

A taxonomic information model for botanical databases: the IOPI Model

Walter G. Berendsohn¹

Summary

Berendsohn, W. G.: A taxonomic information model for botanical databases: the IOPI Model. – Taxon 46: 283-309. 1997. – ISSN 0040-0262.

A comprehensive information model for the recording of taxonomic data from literature and other sources is presented, which was devised for the Global Plant Checklist database project of the International Organisation of Plant Information (IOPI). The model is based on an approach using hierarchical decomposition of data areas into atomic data elements and – in parallel – abstraction into an entity relationship model. It encompasses taxa of all ranks, nothotaxa and hybrid formulae, "unnamed taxa", cultivars, full synonymy, misapplied names, basionyms, nomenclatural data, and differing taxonomic concepts (potential taxa) as well as alternative taxonomies to any extent desired. The model was developed together with related models using a CASE (Computer Aided Software Engineering) tool. It can help designers of biological information systems to avoid the widely made error of over-simplification of taxonomic data and the resulting loss in data accuracy and quality.

Introduction

Data models. – A model is "a representation of something" (Homby, 1974). In the technical sense, a model is the medium to record the structure of an object in a more or less abstract way, following pre-defined and documented rules. The objective of applying modelling techniques is either to describe and document the structure of an existing object, or to prescribe the structure of one to be created. In both cases, the model can be used to test (physically, or, in most cases, intellectually) the function of the object and to document it, for example for future maintenance. Testing is usually done with the purpose of further refining the model, either to perform like the existing object, or to perform according to the functionality desired for the new object.

In the realm of computer science and the creation of computer programs, the descriptive process may be called "system analysis" while modelling the program itself is "system design" (Anonymous, 1990-1995). In reality, both processes usually go hand in hand: an analysis of existing data and (computerized or non-computerized) functions is done with the aim of creating an information system that handles the data and supports the functions. The modelling techniques provided by information science may roughly be subdivided into three types:

- "Function (or process) modelling" techniques analyse the tasks of the system and the flow of information within it, with the aim of separating functions (functional decomposition) and grouping them according to their mutual coherence.
- "Information modelling" focuses on the data used within the system to build a data model, usually in the form of an entity relationship diagram or a hierarchical data structure diagram.

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

- "Object-oriented modelling" tries to unify the first two approaches, taking into account the increased importance of user interfaces and the higher 'data orientation' of modern systems.

Function modelling is a comparatively rapid method to obtain the specification framework for a computer system. However, information systems mainly based on function modelling techniques tend to be rather rigid when additional functions are to be added later on. Information modelling needs a strong, in-depth understanding of the system and its environment and, by consequence, it is laborious and time-demanding, especially if the people conducting it do not have an intimate knowledge of the field in which the system is to be used (Coad & Yourdon, 1991). Object-oriented modelling is a rather new technique, which is why the modelling procedures provided by information science are not yet fully mature, and neither are the tools (especially database management systems) used for system design and implementation.

The information model presented here has been developed over the past six years and is based on more than a decade of experience gathered in designing and handling botanical databases. It can be used as a base for the application of object-oriented techniques, and parts of it have actually served as the basis of function modelling and system implementation. It uncovers the structural complexity of the paradigms underlying the classification and naming of plants. Although parts of the model can be used for didactic purposes, its main aim is to uncover the data structures, not to classify the underlying concepts of taxonomy and nomenclature. The model will help designers of information systems to avoid the widely made error of over-simplification of taxonomic data and the resulting loss in data accuracy and quality.

Scope of the IOPI Model. – The present information model has been designed for the International Organisation for Plant Information (IOPI) with the following aims:

- To provide for all data items identified by the 'Data Definition Group' of the IOPI Checklist Committee (see Bisby, 1994a).
- To facilitate the inclusion of additional data items pre-viewed for an extension of the checklist into the database used for the Species Plantarum Project (Anonymous, 1997).
- To store information imported from existing datasets for later revision.
- To provide an effective 'black-box' taxon/name object for non-taxonomic information systems (e.g. in molecular biology, biochemistry, biodiversity, ecology, etc., and in collection management; see Berendsohn & al., 1997b).
- To allow taxonomists working on the revision of a group to store the information gathered in the process and make it available to others. With the part of the model published here, this refers only to information on taxa or names, and the judgments taken with respect to such information.

The model includes a great number of data items which, at first glance, are not related to the checklist project. They are necessary to ensure future extensibility of the system (and there is no obligation to use all possible functions in a program based on this structure). Other aspects of the IOPI Model have been excluded from this article, particularly the treatment of geographic distribution and of literature references. The central object of botanical information, the naming and classification of taxa, is covered in full.

Data models for botanical databases. – Several attempts were made to develop standards for botanical databases, at least to allow for a data exchange between different systems. An early example is the International Transfer Format (ITF; Anonymous, 1987) and the Australian Herbarium Information Standards and Protocols for Interchange of Data (HISPID; Croft, 1989; Conn, 1996). However, lists of fields or simple data dictionaries have proven to be unable to cope with the intricacies inherent to botanical nomenclature, taxonomy, and collection information management.

Data models for botanical collections or taxonomic databases have been developed at various places since 1992 (e.g. Anonymous, 1992; Bolton & al., 1992; Sinnott, 1993; P. D. Wilson, 1993; Anonymous, 1994; 1995; Lindberg & al., 1996; and Blum, 1996). All represent attempts to bring order into the complex data structures which are involved when plants are named, collected, classified, and investigated as to their properties. Doubtless, many more such unpublished documents exist, and even more systems have been developed without any attempt to publicize the underlying model (be it because the information is considered proprietary or simply because no such model exists outside of the actual implementation).

