CHARACTERS AND DESCRIPTIONS

The Prometheus Description Model: an examination of the taxonomic description-building process and its representation

Martin R. Pullan[1], Kate E. Armstrong[1], Trevor Paterson[2], Alan Cannon[2], Jessie B. Kennedy[2], Mark F. Watson[1], Sarah McDonald[1] & Cédric Raguenaud[1]

[1] *Royal Botanic Garden, Edinburgh EH3 5LR, U.K. m.pullan@rbge.org.uk (author for correspondence); k.armstrong@rbge.org.uk; m.watson@rbge.org.uk; sarahmc28@hotmail.com*
[2] *School of Computing, Napier University, Edinburgh EH14 1DJ, U.K. trevor.paterson@bbsrc.ac.uk; A.Cannon@napier.ac.uk; J.Kennedy@napier.ac.uk; cedric@raguenaud.org*

A model for representing taxonomic descriptive data is presented. The model has been developed in response to the growing requirement for the global exchange of descriptive data. Meaningful exchange of data requires that data be represented in a form that can be consistently parsed and interpreted, requiring a common data model and the constrained and explicitly defined use of descriptive terms. The model presented here is divided into two parts that address both of these issues. A new data model for the representation and storage of taxonomic descriptive data is proposed that builds on and extends the best features of current descriptive data models and formats. An ontology-based model for defining and constraining the use of descriptive terms is also presented. The model is based on an analysis of current taxonomic working practices and the processes involved in generating a description. The model takes a specimen-oriented approach allowing descriptive data to be represented through a range of levels of abstraction from actual measurements of structures on a specimen to abstract descriptions of the features expected to be found on a specimen that is a member of a particular taxon. A comparison and discussion of the important aspects of the new model relative to existing models is presented.

KEYWORDS: descriptions, ontology, Prometheus description model, taxonomic descriptive data.

INTRODUCTION

Expectations of the taxonomic community are changing, driven largely by the desire to make better and more efficient use of earlier research, mediated by the storage and exchange of taxonomic information in digital form. Progress in this area is, however, hampered by the subjective nature of taxonomy and the continued use of natural language to represent complex concepts, in which the use of terminology is undefined and uncontrolled (Sokal & Sneath, 1963; Shetler, 1975; Diederich, 1997). Taxonomists strive to classify organisms into an ordered hierarchical system, reflecting perceived natural relationships as indicated by shared characteristics between separate taxonomic entities and the interpretation of morphological and genetic variation. Traditionally taxon concepts, and the character concepts upon which they are based, are communicated through natural language descriptions. Integral to the process of developing descriptions are the definition of the character concepts used and the recording of observed character states. There is, however, no standard method for achieving this, and the published results of taxonomic revisions vary both in their internal consistency and how well the definitions of the concepts being used are expressed. This causes few problems when the information presented is being interpreted and assimilated by the human reader. Yet, when viewed from the perspective of computerized information management, this lack of standardization creates many problems, and the effective sharing of information from disparate taxonomic sources becomes problematic. This is further complicated by the fact that taxonomy is a subjective reaction to our surroundings and a consequence of a natural unconscious preoccupation of man to categorize the components of his environment (Davis & Heywood, 1963). Outside of the domain of science the motivation for creating a classification will vary from circumstance to circumstance (e.g., dividing plants into those that we can and cannot eat, grouping together those that provide us with useful building materials, etc.). Consequently, the basis upon which differences between plants are recognized will also vary from circumstance to circumstance. Although in principle the formal outputs of taxonomy are intended to produce a single overall classification of organisms independent of intended use, examination of descriptions

from disparate sources indicates that in practice the natural tendency to classify using different approaches for different situations still prevails (Sokal & Sneath, 1963). Typically descriptions (and character concepts) are created and compiled for a particular project (e.g., a revision or Flora) and lack a uniformity of approach when considered across projects. Often this means that data generated within a project are only applicable to that project, and it is difficult to transfer those data to other systems or to merge them with other data in a meaningful way.

