

The Prometheus II Description Model: an objective approach to representing taxonomic descriptions

Sarah M. McDonald¹, Cédric Raguenaud², Martin R. Pullan¹, Jessie B. Kennedy², Gordon Russell² & Mark F. Watson¹

¹Royal Botanic Garden, Edinburgh, EH3 5LR, U.K. Email: s.mcdonald@rbge.org.uk; m.pullan@rbge.org.uk (author for correspondence); m.watson@rbge.org.uk.

²School of Computing, Napier University, Edinburgh EH10 5DT Email: cedric@bassoe.com; j.kennedy@dcs.napier.ac.uk; g.russell@dcs.napier.ac.uk.

SUMMARY

A model for improving the storage and communication of plant descriptions is presented. The model is flexible and yet reduces the ambiguity often present in text descriptions. The fundamental idea behind the model is the necessity for clear representation of the meaning of each term used within a description. The model therefore emphasises the assignment of definitions to all taxonomic terms used. This approach will eventually provide a more objective way of expressing plant descriptions that supports their comparison without making assumptions on the semantics of the terms used.

KEYWORDS: definition, descriptions, character, databases, taxonomy.

INTRODUCTION

This paper presents an abstract data model designed to address some of the current problems with taxonomic descriptions. The ways in which taxonomic descriptions are currently written will be discussed followed by an analysis of the problems with descriptions. The Prometheus II Description model will then be outlined. Examples from existing descriptions will be used both to highlight the problems and to demonstrate how the model works. The model will be used to build a system that will eventually be an archive of descriptive data which others may use and refer to. The data held in the archive will be unambiguous because any terminology used will refer to a definition. This paper does not deal with how descriptive data will be recorded or viewed by the user in the interface of any future system, rather it deals with the data that will be recorded and how those data will be linked to form descriptions.

CURRENT PRACTICES

A taxonomic description is the main way in which a taxon concept is communicated. Character data are used within a description to record the boundaries of a taxon concept, however, there are problems with the way in which these character data are handled and used for communication. There are many instances where this communication is ineffective. For example, in a recent study by Kelly & al. (in press), diatom taxonomists were asked to identify images of diatoms. The ability to correctly identify each specimen ranged from 33.8% to 86.5%

with a mean of 63.3%. It was suggested that one cause of such low accuracy could be deficiencies in the descriptions of species in Floras. As the purpose of taxonomy is to communicate taxon concepts and their relationships in order to provide frameworks within which other disciplines can work, it is important that the way in which these taxon concepts are communicated is improved.

The data in a taxonomic description record information about an organism. ‘Character’ in this instance is often defined as ‘a statement on a feature of the organism’ (e.g., Davis & Heywood, 1963; Blackwelder, 1967; Wiley, 1981; Colless, 1985; Stuessy, 1990; Fristrup, 1992; Bailey, 1999). However, taxonomists in general are not concerned with the exact definition of the term ‘character’, but concentrate on developing character concepts, i.e., defining what they mean by ‘character’ in practice.

How character concepts are developed. – In order to define a model for representing character data, it is important to understand the process by which character concepts are developed. The construction of character concepts involves the partitioning of observed variation into characters and character states (Wilkinson, 1995). From an alpha-taxonomic perspective, character concepts are developed during the taxonomic process (Cannon, 2001). The taxonomic process involves sorting specimens into groups using a mixture of past experience and differences in the overall appearance to differentiate the specimens. Once the initial sort is completed, each group is then examined in turn to decide whether it actually contains one taxon. Each specimen within the groups is examined in more detail to determine how much variation between specimens can be considered simply within-taxon variation and how much is sufficient to warrant separating the specimens into different groups. The taxonomist is looking for comparable structures and the variation between those structures. It is at this point that recognisable character concepts start to appear, i.e., the combination of a structure, the aspect of the structure being described and its possible states. For example, the character ‘leaf shape’ may be recognised and the states ‘obovate’, ‘ovate’ and ‘lanceolate’ observed in a group. The description of that group would then read ‘leaves obovate, ovate or lanceolate’.

The use of character concepts in the description of taxa. – Character concepts are used to formulate a description, which is then tailored to fit the intended purpose. Different types of description are appropriate for monographic works (monographs, revisions, etc.) and floristic works (Floras, field guides, etc.). In all cases, the main purpose of the description is to aid in identification, but the emphasis and level of detail depend upon the type of publication and its

intended audience. New classifications are generally put forward in monographs and revisions and the descriptions tend to be comprehensive, containing (ideally) a full justification and explanation of the author's taxon concepts. Monographic publications usually have introductory chapters discussing the author's character concepts, sometimes relating these to others in the literature. For example, the monographic treatment of the genus *Sanicula* (Umbelliferae) (Shan & Constance, 1951) discusses the characters important for classification and evolution for each section within the genus at the beginning of the treatment. In contrast, floristic works tend to be more concise; they may use and comment on existing treatments, but will generally not develop new classifications. Specimens are usually consulted, but descriptions are often written from a mental summary of the variation seen rather than measuring each specimen individually. Although descriptions in floristic works may convey taxon concepts, this role is secondary to the primary purpose of differentiating between taxa found in a geographical area. Thus the descriptions are often short and only give the diagnostic characters essential for identification. For example, the description of *Sanicula* found in the Flora of India (Mukherjee & Constance, 1993) is less than half the length of the description contained in the monographic treatment of *Sanicula* (Shan & Constance, 1951).

PROBLEMS WITH DESCRIPTIONS

There are many problems with the ways in which character data are recorded in descriptions. The specific issues addressed by the Prometheus II Description model follow this general discussion of the problems.

'Character'. – As mentioned earlier taxonomists are generally not concerned with developing a wide ranging definition for the term 'character'. In fact given the number of uses for character data this would be very difficult, if not impossible. For example, the loose definition of the term 'character' as a feature of an organism is modified to be more specific by including the concept of homology when using features in cladistics (e.g., the definition of character in Kitching, & al., 1998). Colless (1985) consulted 50 publications and found nineteen different explicitly stated, or clearly implied, definitions of 'character'. These definitions included 'character' defined as an attribute, a set of attributes, a feature, a property, a differentia, an aspect (of an organism), a basis for comparison and a set of probability distributions. However, Diederich & al. (1997) concluded that there are two main ways in which the term 'character' is used in descriptions in the literature. One is where the character is a general concept, e.g. leaf shape, which is separated from the score, e.g. obovate. In this situation the character is a combination of a structure (leaf) and an abstract concept or property describing that structure (shape). The score is sometimes called the 'character state'. The other use is where

the structure and the score are combined, e.g. leaves obovate, in which case the property (shape) is implicit. There is little consensus on what the term 'character' actually means, which leads to problems with interpreting descriptions.