History of the model – During the 1992 meeting of IOPI and TDWG (Taxonomic Databases Working Group) in Xalapa, the usefulness of models was demonstrated in talks presented by C. McMahon and by Berendsohn (1993). The Information System Committee of IOPI agreed to work out a detailed model for checklist data. The author provided a data model developed for the Botanical Garden and Botanical Museum Berlin-Dahlem using a CASE system (Anonymous, 1990-1995) part of which was adapted to the Data Definition Subgroup's provisions (IOPI model draft version 1 and 2). The document has since then undergone various changes and some of the resulting drafts (version 3, 4, 5.2, 6.1 and 7.3) have been made available on the Internet; version 6.0 has been distributed as part of IOPI's Global Plant Checklist project plan (Wilson, 1994). It soon became clear that the task at hand – providing a world checklist of taxa – would require massive input from large parts of the taxonomic community. A minimal model incorporating only the data prescribed by the Data Definitions Subgroup (Bisby, 1994a) would lead to a great amount of unjustifiable data loss. Consequently, during subsequent drafts the complexity of the model increased in order to preserve the taxonomic information provided by data sources. From a simple hierarchical model (higher taxon, family, genus, species, optionally infraspecies), a multi-taxonomy, multi-rank model was developed. Here, the revised core model will be offered to a wider community of readers working in the various domains of plant taxonomy.

Further IOPI meetings which deserve special mention include: Data Definition Subgroup meetings in Geneva (June 1993), and Berlin (July 1993), Information Committee meetings in Berlin (Feb. 1993, Jan. 1994), Geneva (June 1993), and Washington (Oct. 1993). Since then, the model has remained essentially stable. Some discrepancies have been solved and additions have been included as a result of prototyping efforts and discussions within the CDEFD ("A common datastructure for European floristic databases") project group.

Modelling methods

CASE techniques in data modelling and system development. – The present document makes extensive use of hierarchical data structure diagrams, which greatly facilitate discussion of the model with ‘non-technical’ participants. To depict the results of the data analysis in an abstract form, entity relationship diagrams are used which represent a logical model of the data and their interrelations. The latter can be transformed into an implementable design model in the form of relational model diagrams. These types of graphical abstractions of the analysis and system planning represent only three facets of the possibilities of a modern CASE system. Function modelling, information flows, module structures, and dialogue design are other analysis and design tools offered, which can be based directly on the results of the data analysis here presented.

The complete model held in the CASE system represents the unified efforts of various projects, particularly IOPI and CDEFD (Berendsohn & al., 1997a, b, c). All diagrams, definitions, etc. are held in a common relational data repository by the CASE system, ensuring optimal congruence between the different functional and information-related areas of the different projects, which are directly influencing each other. In this way a general view of botanical data (user and research data) is forming, which will greatly benefit specialized systems based on it.

Data Structure Diagrams (DSDs). – A DSD (Anonymous, 1900-1995) is a hierarchical tree diagram depicting "may consist of" relationships between "data items" if read from top to bottom following the connecting lines (Fig. 1). Data items may be data fields, attribute fields, parts, or entity types. "Data fields" are the basic building blocks of the model. If not further subdivided, they correspond to attributes in a relation, i.e. field definitions for a database table (marked by "Attr" on the right hand top of the box). A little arrow pointing to the bottom of a shaded data box indicates

Fig. 1. Data structure diagram (DSD).

that it represents a separate DSD. A little arrow pointing to the bottom of an unshaded data box indicates that the box represents a 'foreign key', i.e. a link to the entity type specified in parentheses in the box. The "may consist of" relationship may be modified by conditions of a kind that is specified by an abbreviation on the right hand side on top of the respective box. Such kinds of conditions are:

- "If": data items below this box are to be read only if the condition is satisfied.
- "Excl" (= exclusive alternative): same as "If", but several such conditions exist which are mutually exclusive.
- "Loop": data item is repeated as many times as indicated in the condition.

Because DSDs depict the data with a lesser degree of abstraction than other types of diagrams, they have proven to be a very efficient modelling tool for interdisciplinary communication in the analysis of scientific information.

Entity Relationship Model Diagrams (ERDs). – The ERD represents a higher level of abstraction. An "entity type" (capitalized label) can be thought of as a class of objects which may be described in the form of a table. The column headers of the table are the attributes, every row in the table represents an "entity". Two entity types may be connected by means of defined relationships between attributes (keys) which are present in both. In ERDs, the "relationships" are read along the connecting lines, starting with the entity-type name, followed by the descriptive text nearest to it, then the cardinality (i.e., how many entities of the second entity type are referred to an entity of the first type) and, finally, the name of the second entity type. The cardinality may be "1" (exactly 1), "C" (0 or 1), "N" (1 to many), or "CN" (0 to many). The "C" stands for conditional relationship, i.e. it is possible that no entity is referred to (Fig. 2). The definition of cardinality as here employed also depicts "referential integrity rules", i.e. statements to the effect of guaranteeing that a foreign key always corresponds to a primary key. Referential integrity is usually enforced by the database management system itself, after relationships have been defined. For example, in Fig. 2, the cardinality "1" means that it is impossible to delete an entity of the type Name Rank while there is still an entity of the type Taxon Name related to it. In contrast, "data integrity rules" are semantic rules for the creation, deletion, or modification of records. These usually have to be enforced by programme code written for the specific application. Data integrity rules are given at the end of each item of the model description.

ERDs and DSDs are glued together by the data elements, attributes and entity types which are referred to by both. A detailed list (data dictionary) of the entity types with their attributes and data elements used in the context of this article is available on the Internet [<http://www.bgbm.fu-berlin.de/IOPI/ChecklistModel/Entities.htm>].

Fig. 2. Entity relationship (ER) diagram: relationship and cardinality. – A taxon name has exactly 1 corresponding name rank. Every name rank is assigned to 0 to many taxon names.

1. *Potential taxa*

Linking information to taxon names. – Linking non-taxonomic data to taxa is an implicit idea of any botanical database. However, the linked object, the taxon, provides for an astonishing number of innate problems. The commonly used identifier for a taxon is its name, but the same name may be applied to different, non-congruent concepts of a taxon. The difference may range from outright misapplication of a name to slight disagreement over the circumscription of a taxon (i.e., its boundaries against other taxa). This is not accidental, but rather a direct consequence of the rules of botanical nomenclature (see Art. 47.1. of the *Code*; Greuter & al., 1994).