There are two issues that need to be addressed in order to improve computerised handling of taxonomic descriptions. The first is the establishment of a common model for representing taxonomic description data. The second is the exertion of control over terms used to convey descriptive concepts, so that a common framework of understanding can be established that is accessible to both humans and computers. In order to enable consistent interpretation and retrieval of data both by humans and computers a clear data structure must be established. However, natural language, the common medium for dissemination of taxon descriptions, is fluid and defies strict categorisation. Subtle nuances of language are used to convey differences that are context specific. The intended and actual interpretation of terms used to convey such nuances will vary from individual to individual and are therefore difficult to capture in a highly structured system in which terms are considered to have only a single interpretation. Although placing constraints on language will enhance the consistency of communication, it will inevitably reduce the expression of individual style in descriptions. This does not necessarily lead to loss of information; yet, it is the restricted expression of individual style that many taxonomists find difficult to accept. Nevertheless, if taxonomy is to truly embrace the computer age and accommodate the changing expectations of the broader biological community, it too needs to change. The challenge is to develop a system that minimizes the pain of the transition. It must restrict individual expressiveness as little as possible, yet maximize the level of standardisation, ultimately assisting the taxonomist in generating clear and unambiguous descriptions.

MODELING CHARACTER CONCEPTS FOR DESCRIPTIVE TAXONOMY

The need for a standard means of representing descriptive information has long been recognised and a number of electronic description formats have been developed to allow the storage and analysis of data, for example: DELTA (Dallwitz, 1980), NEXUS (Maddison & al., 1997), and NEMYSIS (Diederich, 1997; Diederich & al., 1997, 1998, 2000). There are also a number of groups that are currently active in developing products in this area such as the Standardised Descriptive Data working group of TDWG (TDWG-SDD, 2003) that is working on a replacement for DELTA, and the BioLink group of CSIRO (CSIRO, 2001) that is working on developing new ways of representing taxonomic description data by extending and revising the DELTA format. Of the electronic description formats, DELTA is the most fully featured for use in taxonomic revisions, Lucid (http://www.lucidcentral.org) being designed for creating keys and NEXUS being a general extensible file format for any systematic information. All of these models have their merits and drawbacks, yet certain issues persist. Specifically, two points need to be considered: the conceptualization of characters (i.e., how character concepts are generated within the mind of a taxonomist and how those concepts are represented), and the granularity of data, (i.e., the degree to which descriptive statements are broken down into their component elements).

Problems with character. — The data formats listed above all attempt to build a representation of the basic element of a description, often referred to as a *character*. The different systems, however, conceptualize character in different ways. These differences have arisen due to the different interpretations of the term character that exist within the taxonomic community (Colless, 1985) and differences in the intended use of the information that is being stored. Inconsistencies also arise in the derivation of character concepts during the taxonomic process when observed variation is partitioned into characters. Therefore, while it will be readily agreed that characters are the basic building blocks of descriptive data, there is little consensus amongst taxonomists as to what the term character actually means. Colless (1985) consulted 50 publications and found nineteen different explicitly stated or clearly implied definitions of character. This plainly illustrates how, with such varied interpretation, the term character has lost most of its meaning and value, making the precise interpretation of taxonomic descriptions problematic.

The DELTA character model. — In the most general sense a character can be defined as a statement on a feature of an organism. This understanding naturally leads to the modelling of character as a two-part construct. DELTA, the first published format for taxonomic descriptive data (Dallwitz, 1980), follows this two-part model of character, and the next generation of DELTA under development by the BioLink group appears to continue to adhere to this practise (CSIRO, 2001). Herein, the character is often named descriptively using a character name such as "striated area on petal apex <presence>", and associated with a set of character states (in this example "present" and "absent") that detail the con-

ditions in which the named character can be found (Dallwitz & Paine, 2004). In the DELTA model, the combination of character name and character states constitutes the character. This approach reflects the traditional way taxonomists conceptualize characters and the way such data are presented in natural language descriptions. It is common practise to consider the characters in a DELTA dataset and the objects they are describing as forming a matrix; each cell being filled with an appropriate character state (Dallwitz, 1980). This matrix-based approach is then used as the foundation for identification systems based on the DELTA format. The NEXUS format (Maddison & al., 1997) is also rooted in the need to represent a matrix of characters and taxa in a computer readable format and is therefore also based upon a two-part character model. Although the internal representation of the information differs significantly from that of DELTA, the underlying conceptual model is the same.

