jakupovic, J., Nimis, P.L., Valdés, B. 1996 [Jul 5]: CDEFD information model for biological collections. [http://www.bgbm.fu-berlin.de/CDEFD/CollectionModel/cdefd.htm]

-, Anagnostopoulos, A., Hagedorn, G., Jakupovic, J., Nimis, P.L., Valdés, B., Güntsch, A., Pankhurst, R.J. & White, R.J. 1999. A comprehensive reference model for biological collections and surveys. *Taxon* 48: 511-562.

Crosby, M.R. & Magill, R.E. 1988. *TROPICOS: A botanical database system at the Missouri Botanical Garden*. St Louis.

Dallwitz, M.J., Paine, T.A. & Zurcher, E.J. 1993. DELTA user's guide. A general system for processing taxonomic descriptions, 4th edition. CSIRO, Australia. [also available at http://biodiversity.uno.edu/delta/].

Duncan, T., Rosatti, T., Morefield, J.D., Beach, J., Jacobsen, S.D., Ballew, R. & Kelley, D. 1995. A relational data model for Botanical Collections. Association of California Herbaria, Inc. Working paper number 2, draft design document version: November 30, 1995. California.

Filer, D.L. 1994. *BRAHMS - Botanical Research and Herbarium Management System. A pocket introduction and demonstration guide*. Oxford Forestry Institute.

Gibbs Russell, G.E. & Arnold, T.H. 1989. Fifteen years with the computer: Assessment of the "PRECIS" taxonomic system. *Taxon* 38: 178-195.

Greuter, W., Barrie, F.R., Burdet, H.M., Chaloner, W.G., Demoulin, V., Hawksworth, D.L., Jorgensen, P.M., Nicholson, D.H., Silva, P.C., Trehane, P & McNeill, J. 1994. International Code of Botanical Nomenclature (Tokyo Code) adopted by the Fifteenth International Botanical Congress, Yokohama, August-September 1993. *Regnum Veg*. 131.

Humphries, J.M., Biolsi, D. & Beck, R. 1990. *MUSE tutorial and reference manual*. Ithaca, NY.

Middleton, D.J. 1996. A revision of *Annodendron* A.DC. (Apocynaceae). *Blumea*, 41: 37-68.

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

- 1993. Taxonomic databases: the PANDORA system. Pp 229-240 *in*: Fortuner, R. (ed.), *Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision*. Baltimore.

- & Pullan, M.R. 1996. DELTA in PANDORA. Pp 16-20 *in:* Macfarlane, T.D., Chapman, A.R. & Launder, N.S. (eds), *DELTA Newsletter, 12 April 1996*. Como, Western Australia. [also available at http://www.calm.wa.gov.au/science/delta/news/dn12.pdf]

Phipps, J.B. & Muniyamma, M. 1980. A taxonomic revision of *Crataegus* (Rosaceae) in Ontario. *Canadian Journal of Botany*, 58: 1621-1699.

Rumbaugh, J., Jacobson, I. & Brooch, G. 1998. *Unified modelling language reference manual*. Addison Wesley Longman, Inc.

Saarenmaa, H., Leppäjävi, S., Perttunen, J. & Saarikko, J. 1995. Object-oriented taxonomic biodiversity databases on the World Wide Web *in:* Kempf, A. & Saarenmaa, H. (eds), Internet Applications and Electronic Information Resources in Forestry and Environmental Sciences. Workshop at the European Forest Institute, Joensuu, Finland, August 1-5, 1995. *EFI Proceedings 3*.

Sinnot, Q. 1993. *Germplasm resources information network (GRIN) taxonomic prototype, Grin 3 schema draft*. U.S. Department of Agriculture, Beltsville.

Skov, F. 1989. Hypertaxonomy - a new computer tool for revisional work. *Taxon* 38: 582-590.

Walter, K.S. & O'Neal, M.J. 1993. BG-BASE: Software for botanical gardens and arboreta. *The Public Garden, October 1993*: 21-22, 34.

Watson, M.F. 1997. On revising a genus. *Plant Talk* 10: 31-34.

Zhong, Y, Jung, S., Pramanik, S. & Beaman, J.H. 1996. Data model and comparison query methods for interacting classifications in taxonomic databases. *Taxon 45*: 223-241.