As herbarium specimens and field observations are the basis for grouping plants into taxa, the ‘cleanest’ way to link data to taxa would be to link them via designated voucher specimens. This makes the data independent of the taxon concept of a particular taxonomist. However, although providers of data should actively be encouraged to collect and deposit vouchers, building the information system on this approach is impractical, for the following reasons:

- A substantial percentage of taxa exists which are well circumscribed, easily identified, and not beset with nomenclatural problems.
- A vast amount of information exists which is linked to taxon names only. Such unvouchered data have to be accommodated by the system.
- There is a consensus view, in IOPI, that a taxonomic and nomenclatural checklist is a top priority, which, when fully funded, will need several years to be completed; a completely specimen-based approach would not be achievable within a reasonable time frame.

As outlined above, the circumscription of a taxon may vary from one author to another, the system the authors adhere to may be different, and a huge amount of extant information about plants is linked to nomenclaturally flawed names, synonyms, or misapplied names. The data model must allow the preservation of these different concepts, and document the errors. In a system which is to be built, at least initially, from data imported and converted from existing databases, it is particularly important that the editing process should work on data in the database rather than requiring evaluation of the information at the time of data input. Thus, simply merging the

information linked to a taxon name does not work. Several name items have to be allowed for, in which the name may coincide but be qualified by a reference citation identifying the concept of the taxon implied. These "Potential Taxa" (Berendsohn, 1995) form the central linkage point for all information about plants in the IOPI Model, and in related botanical information models that have developed over the last years (see Berendsohn & al., 1997a, for a general framework for biological information models).

To build a framework for the model, Fig. 3 depicts the principal taxon-related data areas which have to be considered:

- Potential taxa are built by combining a name (Section 2) and a taxonomic literature citation with a referenced status assignation (Section 3), which links the name to a reference according to which the name is accepted (the "taxon circumscription reference"), and according to which it is classified.
- Other taxonomic information includes all specific taxonomic and nomenclatural judgements passed on the potential taxon by the authors of its circumscription reference. This includes data on the status (e.g. accepted name or synonym, see Section 3), nomenclatural status (including conserved or rejected status), as well as systematic relationships, e.g. the inclusion of a species in a specific family. Because model links the systematic relationships to a potential taxon, it allows for an unlimited number of alternative taxonomic systems (see Section 4).
- Information can be linked to a potential taxon yet have a source reference different from the circumscription reference of the potential taxon. Linked information may include, for example, specimen determinations, herbarium management data, phytochemical items, uses, morphological features, chromosome numbers (Berendsohn & al., 1997c), etc. Such data are treated in some detail by the CDEFD information model (Berendsohn & al., 1997a, b). The only data area of this type included among the IOPI checklist data definitions is the geographical distribution of the potential taxon, which has been modelled in Berendsohn (1994).

Fig. 4. Potential taxa, taxonomic status and classification.

- Auxiliary data, consisting primarily of person-related data (see Section 6) and literature references (Section 7).

Arguably, certain nomenclatural data such as place of valid publication, basionym and type information should be linked to the taxon name rather than to a potential taxon name. However, although only one view will ultimately prove correct, the available information may be in error. The database is to document different opinions (as well as errors) as part of the investigative process. If such erroneous information were linked to a name as such, it would be overwritten in the correction process.

In the model, every name with an acceptable structure (see Section 2) represents a potential taxon, even if the only circumscription reference is the author or presumed author of the name itself.

Entity relationship model of potential taxa. – Fig. 4 depicts the entity types directly involved in the definition of potential taxa. The detailed structure of names is described in Section 2, the different possibilities for status assignments are given in Section 3, and the systematic relationships in the sense of classifications, in Section 4. The relationships of the entity types as depicted in Fig. 4 define the relational integrity rules. For example, for every name there must exist 1 or more ("N") potential taxon names. A potential taxon name always includes exactly "1" name.

2. Taxon names (*scientific names and other designations*)

Data elements in taxon names (Fig. 5). – A taxon name is here defined as consisting of either

- one, two, or three name elements, not counting rank designators (monomials, binomials, and trinomials, respectively), plus the author citation for the name (scientific names); or of
- two taxon names (formulae for hybrids and graft chimaeras); or of
- a scientific name plus the designation of a cultivar and/or a cultivar group; or of
- a scientific name plus a word, symbol or phrase designating an "unnamed taxon".

All taxon names need a designation of their taxonomic rank, which may or may not be explicitly cited as part of the name (as is the case of infraspecific scientific names, and names in ranks between species and genus).

All names fulfilling these pre-requisites may be entered in the database, irrespective of taxonomic or nomenclatural status. Relationships and attributes which express taxonomic or nomenclatural judgement are not included in nor directly related to the entity type Taxon Name, but may be expressed in relationship to a potential taxon name, which allows citation and retention of the source of information. For example, if an author assigns an incorrect basionym to a name, the error is referred to the potential taxon created by that author's action. This is to correctly model the real-world situation, that although ultimately a single correct statement may exist, this will often be at variance with what specific authors write.

Name rank. – Every name has a rank, which cannot be changed without creating another name and a potential taxon name to which it refers. Hybrid formulae, here included under Taxon Names, receive the rank of the taxon they designate, i.e. the lower of the name ranks of their parents.

Name structure. – The database program has to create taxon names by a process of concatenation of data items from various tables. The actual process varies depending on the structural type of the name; retrieval of a genus name, for example, is functionally different from retrieval of a hybrid formula or of an infraspecific name. To facilitate this process, the entity type Taxon Name includes the specific attribute "Name Structure", although the structure might also be computed from other attributes (e.g. an entity of the type Taxon Name with generic rank and with the name element filled in must be a monomial, a name with the hybrid flag set and an empty name element attribute must be a hybrid formula, etc.).