The more serious problem with characters in practical terms is that the selection and definition of character data is ambiguous. Davis and Heywood (1963) highlighted the problem, commenting that "Character selection is the weak link in this whole approach. In assessing similarity the taxonomist working neurally does not have to make consciously the abstractions we call characters; it is only when he wishes to communicate about any particular aspect, e.g. for diagnosis, that he is forced to rationalise what he recognises as a *Gestalt* of many independently varying elements, and break it down into component parts. As a consequence of this the taxonomist may produce a satisfactory division into species *despite* the characters set down in his descriptions." Although Davis and Heywood were referring to numerical taxonomy when they stated that character selection is the weak link, this logic can be applied to descriptions, to comparisons and to cladistics. If the characters are not chosen well then any work based upon them will be of little use. This is demonstrated in the example used by Davis and Heywood. The genus *Biscutella* (Brassicaceae) was monographed by Machatschki-Laurich (1929), but the key and descriptions provided are not sufficiently different for related taxa to allow separation. However, an examination of material cited shows the species accepted are recognisable on the basis of characters not mentioned by the author. Although it is difficult to see how one may produce tools to improve the selection of characters, it is possible to devise ways of making sure the characters chosen are unambiguous and easy to interpret, as will be described later.

Terminology. – A further problem with character data and descriptions is that there are no universal definitions for the terms used and glossaries of terms used are not often given, making objective and meaningful comparisons almost impossible. For example, Lawrence (1951) defines the term 'tomentose' as 'densely woolly or pubescent; with matted soft wool-like hairiness' whereas Stearn's (1983) definition is 'thickly and evenly covered with short more or less appressed curled or curved matted hairs'. In this case, not only is the way in which the term is described different but the meanings also differ to some degree. A term used without a definition will mean that a reader's interpretation of the term will depend on which definition(s) they have previously encountered and how accurately they have remembered them. The reader's understanding of the term may consequently be significantly different from the author's. Taxonomists tend to have a set of terms that they always use and have fixed ideas about their meaning, which may be different from other taxonomists' understanding. There are also

numerous examples of keys and descriptions presenting ambiguous terms which make correct identification and comparison difficult. For instance, a key to the species of *Sanicula* found in Taiwan (Huang, 1993) differentiates *S. lamelligera* from *S. petagnioides* on the basis of the number of spines found on the fruit. The states for this character are ‘densely spined’ and ‘sparsely spined’. The author had a sound taxon concept for each species but any user of this key may find it difficult to distinguish between ‘densely’ and ‘sparsely’. It is only possible to distinguish ‘densely’ from ‘sparsely’ by comparing previously identified specimens from a herbarium, which is frustratingly time consuming and not always possible.

Descriptions. – Taxonomists tend to summarise data obtained from all the material examined when writing a description, which leads to the following issues:

1. A considerable amount of time and effort is spent gathering data, many of which are not included in the summary account. Once the work has been published, these data are often discarded or forgotten, and not made available for reuse or verification (Diederich & al., 2000). The *Biscutella* monograph cited by Davis & Heywood (1963) is an example of this. The author will have no doubt noted the characters for separating related species but chose not record them in the publication, meaning that any reader must return to the specimens and make repeat observations.
2. When examining summary accounts, it is often impossible to distinguish, with any degree of conviction, between actual observation and extrapolation (Watson, 1971). For example, the *Sanicula* description found in the Flora of India (Mukherjee & Constance, 1993) reads ‘leaves palmately or pinnately lobed or divided to compound, rarely entire’, it is not clear whether this is an observation from one specimen with all these states or an extrapolation based on, for instance, some specimens with pinnately lobed leaves, some with palmately divided leaves and some rare specimens with entire leaves. There are numerous arrangements of these possible states, but no indication as to which combinations have actually been observed.
3. The thought processes underlying the summarization of the data are not available for evaluation.
4. It is impossible to distinguish between absence of a feature in the material being described and mere failure to seek or comment on it, which makes using descriptions for identification or constructing keys difficult.

Methodology. – Additional problems arise from the fact that there is no formalised methodology for carrying out the taxonomic process. Each taxonomist develops their own way of working and is unlikely to significantly change their working practices once they are

established. Leenhouts (1968) wrote a guide to herbarium taxonomy, but it has now been largely forgotten, despite some approval and agreement in the literature at the time. On occasion, there are collaborative projects, for example Flora Malesiana or Flora Europaea. However even in these projects, the individual taxonomist is assigned a subset of the subject area and produces descriptions in an agreed format, but does not actually collaborate on the details of the taxonomy. Working in isolation, and the absence of accepted formalised procedures, gives rise to individualistic and subjective working practices that are not conducive to the effective communication of taxonomic concepts.

Comparisons. – Character data and taxon concepts, which are based upon these terms, are not interchangeable with other character data and taxon concepts, without making rather broad assumptions about equivalence of terminology. This means it is difficult to use a description for any purpose, other than identification or comparison within the particular monograph or Flora in which the description appears. For example, the leaves of *Sanicula* are described by Mukherjee & Constance (1993) as ‘palmately or pinnately lobed or divided to compound, rarely entire’ and by Huang (1993) as ‘palmately 3–5 lobed or rarely pinnately lobed’. It is not possible to assess these characters as being equivalent or to say that the authors’ taxon concepts are the same, without making broad assumptions.

The information in a description cannot be used reliably for any of the wide range of methods that use taxonomic data, such as comparative biology and phylogenetic reconstruction. It is often not even possible to construct a simple key using descriptions written by the same author using current working practices. Anyone wishing to do a taxonomic revision or collect data for comparative biology and cladistics generally has to return to the specimens, using the descriptions as a guide rather than a source of data. This is acceptable for situations where there are only a few specimens or species, but makes working on large families (500+ species) almost impossible (Jacobs, 1969). Descriptions are also poor sources of comparative data (Watson, 1971; Anonymous, 2000a) and it may not be possible to deduce from the descriptions whether two taxa share similarities because the same criteria are not used in the construction of the descriptions (Sivarajan, 1991).

Computerisation. – Electronic descriptions have been seen as a way of making descriptions more uniform and informative. An example of this is DELTA, a data format for representing and manipulating taxonomic descriptions (Dallwitz, 1980). The main drawbacks of DELTA are that the user can use any terminology within a ‘character’ and its ‘states’ and that the user is free to place the boundary between ‘character’ and ‘character states’ arbitrarily. Furthermore it is

difficult to describe the interrelationships between ‘characters’ in any way other than to say they are dependent or to use ‘or’ between different characters. These limitations have arisen because in DELTA the focus is on maintaining consistency within data sets, without regard either for the consistent reuse of terms across data sets, or for the comparison of data sets from disparate sources.