As pointed out by Diederich (1997), the two-part character model leads to problems with the consistent parsing of descriptive data. In the DELTA model the allowed content of character is loosely defined with no clear distinction between the information that should be included in the character name, that which should be included in one or more of the character states, or how information should be divided between characters within a dataset. The example given above serves to illustrate this point. The character provided is "striated area on petal apex, present". Within this statement a combination of structure (petal apex), property (presence) and state information (striated) are found, although the state information does not relate to the property actually being described by the character. As a result, although the apex of the petal is the central object being discussed, it is cryptically encoded within the character and not directly accessible using the standard DELTA parser. Furthermore the same information could have been encoded in a different way to the same effect, e.g., "area on petal apex <texture>" with an associated score of "striated", wherein the presence of the striated area is cryptically encoded. Different DELTA users can, therefore, code the same information in a wide variety of ways. Although this approach has the advantage of not constraining the user to any particular way of thinking or mode of working, and therefore does little to limit the expression of individual style, it is unsatisfactory in terms of the transfer of information between systems. This is especially true when DELTA sets from disparate sources are to be merged. That is not to say that it would not be possible to produce compatible data sets using DELTA, but, it would be the user's responsibility to ensure compatibility before merging or comparing DELTA data. The fact that DELTA was modelled in this way is not surprising given that at the time the format was initially conceived, electronic information exchange was still in its infancy and the problems arising from the need for widespread exchange of information were not an issue. The creators of the DELTA model were more concerned with the basic problem of representing descriptive data in a computer readable format, which could be used for a range of purposes on a single system. DELTA, therefore, provides a useful model for the organization and scoring of data, yet its flexible use of language and character construction prohibits standardized data-sharing, making it unsuitable for our purposes.

Diederich's character model. — Diederich (1997) suggested a solution to this problem by increasing the granularity of the character model, moving from a two-part to a three-part representation consisting of a *structure* + *property* + *state*, and imposing a controlled vocabulary with which to populate these triplets. Although this requires the user to adhere to the vocabulary, this approach overcomes a number of issues by clearly partitioning the constituent elements of a descriptive statement into its fundamental parts; specifically the object being described (*structure*, e.g., leaf), the aspect of the object being described (*property*, e.g., shape) and the state of the object (*state*, e.g., lanceolate). Terms belonging to the different categories of structure, property, and state cannot be compounded. As such, state terms cannot be used within a structure identifier and thus the term "striated area" as used in the DELTA example above would not be allowed. Each element of the triplet is coded using one or two words rather than the verbose coding found within DELTA. In so doing, data can only be coded one way, even when entered by different authors. For example, the only way to represent the data encapsulated in the statement "striated area on petal apex, present" would be to create two triplets, one declaring the area on the petal apex to be present and the other declaring the area to be striated. A consequence of this approach is, however, that the expression of complex statements is often more laborious than it would be in DELTA. The Diederich model may require the construction of multiple triplets in order to express the same information that could be encapsulated in a single compound DELTA character state or character name. This also means that the possibility of recreating aesthetically pleasing natural language descriptions reflecting individual style is reduced, although as stated earlier, computerised parsing of the information is greatly simplified, and the potential for reuse of the data for purposes other than those which they were originally collected is increased.

The Prometheus character model. — As a result of the factors described above relating the ability to consistently parse descriptive data, when developing the Prometheus character model it was decided to follow

the principles laid down by Diederich. Therefore, in Prometheus character model the three-part approach to representing characters has been adopted. However, the semantics of the basic and modified triplet presented by Diederich have been extended to create a richly featured format for storing descriptive information. Furthermore, because of the confusion surrounding the precise meaning of the term character, its use has been avoided in this model. The term *description element* has therefore been used instead; the basic form of which is shown in Fig. 1A.