Binomials and trinomials. – Every entity of the type Taxon Name includes exactly one name element, so that binomials and trinomials require a relationship to the obligatory higher ranking name part (i.e. a relationship to a species name for an infraspecific epithet, and a pointer to a genus name for elements belonging to a binomial). Although this hierarchy represents an element of classification, here it is kept completely separate from the classification in the systematic sense. The latter is based on potential taxa while the former represents a "grammatical" rule which must be solved within the entity type Taxon Name. However, integrity rules must ensure that the two hierarchies do not conflict.

Fig. 5. Data structure of taxon names as used in botany. – Nothotaxa may include a "normal" name structure in addition to the hybrid-specific attributes, therefore the "hybrid formula or nothotaxon" condition is non-exclusive.

Recursive relationships such as the one depicted for name elements in bi- and trinomials are somewhat difficult to implement, because they carry a performance penalty in current relational database systems. Consequently, most database designers will prefer to "flatten out" the structure by including three attributes in the entity type Taxon Name, namely monomial, infrageneric, and infraspecific epithet.

(Micro)species, Aggregates, "sensu lato" taxa. – Aggregates and "sensu lato" taxa have been created by taxonomists to avoid the mandatory need for clear decisions on taxonomic identity at the basic level, or sometimes to express doubt or disagreement as to the segregates' (microspecies, included species) taxonomic distinctness at that level. The botanical *Code* does not provide for the formal recognition or handling of these terms, and no uniform policy exist in this respect within the taxonomic community.

Since aggregates and their likes have been and still are used in taxonomic literature, the data model must provide for them. The solution here followed is to include a rank "Aggregate" immediately above species level and use the classification mechanism to provide the necessary links (see Section 4). A possible alternative would be a separate entity type linking aggregates and (micro)species, but apart from a slight advantage in integrity enforcement, no immediate benefit is derived from that solution.

Cultivars. – Names of cultivars and cultivar groups consist of a scientific name (as defined above, but without author citation) and a cultivar epithet of one or more words. Cultivar names may include the two- to four-word epithet of a cultivar group, which is inserted in round or square brackets next to the cultivar epithet. As for other taxon names, all cultivar and cultivar group names, irrespective of their conformity with the conventions recently laid down in the *ICNCP* (Trehane & al., 1995), are to be stored as published. Any variation of cultivar names must be dealt with by creating potential taxa. To include a detailed treatment of cultivated plants according to the *ICNCP*, the present model will have to be extended by the introduction of several new entity types (e.g. "Registration Authorities", "Denomination Class", "Trade Designations", and "Selection or Maintenance"), but the present entity types will remain unchanged.

Hybrids and hybrid formulae. – The data model closely follows the rules and recommendations given in Appendix I of the *Code* (Greuter & al., 1994). Hybridity is indicated by an attribute in the Taxon Name entity type (attribute "Hybrid Qualifier", value: multiplication sign). This attribute may also be used to indicate graft chimaeras (addition sign) and somatic hybrids, which are supposed to follow the same structural rules as hybrids. Hybrid formulae are expressed by a relationship to the entity type Hybrid Formula which in turn points to two taxon names (those of the parents). The parents themselves may be designated by hybrid formulae, so that multiple parentage can be expressed.

To maintain data integrity, the following rules have to be enforced:

- Hybrid formulae are entities of the Taxon Name entity type, with the attribute "Name Element" remaining empty. In parallel, an entity of the Hybrid Formula entity type must be defined, which provides links to both parents.
- In contrast, nothotaxa (named hybrids) have the "Name Element" attribute filled in. A corresponding entity of the Hybrid Formula entity type may exist, which

points to one or two parents. (Art. H 3.2 of the *Code* prescribes that at least one parent be known or postulated to designate a nothotaxon.)

- For output, nothotaxa of genus and species rank as defined in the *Code* must be indicated by inserting the multiplication sign before the generic name or the epithet, respectively. Other nothotaxa are indicated by prefixing the term "notho" or "n" to the term denoting the rank of the taxon (Art. H.3.1)

Attributes to indicate the sex of the respective parent are included in the entity type Hybrid Formula (Rec. H.2A).

Unnamed taxa. – These are taxa recognized by a taxonomist which have not yet been named but have been referred to otherwise (for example, by a herbarium specimen citation and/or a location description). It may seem frivolous to introduce a naming convention for unnamed taxa. However, as the Australian members of the IOPI's Information Systems Committee pointed out, up to a third of the plant taxa present in Australia may belong to this category. It could be argued that these taxa do not possess a name structure at all. However, the point in separating taxon names from potential taxon names is that several taxonomic and/or nomenclatural opinions may exist with respect to any taxon name, including designations of unnamed taxa. For example, a recently observed population of trees in Central America might be stored as [*Crateva* "Population in the El Imposible National Park, El Salvador, C.A. with (a character combination of) blue corolla, petioles > 2 cm, and fruits > 5 cm in length; *Berendsohn & al.* 551, 743 (B, LAGU, MO, WIS)" sec. Berendsohn, pers. comm.] When, upon examining the specimens, a specialist in *Capparaceae* (Iltis, pers. comm.) comes to the conclusion that the material represents yet another variant

of *Crateva tapia* L., he will create a synonym record and assign *Crateva tapia* L. sec. Iltis as the accepted name.

As a side effect, the unnamed taxon construct allows for the inclusion of cladistic nodes in the model. In conjunction with the reference provided by status assignment, it also gives room for the entry of potential taxa with "Candidatus" status in prokaryotes (Murray & Schleifer, 1994; Murray & Stackebrandt, 1995).

Taxonomic reference citations. – Monomials, binomials, and trinomials have an associated author citation. In the case of taxa which have been altered in rank, or transferred to a new position, authorship of the basionym is to be added parenthetically (*Code*, Art. 49). In addition, "ex authors" (the author(s) to whom a name was ascribed, their name(s) preceding the particle "ex") as well as "emend." citations may occur.