The vast majority of record keeping is paper-based and takes the form of notes, drawings and diagrams. Newly qualified taxonomists tend to utilise computer techniques to an extent, but there is no consensus among taxonomists on which of these techniques are the most effective. Spreadsheets of varying degrees of sophistication are often used to record character observations during the taxonomic process. Some of these spreadsheet packages are specifically designed to be able to export to file formats such as DELTA (Dallwitz, 1980) and NEXUS (Maddison & al., 1997). However, the focus has been on personal data organization rather than creating an archive of taxonomic data to which others may refer.

Objectives for the Prometheus II Description Model. – Allkin (1984) suggested that descriptions could be made more useful in the following ways:

1. Any terminology is used consistently throughout all descriptions.
2. The criteria used to describe one organism are used in all other descriptions.

Other authors have also suggested that there should be a standard approach to taxonomic descriptions (Watson, 1971; Sivaranjan, 1991; Diederich & al., 1997; Anonymous, 2000b). However, so far it has not been possible to agree a standard descriptive terminology or structure for the whole of botany, let alone for all branches of taxonomy (Anonymous, 2000b). The TDWG Descriptors Group attempted to finalise a list of universal characters that should be scored for every plant description, but were not able to agree on the list’s contents. A fixed character list would deal with the problem of distinguishing between the absence of a feature in the material being described and mere failure to seek or comment on it. However, a fixed character list is extremely limiting and will inevitably lose the expressiveness of a taxonomic description. It would also be difficult to address the problem of the thought processes underlying the summarization of the data not being available for evaluation, as this would require explicitly recording what the taxonomist thinks. Consequently the model will address the following points:

- The model is designed to help clarify what the taxonomist means by each term used, by being explicit about the types of information that should be included and by defining any terminology used.
- The loss of data and the difficulty in distinguishing between extrapolation and actual

observation will be addressed by designing the model to store character data easily during the taxonomic process rather than capturing data after the description has been published. The model will also store the combinations of states by explicitly recording variation as either AND or OR and by encouraging the user to work from specimens rather than recording taxon summaries.

- The recording of definitions and the relationships between possible states will promote the reuse and comparison of the data held within taxonomic descriptions.

THE PROMETHEUS II DESCRIPTION MODEL

This section is intended to give an overview of the basic concepts of the model. Certain aspects of the model will then be dealt with in more detail later. The model is intended mainly for recording the information collected for new descriptions, but it can also be used to record an interpretation of an existing description.

In order to avoid confusion with alternative definitions of the term character, a definition from Diederich & al. (1997) will be used in the Prometheus model. Character is specifically split up into structure, property and score. This definition is flexible enough to apply to both qualitative and quantitative statements, while enabling the taxonomist to be explicit about every aspect of each statement.

Defining terms. – The Prometheus approach emphasises the definition of **terms** and requires each use of a term to be placed into a context determined by the definition. Terms are simply the bare terminology found in a standard description, for example ‘leaves’ or ‘pubescent’. At present, the model requires that these terms be defined to create **defined terms** in order to make explicit statements about features that can be easily interpreted. The way in which the definitions are formulated is the subject of future research and is not dealt with in this paper. However, it is possible to discuss the information that must be included in a defined term in order to make it explicit.

Each defined term must include a textual definition and a literature reference (including the author of the definition) to differentiate the alternative definitions of a term. For example, ‘pubescent’ has been defined as ‘pubescent = adj. of pubescence, a somewhat dense cover of short, weak, soft hairs (Hewson, 1988)’. A defined term should also include aids to interpretation such as pictures and drawings whenever possible. For example, the Hewson (1988) definition of pubescence includes a figure illustrating ‘pubescent’ on *Lepidobolus preissianus*.

The types of term are **structure terms** (e.g. leaf), **property terms** (e.g. texture), **qualitative state terms** (e.g. pubescent), **unit terms** (e.g. cm) and **modifiers** (e.g. rarely). After definitions are assigned, these become **defined structures**, **defined properties**, **defined qualitative states**, **defined units** and **defined modifiers**. Fig. 1 shows five kinds of term and five kinds of defined term that in various combinations allow the building of descriptions. Table 1 gives an explanation of the notation used within this paper. The notation is following that used within Pullan, & al. (2000) and Raguenaud, & al. (2002).

The five kinds of defined term require slightly different information to be included. A defined structure requires a definition, a reference and an image. When making statements about dimensions, such as length, ambiguity arises. It is not always clear where the measurement has been made. For example, leaf length could be measured in a number of ways. The defined property therefore must include information about where a measurement has been taken, i.e. the start and end points (e.g. the apex of the leaf and the point where the petiole joins the lamina). The set of property terms, from which defined properties may be formed, is fixed (Table 2). It contains all the possible properties used to describe structures. However, additions to the property term list are possible when specific plant groups require additional vocabulary. The addition of new properties must be performed with care. For instance, it would be a mistake to add 'smoothness' as a property when it should be described as a state under 'texture'. The catchall property 'form' is included to cover terms, such as acephalous or centripetal, which do not easily fit into more specific properties.

Defined qualitative states must include a reference to one or more property terms (Fig. 1). Associating a defined qualitative state with a property term allows queries such as 'find all the states that are shapes' to be handled. Allowing a defined state to reference multiple property terms handles state terms such as 'radiating', where it is difficult to conclusively assign the term to only one property.

Defined units are used in conjunction with values. When descriptions are created, taxonomists may use different units for their measurements (e.g. imperial or metric units) or different scales (e.g. metres, centimetres). Each unit term has only one definition as these are internationally agreed systems. The inclusion of units in a value/property definition would be too restrictive; therefore taxonomists are responsible for their choice of units when recording scores.

Modifiers are different from all other terms in that the list of modifier terms is both fixed and that each term can only have only one definition, in order to eliminate ambiguity and promote comparability.

Building descriptions. – In order to create a description, the various types of defined terms are linked together to make statements on features of an organism, i.e. to form characters. As discussed earlier there are many problems with the term ‘character’, so in order to be explicit, this model will refer to these statements on features as **description elements**. There are two kinds of description element, both of which will be discussed in more detail later: **qualitative description elements** describe a feature in terms of a defined qualitative state, and **quantitative description elements** describe a feature in terms of a value and a defined unit. When creating a description, description elements are grouped together into **description units** (see Fig. 2). Description units are everything that is said about one defined structure, for example, a leaf description unit might contain various description elements, each describing one aspect such as colour, texture and shape. By grouping description elements that describe the same defined structure, a user may quickly find all the information in a description about that defined structure. Description units are then grouped together to form a **description**, which is everything that is said about a specimen or taxon in a publication. The use of each defined term in a description is a subjective issue and in order to recognise this the user of the defined terms is noted by recording the author of each description.