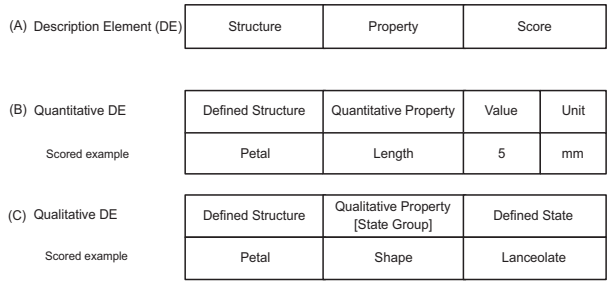

Fig. 1. A, Description Element (DE) is used to represent an atomic character statement. A conceptual DE is composed of a structure, property and score triplet. Quantitative DEs (B) record the defined structure term, defined property term and the score as a numerical value, which may have an associated defined unit term. The property for a qualitative DE (B) is represented as a state usage group (see later) and the score is recorded as a defined state term.

REPRESENTING MODIFIERS AND RELATIVE STATEMENTS

The basic description element is sufficient to record simple descriptive data. However, it is common practice among taxonomists to embellish their data with qualifying adjectives such as "rarely" and "often". It is usually intended that such adjectives enhance the communicative power of a description, however, they are highly subjective and have no clearly defined interpretation and so usually introduce fuzziness rather than clarity into the information being communicated. Nevertheless, so as not to overly restrict the expression of individual style, a mechanism for incorporating such modifiers into the basic description element has been elaborated. The mechanism is designed so that when querying descriptions, it is possible to ignore these modifiers without detrimental effect to the results.

The most basic modifiers are *frequency modifiers*. These are simple terms that can be added as an attribute of an abstract description element in order to indicate relative occurrence (e.g., mostly, often, usually, sometimes, rarely, etc.). Other, more complex modifiers capture how one state relates to another (e.g., leaf length in comparison to leaf width). These modifiers link source and destination description elements, and have associated terms to indicate the nature of the relationship being expressed. Three forms of such statement are distinguished: *relative modifiers*, *spatial modifiers* and *temporal modifiers*.

Relative modifiers allow two undefined scores to be related using the terms greater-than, less-than, equal-to, ratio, not-equal-to, less-than-or-equal-to, greater-than-or-equal-to (e.g., leaf length "less than" leaf width). A relative modifier may include a value to indicate the magnitude of the relationship (e.g., length is twice width: length "ratio: 2" width). *Spatial modifiers* allow the location of measurements to be more accurately specified by identifying a structure on the structure being measured and specifying the relative position of the two objects using at, above, below or between (e.g., diameter "at" branch). Often, however, it is not possible to localise a measurement using a structure; to accommodate this it is possible to use a *landmark statement* instead. A landmark statement is simply a descriptive phrase; for example, trunk diameter "at" <breast height>, where breast height would be recorded as a landmark statement. *Temporal modifiers* allow the time of year or sequential order of events to be recorded. A temporal modifier either relates two description elements (e.g., fruit green before fruit red) or a description element and a temporal statement (e.g., flowers present "during" <spring>) using after, before, during or while. The various forms of modified description element are illustrated in Fig. 2.

The description element forms the basic building

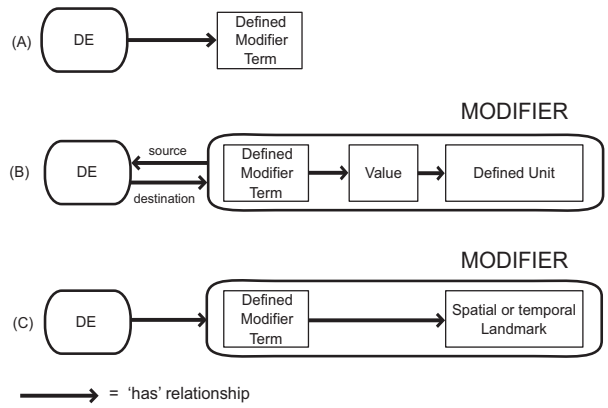

Fig. 2. In order to represent the nuances of natural language character descriptions, description elements (DEs) can be modified in a variety of fashions. A, A defined modifier term can be stored as an attribute of a DE (e.g., usually, rarely). B, two DEs can be compared via a modifier element which records a modifier term (e.g., relative modifiers and values: ratio, less than, etc.) and may have an associated score. C, a modifier can contain a modifying statement (e.g., <at> "chest height").

block of the Prometheus description model. By creating combinations of description elements a complete description of a specimen or taxon can be constructed. However, in order to determine the appropriate ways of combining description elements, it is first necessary to understand the general description-building process employed by working taxonomists.