In the model, the true character of the author citation as a bibliographic reference is directly mirrored (Fig. 6 & 13). All author citations are treated as links to a reference detail (exact page citation) which points to a reference title, which in turn has one or more authors. The only exception are the "ex authors", which are not authors

in a bibliographical sense. If the exact citation is unknown, a dummy reference in both entity types, Reference Detail and Reference Title, may be used to provide the link to the author team. Management of dummy references has to be assumed by the application program.

Details on references, and of the respective ER model, are provided in the treatment of bibliographic references in the IOPI Global Plant Checklist Project Plan (Berendsohn, 1994).

Entity relationship model of taxon names. – Fig. 7 summarizes the entity types used to express the data structures detailed in Fig. 5. Data integrity rules for taxon names include:

- A taxon name with a rank below genus and above subspecies must point to a genus name as obligatory name part.
- A taxon name with a rank of subspecies and below, cultivars excepted, must point to a species name as obligatory name part.
- A taxon name designating a graft chimaera cannot be part of a scientific name or hybrid formula.
- Aggregates may not exist on their own: for every aggregate at least 2 (micro)species have to be defined.
- Aggregate-like names ("s.l.", "group", etc.) are to be entered as unnamed taxa of aggregate "rank".
- A monomial must have a rank of genus or above.
- A binomial must have a rank of species to subgenus (or be a cultivar).
- A trinomial must have a rank below species.

3. Referenced status assignments: from name to potential taxon and synonym

Status assignments. – The entity type Referenced Status Assignment links names of potential taxa with their status reference (Fig. 4; data structure: Fig. 8). The title of the reference should suffice, but if deemed necessary, the model provides the entity type Reference Detail to accommodate page or figure numbers etc. within a reference title. The status assignment combines two functions: the reference may on the one hand create a potential taxon name, i.e. assign accepted status to a taxon name. On the other hand, it may assign a status implying non-acceptance to an existing potential taxon name, link existing potential taxa and their data to another potential taxon, or express a specific nomenclatural relationship between potential taxon names.

Assigned status. – This entity type cites, describes, and classifies the decisions taken on the status of a potential taxon name. It has three attributes: "Status Class", "Status Detail", and "Status Citation".

The attribute "Status Detail" holds the actual status information, e.g. "misapplied name". The attribute "Status Citation" holds a phrase or symbol needed for output which expresses the relationship assigned, e.g. "misapplied for" or "≡". Internationalization applies to this attribute.

The "Status Class" categorizes the entities of the type Assigned Status according to the basic types of relationships between Potential Taxon Name and Referenced Status Assignment. This implies that different data integrity and output rules can be

applied depending on the status class. Three basic criteria have here been used to formulate status classes:

- The number of relationships between an entity of the Referenced Status Assignment type and the entity type Potential Taxon Name. Three cases exist: (1) only the 1-to-1 relationship which assigns the status exists (e.g. in accepted names); (2) in the case of a full (in-toto) synonymy, another 1-1 relationship is added (for the accepted name), which (3) turns into a 1-n relationship in pro-parte synonyms (however, this relationship has been denormalized in this model).
- The presence or absence of a direction in cases (2) and (3) above. An example for a directed relationship is the assignation of an accepted name to a synonym. The status citation "is synonym of" is correct only in one direction. An example for undirected relationships is a list of homotypic synonyms.
- The necessity for the second potential taxon name to be accepted (according to the same reference that assigned the status to the first one). Again, this applies only to synonym-type relationships.

For the integrity of the system it is insignificant whether a potential taxon A is a synonym of B, or is a misapplied name wrongly used in the place of B. Both belong to the same status class "S" (see below). A pro-parte synonym (status class "L"), however, relates to more than one accepted name, which implies quite different handling procedures than the in-toto synonymy situation, especially for data output.

Further specifications of the status may be added by a relationship to the entity type Nomenclatural Status Detail (see below), and/or by means of a modifying flag (doubtful or tentative flag) in the entity type Referenced Status Assignment itself.

The separation of status assignments into an entity type of its own allows for future additions and further refinement. Presently, the following status classes and corresponding status details have been defined:

- "A": accepted according to the status reference (circumscription reference for a potential taxon). By means of the "Doubtful Flag", provisionally accepted names may be differentiated from fully accepted names.
- "U": unresolved. According to the status reference, the potential taxon name does not refer to an accepted taxon (doubtful names in taxonomic monographs etc.), or the status reference simply does not make clear which status is to be assigned. The two cases may be differentiated by means of a status detail.
- "S": a directed, in-toto relationship. According to the status reference the provided accepted name should be used instead of the potential taxon name to which the status is assigned. The former must be accepted in the same reference that declared the synonymy. Two principal cases may be differentiated: the underlying taxon names may be different ("true synonyms"), or the taxon name may be the same and the potential taxa differ only by their circumscription reference ("potential concept synonyms"). Status details that can be assigned to true synonyms: unspecified synonym; illegitimate or legitimate homotypic synonym (special case: basionym); heterotypic synonym; misapplied name; alternatively accepted name; correctable variant. Status details for potential concept synonyms: actually identical potential taxa (e.g. database records stemming from a common primary source record); coextensive potential taxa only differing in their classification (see below); presumed congruence because no further details are available (e.g. non-taxonomic sources citing only the taxon name); presumed congruence based on comparison of synonymy and geographic distribution given. Linked information (as defined in Section 1) related to the potential taxon name put into synonymy may be directly and unmodified transferred to the accepted potential taxon after the author assigning the status has scrutinized it. This applies particularly to the process of taxonomic coordination of the checklist (cf. Wilson, 1994). The system is instructed as to how to proceed by means of the attribute "Inherit Flag" in the entity type Referenced Status Assignment.
- "L": a directed, pro-parte relationship. According to the status reference, several accepted potential taxon names apply instead of the one the status was assigned to. The same rules and status details apply as for status class "S", except that the details cited for concept synonyms do not apply and that at least two entities of the type Referenced Status Assignment must exist to provide accepted potential taxon names.