MEASUREMENT VERSUS CATEGORISATION

This section deals with qualitative and quantitative description elements in more detail. In an ideal world all taxonomic data would be recorded through measurement in quantitative description elements. However, this is not always practicable and taxonomists tend to assign qualitative states by breaking up continuous variation into more easily handled discrete states. For instance, leaf shape is usually described in terms of discrete states such as linear or lanceolate, although in reality leaf shape is a continuum (Hickey, 1973). Within this model, description elements are split into quantitative and qualitative to reflect these different modes of working. These types of description element are differentiated because they require different pieces of information to be included in order to be explicit.

Qualitative Description Elements. – In order to correctly interpret a statement such as ‘leaves obovate’ all a reader needs to know is what leaves are and what obovate means. Therefore, a qualitative description element consists of a defined structure and a defined qualitative state (Fig. 3A). When selecting a defined structure for inclusion in a description

element there are three options:

- draw from a list of defined structures
- create a new defined structure using an existing structure term
- create a new defined structure and a new structure term

Therefore the list of structure terms is open-ended and user defined. Defined qualitative states are selected in the same manner, so the list of state terms is also open-ended and user defined. In a qualitative description element it is not necessary to specifically highlight the property being recorded. The fact that ‘obovate’ is a shape becomes apparent because it is recorded in the definition of the defined qualitative state. Fig. 3B shows how the qualitative description element ‘leaf obovate’ would be represented in the model. Here the defined structure ‘leaf’ is linked to the defined qualitative state ‘obovate’ via the relationship description state. The property ‘shape’ is implicit.

Quantitative Description Elements. – In order to interpret the statement ‘leaf length 5 cm’, a reader must know what a leaf is, what leaf length is and what a ‘cm’ is. This is captured in a quantitative description element by including a defined structure, a defined property, a value and the appropriate defined unit (Fig. 3C). In contrast to a qualitative description element, the property must be explicit in a quantitative description element. For example, the statement ‘leaf 5 cm’ is meaningless, the statement ‘leaf length 5 cm’ is more explicit. The property term is associated with one or more values, which are individual numbers (e.g. 5) and must be associated with a defined unit. The taxonomist is free to choose whichever unit applies to their score. For quantitative statements that do not have units, for example number of petals, ‘count’ is defined as a unit. An example of a quantitative description element is ‘leaf, length, 5 cm’. As shown in Fig. 3D, the defined structure ‘leaf’ is associated with the property term ‘length’, which in turn is associated with the value ‘5’. The defined unit ‘cm’ is linked to the value.

INTERPRETATION OF AND AND OR

Within the model, the semantics of AND and OR will be recorded using different methods for describing specimens and for describing taxa.

AND and OR within specimen descriptions. – Fig. 4A shows how AND is recorded. A single description element is created with two or more states that must be assigned to the same property term. Fig. 4B illustrates the example ‘leaf green and brown’ (i.e., variegated with two different colours). OR within a specimen is shown in Fig. 4C. The two or more alternative states are recorded as separate description elements, but the states must be assigned to the same defined structure and property term and therefore be in the same description unit. An example of

this kind of OR is ‘leaf green or brown’, which means that the specimen has some leaves that are green and some that are brown (Fig. 4D).

AND and OR within taxon descriptions. – In a taxon description AND is represented in the same way as in a specimen description (see Fig. 4A). However, the semantics of OR are more complicated. In a taxon description there are two kinds of OR: within-specimen variation and within-taxon variation. For example, the statement ‘leaves obovate or ovate’ could mean that either each specimen within the taxon has some obovate and some ovate leaves (within-specimen variation) or that within the taxon there are some specimens with obovate leaves and some with ovate (within-taxon variation). Within-specimen variation will be represented in the same way as OR in specimen descriptions (see Fig. 4C). Ideally within-taxon variation will be represented by constructing a taxon description from descriptions of two or more varying specimens assigned to that taxon. Any variation would then be represented by a real specimen. However, when interpreting past descriptions it is not always possible to attribute any variation to a particular specimen, especially if a list of cited specimens has not been included. Therefore the taxonomist making the interpretation must decide whether any variation is within-specimen or within-taxon. Within-specimen variation is then recorded as illustrated in Fig. 4C and within-taxon variation is represented by assigning the variation to **virtual specimens**.

Virtual specimens are artificial constructs that simply serve to represent variation and do not reference real specimens. Fig. 5A illustrates how a description can be either a taxon description or a specimen description, and that a taxon description can be made up of either specimens or virtual specimens. Fig. 5B illustrates an example of an interpretation of an existing description of *Torilis* (Umbelliferae) from the Flora of China (Sheh & Watson, in preparation) using virtual specimens. The description of *Torilis* includes the sentence ‘annual or perennial, loose compound or capitate umbels, lateral or terminal and lateral’. This contains three separate statements that describe variation, but it is unclear whether this variation is within specimens or between specimens. Obviously ‘annual or perennial’ is between-specimen variation as one specimen could only be either annual or perennial. However, when reading the next two statements, it is not clear whether any one specimen will have umbels that are either loose compound or capitate or whether some plants have all capitate umbels and some have loose compound umbels. An experienced umbellifer taxonomist may be able to interpret this description in terms of between- and within-specimen variation, but this requires information not contained in the description. The virtual specimen description mechanism gives the taxonomist a way of grouping statements to highlight the two kinds of variation. Fig. 5B shows how ‘annual or perennial, loose compound or capitate umbels’ could be represented. The

example shows the taxon description reading ‘umbels capitate or umbels loose compound’, and two virtual specimen descriptions reading ‘plant annual or plant perennial’, meaning that any one specimen will be either annual or perennial, but the two options will not be found on the same specimen. It may not be possible to say whether the statements relating to the umbels refer to within- or between-specimen variation so these statements are recorded as within-specimen variation, which seems the most likely option. It must be recognised that the arrangement in Fig. 5B is an interpretation and includes information that is not within the original description. This is acknowledged by recording the author of the interpretation as well as the source of the original description.

COMPOUND STRUCTURES

It is often necessary to break structures up into their parts and to describe each part separately. Description elements, therefore, may be divided into those that contain simple structures and those that contain compound structures. The creation of compound structures is performed during the description construction process. Compound structures are therefore part of a description and do not represent a creation of new defined structures. Individual defined structures are arranged in a hierarchy to allow the description of compound structures using a ‘part of’ relationship (see Fig. 6A). For example, as shown in Fig. 6B, ‘leaf’ is related to ‘margin’ which is in turn related to ‘teeth’ via a ‘part of’ relationship to form the compound structure ‘leaf margin teeth’. The compound structures are arranged into description units according to the structure at the top of the hierarchy even if the description element actually describes the structure at the bottom. Therefore, the example shown in Fig. 6B would form part of the ‘leaf’ description unit not the ‘teeth’ description unit.