The two following status classes serve to accommodate nomenclatural information involving relationships between potential taxon names (as opposed to the Nomenclatural Status Detail, see below). Their separation from the previous classes has been effected primarily to provide a straightforward way to hide this information which is of interest mainly to specialists. For the same reason, no separate classes for the respective multiple relationships were defined, this information can be extracted by a query.

- "D": directed relationship between synonyms. Status details for "D": illegitimate homotypic synonym; correctable variant; incorrect author citation; earlier homonym; basionym.
- "N": undirected relationship between synonyms. Status details for "N": legitimate homotypic synonym; acceptable variant.

Nomenclatural status. – The assignation of a nomenclatural status adds information which may or may not be the actual reason for the status class assignation but has no direct structural impact (e.g. "nom. nud.", "nom. rej.", "nom. cons.", etc.). The nomenclatural status has been separated as an additional entity type to facilitate the addition of new data and to allow for multiple designations. Because different nomenclatural codes apply to taxon names, and because codes evolve, the abbreviated name of the code (*Code*, *ICNCP*, *ICNB*) and ideally the edition should be cited for every statement made.

Basionyms. – Structurally, a basionym is just another synonym (a true synonym with the status detail "basionym" assigned to it). Authors may use the name of a combination correctly and cite the wrong basionym. The combination has thus to be a potential taxon and basionymy cannot be treated as a direct relationship between two taxon names.

Homonyms. – Scientific names in which all name elements excluding the author citation are identical are here considered homonyms (Bisby, 1994b). The homonym warning flag in the entity type Taxon Name serves to draw attention to the existence of homonyms which otherwise could be erroneously linked to new potential taxon names. Because of the existence of parahomonyms ("confusingly similar names" based on different types, Greuter & al., 1994) and because homonyms for a given potential taxon may be cited in various forms ("non", "vix") and to varying degrees of completeness, the homonym declaration must be explicit (i.e., not only handled by a flag but also by a separate entity type).

The declaration of a homonym (at least of a parahomonym) involves judgement, so this is a property of a potential taxon rather than a name (Fig. 9). The source of the homonym assignment is given by the circumscription reference of the potential taxon name it is assigned to. The entity type Homonym Assignment Type provides the original form of the homonym citation for the potential taxon (e.g. "non", "vix") and thus enables the database program to reassemble the original citation of the name and its homonyms.

Fig. 9. Entities involved in homonym declarations.

Synonyms. – Structurally, a synonym entity is similar to an entity designating a potential taxon, i.e. it consists of a taxon name, a status assignment and a reference according to which the synonym status was assigned. However, a synonym cannot have non-taxonomic information related to it, and it must always be connected to an accepted potential taxon name.

Rather than creating a separate entity type for synonyms, the model treats synonymy as a new referenced status assignment to an existing potential taxon, i.e. a synonym is considered to be a state of a potential taxon name rather than a separate kind of name. Due to the existence of pro-parte synonyms, i.e. synonyms which have more than one accepted name, a many-to-many relationship exists between the status assignment and the entity type Potential Taxon Name. However, this relationship was denormalized in the present model (Fig. 10), so that several status assignments are created to the same effect. The principal reason is not the simplified query mechanisms implied, but that different status details may apply to parts of the relationship (e.g., the synonym may be the basionym of one of the accepted names).

Data integrity rules. – The following rules related to the status class have to be enforced. Ideally, those dependent on a status class should be implemented as data, i.e. as a further attribute or attributes of the Assigned Status entity type.

- One reference title may only assign one status class to a potential taxon name.
- If the status class is "A" or "U", no relationship to an accepted name may be defined.
- If the status class is "S" or "L", a link to another potential taxon name has to be provided, which must be an accepted name according to a status assignment with the same circumscription reference.
- If the status class is "S", "L", or "D", and the status detail is "basionym" the combination author of the first potential taxon name must be the same as the basionym author of the second (or, expressed for taxonomists: the basionym authors of a combination and the combination authors of its basionym are the same).
- If the status class is "L", at least one other entity of the type Referenced Status Assignment must exist which assigns status to the same potential taxon name and has the same reference and status class.
- If a potential taxon name record with "misapplied" status detail is assigned, deleted, or changed to another status, the respective misapplied warning flags in the Taxon Name entity type have to be re-evaluated.
- This applies by analogy to the homonym warning flag.
- The "Inherit Flag" in the entity type Referenced Status Assignment can only be set when the status class = "S".
- If the status class is "V", the accepted name link provides a potential taxon which in turn is linked to a taxon name entity of the correct spelling.

Fig. 10. Entity relationship model of synonyms.

- Designation of certain nomenclatural status details may imply non-accepted status (e.g. "nom. inval.", "nom. nud.", or "nom. rej. ").
- If the status class is "N" or "D", a link to another potential taxon name has to be provided, which must be a synonym according to a status assignment with the same circumscription reference.

Alternatives. – The concept of a potential taxon name is rather simple. It is, in fact, a relationship between a taxon name and a source reference. Why, then, is this relationship not explicitly represented in the model, but rather "subsumed into the more complex set of relationships used to record subsequent status assignments", as a reviewer (S. Blum) posited?

The declaration of a potential taxon is a status assignment. Moreover, the circumscription of the potential taxon is not only recorded by referencing a publication. By defining synonymy and by relating to status details, the Referenced Status Assignment entity type accomodates part of the taxonomic decision process contained in the publication, which forms part of the taxon's circumscription. Treating the accepted status assignment by the same entity type as subsequent assignments is thus conceptually correct. Blum's alternative may indeed have advantages when implementing the model: as he correctly points out, it separates the ternary relationship [A – is accepted – according to reference R] from the quaternary relationship [A – is a synonym of – B – according to reference R]. However, accommodation of status details relating to the accepted status declaration will then have to be solved, and the following additional data integrity rule will be necessary:

- Every taxon name serves at least once as a potential taxon name. If a new name is entered from a reference which assigns synonym status to it without giving a reference according to which the name has been accepted, the reference to the nomenclatural protologue is used instead. The thereby created potential taxon can then be declared a synonym. In the case of unpublished names cited in synonymy, a dummy reference must be created.

4. Classification

Definition. – The term "classification" is here understood as "the placing of a plant [an organism] (or group of plants [organisms]) in groups or categories according to a particular plan or sequence ..." (Lawrence, 1951). Classification is thus defined as the storage of assignments of organisms to taxa, and of taxa to other taxa of higher rank. The IOPI Data Model is aimed at recording results of classification processes (Fig. 11), not at directly aiding in the process of classification (see Beach & al., 1993, and Zhong & al., 1996, for advanced taxonomic models to assist in that process). The model must take into account that classification, in addition to simple placement, may have a component related to the circumscription of the potential taxon itself, as in the case of a phylogenetic tree.

Alternative taxonomies. – Most taxonomic databases that presently exist follow a pragmatic data model that is based on a fixed hierarchical classification according to a definite system of higher taxa. The self-set aim of the IOPI model is to keep it open for any view (including erroneous ones) expressed by records used to build the database, although a preferred view may be defined for the released product (see

Section 5). Classifications relate to [potential] taxa, not to names. The source of the placement of any taxon in the classification is the reference defining the potential taxon, i.e. the references assigning an accepted status class to each name. As a consequence, a new (alternative) classification of a taxon creates a new potential taxon entity, and an unlimited number of alternative taxonomies must be supported by the model. A potential taxon may be classified as belonging to any potential taxon of higher rank, but to at most one at each rank (except in the cultivar/cultivar group relation, which is not a true hierarchy: see Hettterscheid & Brandenburg, 1995). This arrangement ensures that information can be completely incorporated, even though only an incomplete hierarchy may be explicit in the source. The program based on the data model may be used to restrict these possibilities, e.g. for the editing process. The use of a specific classification may be enforced (e.g. Brummitt, 1992, for the initial edition of the IOPI checklist), and the input of a complete classification hierarchy, or of specific levels (e.g. family level), may be required in the process of manual data entry.

Systematic sequence. – A specific classification system may prescribe a definite sequence of taxa. Output from the database may follow this sequence. For sorting purposes, an attribute in the form of a systematic sequence number ("phylogenetic sequence number"; Sinnott, 1993) must be included within the potential taxon entity type. The specific taxon must inherit the position of the higher taxon in which it was placed. The sequence number may therefore consist of either a running number (sorting would have to look up classification relationships from the top) or an index mirroring the classification at all pertinent levels.

Entity relationship model of the classification of potential taxa. – Classification is realised in the model by means of a recursive relationship between entities of the "Potential Taxon Name" type (Fig. 4). Every potential taxon may be classified as belonging to a single next higher taxon as reflected by the reference given for its accepted status. It then "inherits" the further classification of that next higher taxon. Data integrity rules for classifications are:

- If the next higher ranked potential taxon itself had already been classified as belonging to another, higher potential taxon, this classification is automatically extended to the first potential taxon, which thus inherits the position within the full systematic sequence.
- If a species is classified within a potential taxon of generic rank (i.e., if the potential genus name is explicitly cited in the circumscription reference of the potential species), the genus name must coincide with the first part of the potential species name. This applies by analogy to the classification of infraspecies in species and genera.
- The classification of a potential taxon cannot be changed, i.e., a new classification of a taxon, according to another source, creates a new potential taxon with that other source as its circumscription reference. Data attached to the first potential taxon may be attached to the second by setting the "Inherit Flag".

5. The "preferred taxon view"

The model allows for a multitude of homonymous potential taxon names to coexist in the database. Although this is deemed necessary for an effective storage of biological information, it is confusing for the user of that information. It must therefore be possible to define a preferred taxon view, in which one potential taxon name is accepted and all differently defined potential taxa of the same name rejected.

Potential taxa are defined by means of reference citations, which may include informal references along with literature citations (see Section 7). In the context of the IOPI Checklist, a reference citation may, for example, merely consist of the name of the taxonomic coordinator and a title like "IOPI Checklist coordinator, 1996". A hierarchy can be established by means of the entity type Taxonomic Coordination and Preference, which defines which reference (and thus, which potential taxon) takes priority (Fig. 12). Thus, a priority list of references is defined that must be followed, whenever conflicts arise, to obtain a preferred taxon view of the data. For the Checklist, this has the additional advantage of making it possible to include standard references like taxonomic treatments as coordinator references. Information affected by the Taxonomic Coordination History is detailed in Fig. 12.

6. Names of persons

Personal names are an important data area in almost any database. In the IOPI model, persons take the role of, e.g., authors of scientific names, or authors, editors, and compilers of bibliographic items, or responsables for unpublished taxonomic decisions (e.g., taxonomic coordinators of checklist entries), and administrative tasks like data entry (Fig. 13). The de-facto standard data structure for personal names is to use their "Western" subdivision and sequence, i.e., first the given name(s), or their initials, followed by a particle (if any) and the family name or names. These name elements should be atomized as far as possible to facilitate concatenation of strings adhering to specific formats, as required for example by specific journals for bibliographic reference lists.

For the IOPI model the attribute structure (Fig. 14) of personal names follows the electronic edition of the *Authors of plant names* (the printed publication of [Brummitt & Powell, 1992] is a TDWG standard). Subtypes may be defined according to the role of the person. For example, a separate entity type should be used to define the personal name abbreviations used in abbreviated citations (according to the standards in Halliday & al., 1980; Brummitt & Powell, 1992; Stafleu & Cowan, 1976-1988, and Stafleu & Mennega, 1992-1995), or to store the addresses of persons (if needed). In any case, program developers have to be observant of national legislation regulating the use and storage of person-related information in databases.