RANGES

Qualitative Ranges. – When recording qualitative ranges, the categorisation of continuous variation into discrete states becomes a problem. Descriptions at the moment do not give any idea of the author’s categorisation so the reader may have difficulty understanding the intermediates in a qualitative range. For example, a description of *Paspalum biaristatum* (Gramineae) (Filgueiras & Davidse, 1994) describes the apex of the upper lemma as ‘minutely papillate to conspicuously ciliate’. It may not be difficult to imagine the states that could fall between these two extremes, but it will usually be unclear whether these states are the same as the author had in mind. A description such as this is almost useless for identification because a user cannot be certain that the specimen they have falls between these two extremes. It has therefore been decided that qualitative ranges such as acute to acuminate and racemose to capitate will not be handled in this model. Instead the user will be encouraged to either record

the range quantitatively where possible or define states that break up the continuum of variation without leaving any gaps. For example, a specimen may have leaves with apices that range between acute and acuminate. Ideally this would be recorded quantitatively as a range of angles. However, this range could also either be defined as two states with wide variation and an arbitrary cut off point between the two, for example all leaves with an apex angle of $<50^\circ$ are acute and all leaves with an apex angle of $>50^\circ$ are acuminate, or the user could define states that describe the intermediates without leaving any of the possible angles undescribed by a defined state. The structure leaf apex is then scored with a series of OR description elements that explicitly describe all the possible variations.

Quantitative Ranges. – A quantitative range will be recorded as two linked quantitative description elements representing both extremes of the range. The range is recorded as a relative modified description element (see later section).

MODIFYING DESCRIPTION ELEMENTS

Quantitative and qualitative elements are the basis of a description and will be the most frequently used format. However, occasionally it will be necessary to add further information to description elements. The model, therefore, includes a mechanism by which qualitative and quantitative description elements can be modified. For example, the description element ‘petal, red’ could be modified to ‘petal, usually red’. It is also possible to modify a description unit by using a partially formed description element (i.e. a description element where a defined structure is not related to a defined qualitative state or is only related to a property term without a value). For example, if it was necessary to describe the fact that one structure appeared before another, such as flowers appearing before leaves.

There are four kinds of modified description elements: **frequency**, **relative**, **spatial**, and **temporal**. A description element may have one or more different kinds of modifier. These four kinds of modified description element are discussed below.

Frequency Modified Description Elements. – In order to allow statements that relate to the frequency of assigned scores, description elements can be qualified using **frequency modifiers**. A frequency modifier can be attached to each description element (as shown in Fig. 7A). The fixed list of frequency modifiers is shown in Table 3. An example of a frequency modified description element is ‘flowers rarely white’. As shown in Fig. 7B, the description element ‘flowers, white’ is related to the frequency modifier ‘rarely’ via the description element modifier relationship.

Relative Modified Description Elements. – A relative statement is one that compares the structure being described to another structure, e.g. bracteoles shorter than the flowers. As for frequency description elements, a relative modifier is attached to the first description element. However, unlike frequency description modifiers, the relative description modifier must give direction to the statement. This will allow the identification of the structure that is being referred to and the structure that is making the reference. For instance in the statement ‘bracteoles shorter than flowers’ the direction of the reference is from bracteoles to flowers. No measurements are taken for the property terms that appear in the description elements so the value is ‘undefined’ to show that one structure is simply shorter than the other. Fig. 7C shows how the model captures this. The two description elements included in a relative statement can describe different defined structures or the same defined structure. The finite list of relative modifiers is shown in Table 3. This list includes the relative modifier ‘to’ which is used to link two quantitative description elements that record the extremes of a quantitative range. Fig. 7D shows an example of a relative description element describing a ratio, the length:width ratio 5:1 is represented.

Spatial Information – Landmarks. – Landmarks allow the location of a measurement on a structure to be recorded. For example, the diameter of a tree trunk could be measured at various points (e.g. breast height). In order to interpret the measurement and allow meaningful comparisons with similar measurements, it is crucial that the location at which the measurement was taken is noted. The model handles this by associating description elements with defined landmarks via a landmark modifier (Fig. 8A). The modifier can then target two kinds of objects. One kind is a defined structure, for example, ‘leaf thickness at midrib’ in which ‘midrib’ is the landmark. Alternatively, the modifier could target a defined landmark, which can be any statement defined by the user, for statements such as ‘trunk diameter at breast height’. The landmark modifiers can only take the values shown in Table 3. As an example Fig. 8B illustrates how the statement ‘trunk diameter at breast height, 3 m’ would be handled. The description element representing the statement ‘trunk diameter, 3 m’ is referred to the landmark modifier ‘at’, which in turn refers to the defined landmark ‘breast height’.

Temporal Information. – Some phenomena only appear at certain periods of the year (e.g. flowers in spring) or when other phenomena have already appeared (e.g. fruit after flower). It is therefore important that the model allows the recording of the point in time at which a structure has a particular state. Temporal modifiers relate a description element to another description element or to a temporal statement. A temporal statement is a free text object that allows the

representation of abstract temporal concepts such as seasons, years etc. (e.g. 'in spring'). Fig. 8C shows how the model captures this. The list of temporal modifiers is shown in Table 3. 'Flower green before flower brown' is an example of a temporal description element. Fig. 8D shows that the two description elements 'flower, green' and 'flower, brown' are linked using the temporal modifier 'before'.

CONCLUSIONS

We have presented a model of taxonomic descriptions which will minimise ambiguity in recording character information. The model presented appears to add to the taxonomist's burdens rather than alleviate them, but this system is more explicit than past models and does offer the benefits of data reusability. An important part of any future system will be an effective, easy to use interface to allow the taxonomist to enter data without feeling like they are doing significantly more work than they are at present. It is felt that the advantages of this system outweigh the initial effort in data capture.

The Prometheus II Description model will be used to create an archive of taxonomic data to which others may refer. Ideally, workers will be able to use this archive to write monographs and Floras, or produce cladistic characters for phylogenetics or write a key, without the need for the re-examination of specimens. By creating an archive, it is hoped that a reduction in the loss of data at publication and in the time needed to complete a taxonomic work will be achieved.

The way in which definitions are formulated and used is the subject of future work. The development of an interface to allow easy recording and logical viewing of taxonomic data must be addressed. The ways in which taxonomists make comparisons between 'characters' and taxa and the uses to which taxonomic data are put also investigation.

ACKNOWLEDGEMENTS

We would like to thank Micha M. Bayer and David G. Mann for their constructive comments on the manuscript. The Prometheus II project is funded by BBSRC grants 754/BIO14354 and 95/BIO14353.

LITERATURE CITED (INCLUDING *ELECTRONIC PUBLICATIONS)

Allkin, R. 1984. Handling Taxonomic Descriptions by Computer. Pp. 263–278 *in*: Allkin, R. & Bisby, F. A. (eds.), *Databases in Systematics*. London.

*Anonymous. 2000a. Multiflora: Automatic compilation of taxonomic databases from multiple botanical texts. [<http://www.cs.man.ac.uk/ai/MultiFlora/>].

- *Anonymous. 2000b.** TDWG subgroup: Structure of Descriptive Data, subgroup session report at the TDWG meeting in Frankfurt, 12.11.2000. [<http://www.tdwg.org/tdwg2000/sddreport.htm>].
- Bailey, J.** (ed.) 1999. *The Penguin Dictionary of Plant Sciences*. London.
- Blackwelder, R. E.** 1967. *Taxonomy: A text and reference book*. New York.
- *Cannon, A.** 2001. Prometheus II Project Report: User Requirements Report. [<http://www.prometheusdb.org>]
- Colless, D. H.** 1985. On 'character' and related terms. *Systematic Zoology* 34: 229–233.
- Dallwitz, M. J.** 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–46.
- Davis, P. H. & Heywood, V. H.** 1963. *Principles of Angiosperm Taxonomy*. Edinburgh.
- Diederich, J., Fortuner, R. & Milton, J.** 1997. Construction and integration of large character sets for nematode morpho-anatomical data. *Fundamental and Applied Nematology* 20: 409–424.
- Diederich, J., Fortuner, R. & Milton, J.** 2000. Genisys and computer-assisted identification of nematodes. *Nematology* 2: 17–30.
- Filgueiras, T. S. & Davidse, G.** 1994. *Paspalum biaristatum* (Poaceae: Paniceae), a new serpentine species from Goiás, Brazil and the second awned species in the genus. *Novon* 4: 18–22.
- Frstrup, K.** 1992. Character: current usages. Pp 45–51 in: Keller, E. F. & Lloyd, E. A. (eds.), *Keywords in evolutionary biology*. Cambridge.
- Hewson, H. J.** 1988. *Plant Indumentum: A handbook of terminology*. Canberra.
- Hickey, L. J.** 1973. Classification of the architecture of dicotyledonous leaves. *American Journal of Botany* 60: 17–33.
- Huang, T.-C.** (ed.) 1993. *Flora of Taiwan 2nd Edition Volume 3*. Taipei.
- Jacobs, M.** 1969. Large families - not alone! *Taxon* 18: 253–262.
- Kelly, M. G., Bayer, M. M., Hürlimann, J. & Telford, R. J.** In Press. Human error and quality assurance in diatom analysis. Pp. 75–92 in: du Buf, J. M. H. & Bayer, M. M. (eds.), *Automatic Diatom Identification*. Singapore.
- Kitching, I. J., Forey, P. L., Humphries, C. J. & Williams, D. M.** 1998. *Cladistics: The theory and practice of parsimony analysis*. Oxford.
- Lawrence, G. H. M.** 1951. *Taxonomy of Vascular Plants*. New York.
- Leenhouts, P. W.** 1968. *A guide to the practice of herbarium taxonomy*. Utrecht.
- Machatschki-Laurich, B.** 1926. Die Arten der Gattung *Biscutella* L. sectio *Thlaspidium* (Med.) DC. *Botanisches Archiv Koenigsberg* 13: 1–115.
- Maddison, D. R., Swofford, D. L. & Maddison, W. P.** 1997. NEXUS: An extensible file format for systematic information. *Systematic Biology* 46: 590–621.
- Mukherjee, P. K. & Constance, L.** 1993. *Umbelliferae (Apiaceae) of India*. New Delhi.

- Pullan, M. R., Watson, M. F., Kennedy, J. B., Raguenaud, C. & Hyam, R.** 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* 49: 55–75.
- Raguenaud, C., Pullan, M. R., Watson, M. F., Kennedy, J. B., Newman, M. F. & Barclay, P. J.** 2002. Implementation of the Prometheus Taxonomic Model: a comparison of database models and query languages and an introduction of the Prometheus Object-Orientated Model. *Taxon* 51: 131–142.
- Shan, R. H. & Constance, L.** 1951. The genus *Sanicula* (Umbelliferae) in the Old World and the New. *University of California Publications in Botany* 25: 1–78.
- Sheh, M. L. & Watson, M. F.** In preparation. *Torilis*. in Wu, Z. Y. & Raven, P. H. (eds). *Flora of China, vol 14*. Beijing.
- Sivarajan, V. V.** 1991. *Introduction to the Principles of Plant Taxonomy*. Cambridge.
- Stearn, W. T.** 1983. *Botanical Latin: History, Grammar, Syntax, Terminology and Vocabulary*. London.
- Stuessy, T. F.** 1990. *Plant taxonomy: the systematic evaluation of comparative data*. New York.
- Watson, L.** 1971. Basic taxonomic data: the need for organisation over presentation and accumulation. *Taxon* 20: 131–136.
- Wiley, E. O.** 1981. *Phylogenetics: the Theory and Practice of Phylogenetic Systematics*. New York.
- Wilkinson, M.** 1995. A comparison of two methods of character construction. *Cladistics* 11: 297–308.

Table 1. Key to the notation. The notation used within this paper follows that used in Pullan & al. (2000) and Raguenaud, & al. (2002).



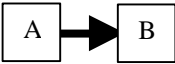
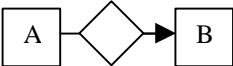
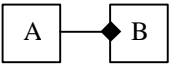
| Symbol | Explanation |
|---|--|
|  | An object within the system |
|  | A relationship between two objects. The relationship is also viewed as an object to which properties, such as ordinality, may be added and to enable appropriate navigation of the system. |
|  | A is a kind of B |
|  | A is related to B |
|  | A is a part of B |
| | Description element |
| _____ | Description unit |
| - - - - - | Description |

Table 2. The fixed list of property terms. Each property term may have multiple definitions.

| Qualitative | Quantitative |
|-------------|--------------|
| Arrangement | Angle |
| Colour | Density |
| Dehiscence | Diameter |
| Development | Height |
| Form | Length |
| Fusion | Number |
| Habit | Width |
| Life Cycle | |
| Orientation | |
| Persistence | |
| Presence | |
| Sex | |
| Shape | |
| Smell | |
| Symmetry | |
| Texture | |

Table 3: The fixed list of modifiers divided into frequency, relative, spatial (landmarks) and temporal. Each modifier has only one definition.

| Kind | Modifier |
|-----------|---|
| Frequency | often usually sometimes mostly rarely |
| Relative | greater than less than greater than or equal to less than or equal to equal to not equal to ratio |
| Spatial | at above below between |
| Temporal | after before while |

Fig. 1. A diagram illustrating the relationships between terms and defined terms. For a key to the notation see Table 1. Structure terms, property terms, qualitative state terms, modifier terms and unit terms are kinds of term and when defined become types of defined term. A defined qualitative state term must be assigned to a property term. Modifier terms and unit terms have only one definition per term, whereas all other terms can have multiple definitions.