Often several persons are referred to jointly as a unit. For practical reasons, the model refers to single persons as if they were teams, or, rather, it does not recognize persons but teams of one or more persons. The data structure of such person teams is depicted in Fig. 14. For an entity relationship diagram of person teams in the IOPI and CDEFD models and an example of an actual implementation, see Elankovan & al. (1997).

7. Non-standard features of taxonomic source references

Bibliographic data are central to scientific research. A well structured, highly atomized partitioning is needed to provide for the various needs of scientists and other users. All information about plants, taxon names, status assignment, classification, etc. in the model is tied directly or indirectly to its source by means of reference citations. Library systems and programs managing scientific literature citations abound, and many of the structural elements found in the present model are also found in

these programs. However, in the context of a taxonomic data model, several special requirements have to be met:

- Inclusion of standard abbreviations used in taxonomic short citations. Fortunately, a broad consensus exists among botanists to adopt the following abbreviation standards sanctioned by the IUBS Commission for Plant Taxonomic Databases (TDWG): for authors of plant names Brummitt & Powell (1992); for names of periodicals Lawrence & al. (1968), and Bridson & al. (1991); and for book titles Stafleu & Cowan (1976-1988) and Stafleu & Mennega (1992-1995).
- Inclusion of databases as referenced sources. In contrast to printed information, standards for the citation of databases and the structuring of such citations have yet

Fig. 13. Entity relationship model of person teams and their role in references, taxon names and taxa.

to evolve. As databases change continually over time, the reference date is an important new attribute. Some commercially available databases produce 'editions', which may be distributed in durable form (e.g. as hard copy).

- Inclusion of reference to unpublished data sources. Taxonomic information may be derived from a wide variety of unpublished sources, such as herbarium sheets, unpublished theses, personal comments, etc. Some of these, such as manuscripts and theses, can be handled by means of the attributes provided for printed publications and can be treated by analogy, with the addition of an "unpublished" flag to clarify their status. Others, such as information from herbarium labels or personal comments, require a specific category, the Informal Reference entity type. Informal references should be accessible, so an attribute is included to specify the place where the reference is deposited (e.g., notes of personal comments stored in an archive).
- For nomenclatural citations, an exact page citation within a title must be possible, to refer to the page or pages on which the protologue is found.

A detailed model of reference citations is included in the IOPI data model as published in Wilson (1994) and is available on the World Wide Web (see Berendsohn, 1994).

8. *Comments*

Although throughout the model the data have been atomized into well defined data elements, comment in free text format will often be necessary to accommodate

special cases. Such comment fields introduce the risk that an ill-designed user interface may encourage to place data there rather than in the apposite fields, thus making search and sorting processes difficult and in effect hiding information from users of the database. Inclusion of comment fields for a particular data area should therefore carefully considered, and, avoided whenever possible.

Comments can be linked to any entity type in the model by means of a comment list (associative entity type). The IOPI data definitions (Bisby, 1994a) prescribe comments for taxa and for geographical data in the IOPI checklist. This is implemented in the present model by allowing comments on names, potential taxa, and global distribution summaries.

The structure defined by the data definition restricts comments to short remarks, which can (and should) be accompanied by citation of a reference source. This is a structure which has been used successfully in the International Legume Database Information System project (ILDIS; Bisby, pers. comm.). It has been included in the present model; however, since modern database systems do not restrict field size, and since the ILDIS structure may be perceived by editing taxonomists as being overly patronizing, a free text field has been added to store longer comments. Again, it is up to the actual implementation to impose restrictions.

Implementation of the model: building the checklist

Realizing a database based on a model as complex as the one presented here demands considerable effort. Design of the central storage facility and of taxonomic workbench programmes are only one side of the task, the administrative and coordination facilities to create and maintain the database have also to be considered. Despite intensive efforts, IOPI has yet to secure funding for an implementation of the global plant checklist database, even though the taxonomic community had volunteered to provide the input, i.e. funding was sought only for the technical and administrative infrastructure necessary to build and maintain the checklist. Lamentably, funding organizations appear to be reluctant to spend even comparatively modest amounts of money on truly international efforts. With the current trend of shoving ever-increasing amounts of unstructured and often unrevised information on organisms onto the international networks, the demand for high-quality data in structurally well defined databases is more urgent than it was ever before. Devising and operating an authoritative but at the same time non-discriminating information system for taxonomy thus remains a top priority for the IOPI.

Conclusions

The IOPI Model's core is formed by the entity types Name, Potential Taxon Name, Referenced Status Assignment, and Reference Title, and their relationships as depicted in Fig. 4. By means of minor changes to data integrity rules and insertion of some additional attributes, it can accommodate the zoologists' and bacteriologists' requirements, as well as the needs under the proposed BioCode (Greuter & al., 1996). This structure is able to handle a multitude of tasks, among them conceptual circumscription of taxa, acceptance, synonymy, classification in alternative taxonomies (even with incomplete trees), and systematic taxa sequence. Any of these tasks may be singled out and modelled in a seemingly more straightforward way. How-

- & Mennega, E. A. 1992-1995. Taxonomic literature, ed. 2, Suppl. 1-3. *Regnum Veg.* 125, 130, 132.
- Trehane, P., Brickell, C. D., Baum, B. R., Hetterscheid, W. L. A., Leslie, A. C. Spongberg, S. A., Vrugtman, F. & McNeill, J. 1995. International code of nomenclature for cultivated plants – 1995. *Regnum Veg.* 133.
- Wilson, K. 1994. *Global plant checklist project plan, version 1.2*. International Organization for Plant Information (IOPI), Sidney.
- *Wilson, P. D. 1993. *Missouri Botanical Garden research data model. Prepared for the Research and Steering Committee. Release 4.0*. Missouri Botanical Garden. [Gopher://cissus.mobot.org:70/11/pub].
- Zhong, Y., Jung, S., Pramanik, S. & Beaman, J. H. 1996. Data model and comparison and query methods for interacting classifications in a taxonomic database. *Taxon* 45: 223-241.