Fig. 2. A diagram illustrating the relationships between description element, description unit and description. A description element is part of a description unit, which in turn is part of a description. A description must have an author. (Table 1 for key to notation.)

Fig. 3. Diagrams illustrating the differences between qualitative and quantitative description elements. A. a diagram illustrating a qualitative description element, where a defined structure is linked to a defined state. The property is recorded within the definition of the state. B. a diagram illustrating the qualitative description element, 'leaf, obovate', where the defined structure 'leaf' is linked to the defined state 'obovate'. C. a diagram of a quantitative description element with the defined structure linked to a defined property, a value and defined units. D. a diagram of the quantitative description element, 'leaf, length, 5 cm', where the defined structure 'leaf' is linked to the defined property 'length', the value '5' and the unit 'cm'. (Table 1 for key to notation.)

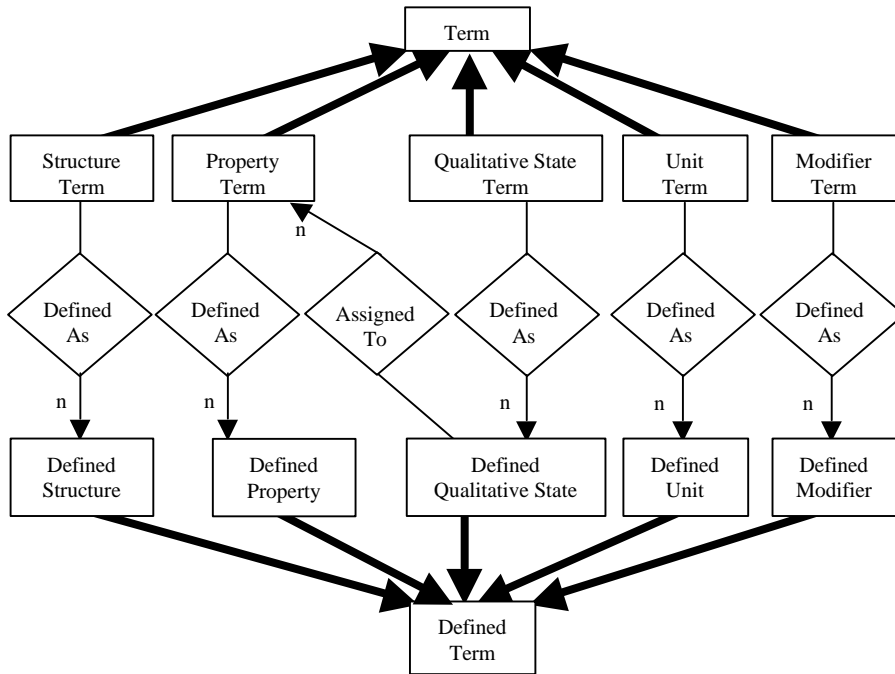
Fig. 4. Diagrams depicting AND and OR in specimen descriptions. A. shows a diagram of AND in which one description element has two defined states assigned to the same property term within their definitions. B. illustrates an example, using the description element 'leaf green and brown'. C. depicts OR showing two separate description elements with defined states assigned to the same property. D. illustrates the example 'leaf green or brown'. (Table 1 for key to notation.)

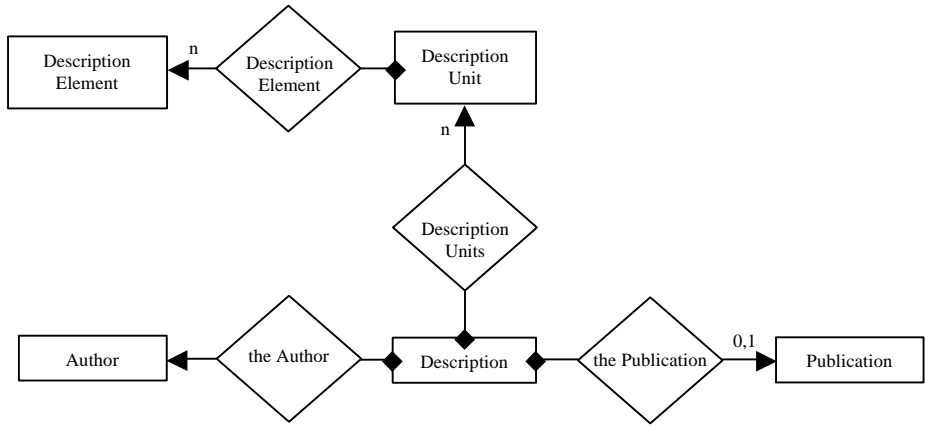
Fig. 5. A. a diagram illustrating specimen and taxon descriptions, which are kinds of description, and virtual specimen descriptions, which are a kind of specimen description. This diagram also shows that a taxon description must contain specimen descriptions, which circumscribe that taxon. B. a diagram showing how virtual specimen descriptions are used to record between-specimen variation within an interpretation of a previously published taxon description. The diagram shows the statement 'umbels perennial or annual' recorded as between-specimen variation by constructing two virtual specimen descriptions, one reading 'umbels perennial' and the other 'umbels annual'. (Table 1 for key to notation.)

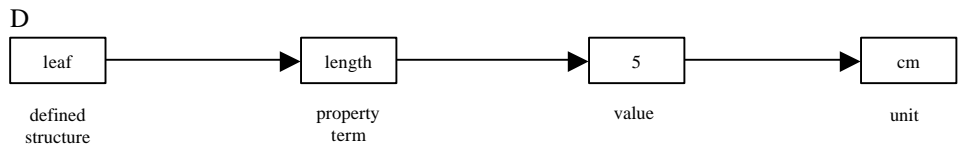
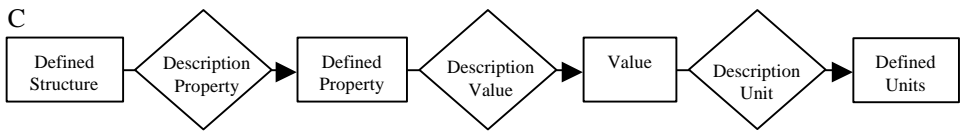
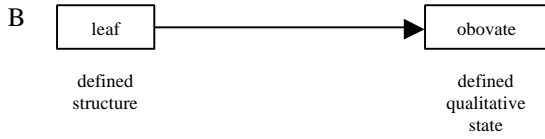
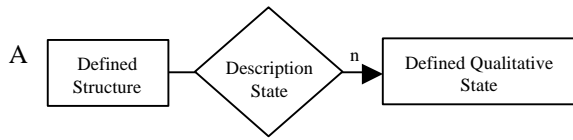
Fig. 6. Diagrams illustrating compound structures. A. shows a defined structure linked to a defined structure to form a compound structure. B. depicts the example, 'leaf margin teeth'. (Table 1 for key to notation.)

Fig. 7. Diagrams illustrating frequency and relative modifiers. A. a diagram of frequency modifiers showing that a description element is linked to a frequency modifier. B. a diagram of the example 'flowers rarely white'. C. a diagram of relative modifiers showing that a description element is linked to a relative modifier, which may have a value with defined units. D. a diagram showing the example 'leaf length:width ratio 5:1'. (Table 1 for key to notation.)

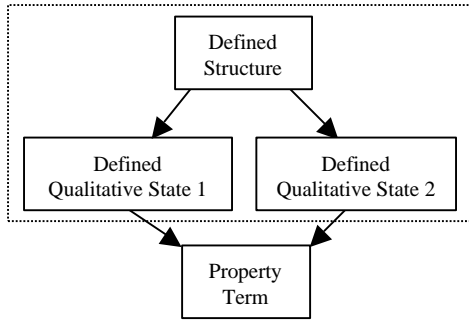
Fig. 8. Diagrams depicting landmark and temporal modifiers. A. a diagram of landmarks showing that a description element is linked to one or two landmarks, which may be defined structures or defined landmarks. B. an example of the landmark modified description element 'trunk diameter 3 m at breast height'. C. a diagram of temporal modifiers showing that a description element is linked to a temporal modifier, which may be another description element or a temporal statement, via a modifier value, which is chosen from a fixed list. D. illustrates the temporal modifier example 'flowers green before flowers brown'. (Table 1 for key to notation.)



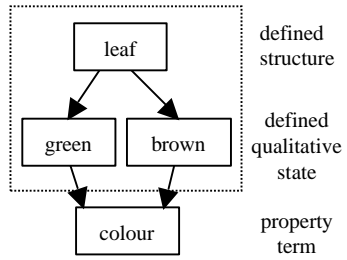




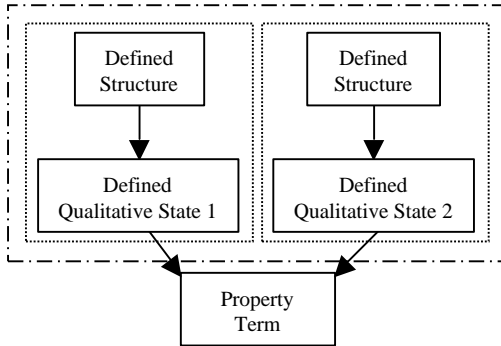
A AND



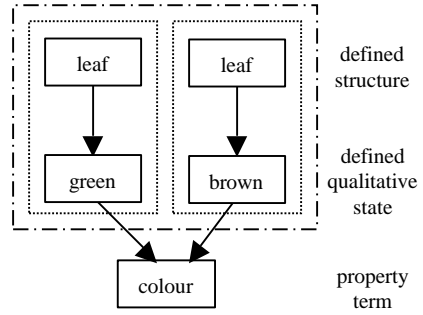
B

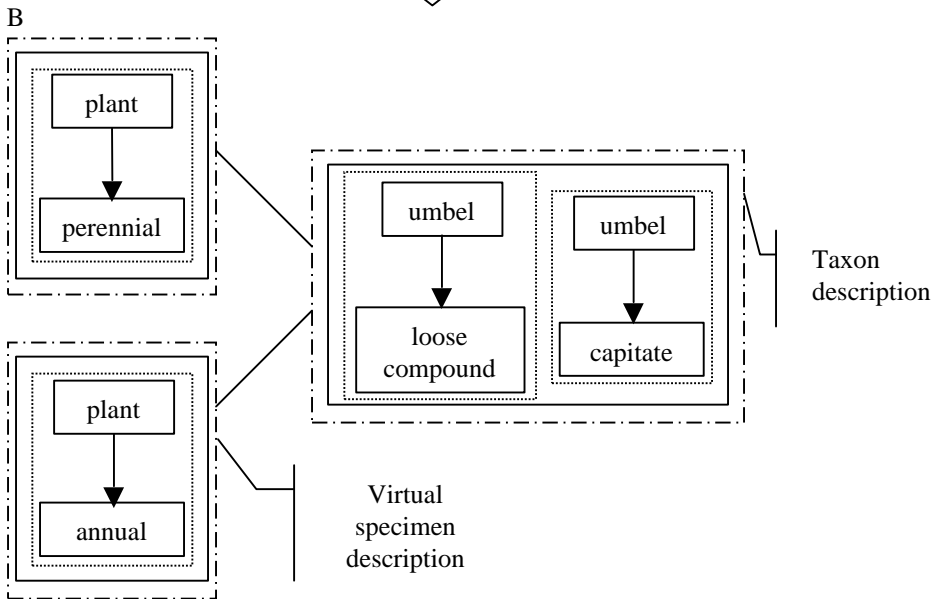
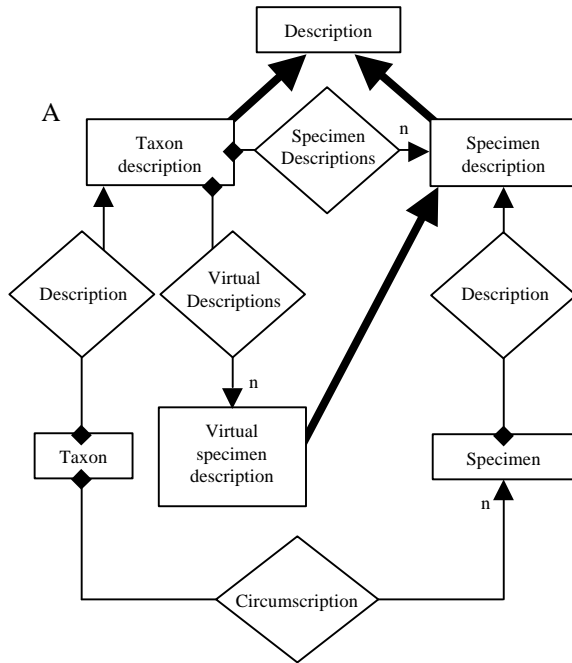


C OR

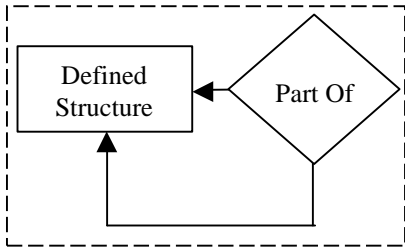


D





A



B