UNDERSTANDING THE DESCRIPTION-BUILDING PROCESS

The process by which a description is constructed is one of abstraction. In general there are three ways in which abstraction is apparent in this process:

- the transition from describing specimens to describing taxa;
- the transition from recording quantitative measurements to recording qualitative scores;
- allowing statements about ranges either qualitative or quantitative.

Considering the various combinations of these influences, a number of degrees of abstraction can be identified in the general taxonomic process:

1. Concrete statements. — At the most concrete level a statement records the measurement of a single instance of a structure; for example, the length of a single leaf on a specimen. The description of the specimen may contain one or more such statements, each recording the length of a different leaf.

2. Abstract structural statements. — Concrete scores can be abstracted into a single statement about a specimen. This happens when a generalised measure for an abstract structure is recorded. For example the average, mininum, maximum or modal leaf length may be recorded, or the upper and lower bounds of leaf length. Such a statement is considered abstract because it relates to no specific leaf on the specimen and generalised because the measure recorded is only a summary of the actual values.

3. Specimen-based structural contextualisation. — Scores can be partitioned within a specimen. For example a distinction may be made between the measurements of basal and apical leaves on a stem. In this case the abstract concept of leaf has been subdivided into two kinds of stem leaf, one basal and one apical.

4. Specimen-based abstract (qualitative) statements. — When contextualisation of the abstraction of a structure has occurred, the scores obtained from each of the kinds of structure can themselves be abstracted into a qualitative score such as basal leaves short, apical leaves long.

5. Taxon-based abstraction. — A further level of abstraction can be obtained by amalgamating specimen descriptions into the description of a taxon. An increase in abstraction is caused here because both the structure being described (e.g., leaf), and the object on which the structure is found (the taxon), are abstract. The reference to the structure relates to no particular structure and the reference to taxon relates to no particular specimen.

6. Taxon-based abstract (qualitative) statements. — The final level of abstraction occurs when the descriptions of multiple taxa are considered together as is required for comparative biology, i.e., the recognition of common traits and variation between taxonomic entities required for their identification. During this process the measures found in the previous levels are often abstracted into qualitative scores reflecting discontinuities in the observed variation in the study material. For example the leaf length measurements obtained from a series of taxa may be subdivided into the qualitative scores short and long reflecting a single discontinuity in the variation of leaf length between the taxa. These scores are abstract because they relate to no particular measurement on any single specimen and are context dependent reflecting only a relative difference in the length of the leaves without indicating a specific value for the length.

In order to produce a new specimen-based taxon description it is necessary to pass through most of the stages of abstraction identified above even if the results of each stage are not explicitly recorded. The end result of the descriptive process is a collection of statements indicating the state of various structures in the object being described, potentially representing various levels of abstraction.

REPRESENTING DESCRIPTIONS IN THE PROMETHEUS DESCRIPTION MODEL

The Prometheus description model allows for the collection of both abstract and concrete data, and permits the recording of multiple concrete scores where useful. The model is intended to allow the capture of such information so that in future alternative abstractions could be applied to them, and it is hoped that any software based upon it will encourage the capture of concrete scores. This model does not, however, provide a mechanism by which the abstraction process can be performed automatically. In the model each description element contains a reference to a description object, and therefore the collection of description elements that reference the same description object can be considered to represent a description. The description object contains a reference to the entity being described e.g., a specimen. In this way

the model allows for the recognition of the various forms of abstraction that are to be found in a description and permits multiple descriptions of the same entity to be stored (Fig. 3).

Concrete vs. abstract data (abstraction levels 1 & 2). — In the Prometheus description model description elements are explicitly recorded as being concrete or abstract. Once this distinction has been made, certain behavioural constraints are placed upon the element if it is declared to be *concrete*:

• A concrete description element must link to a specimen not a taxon.

• A concrete description element may not contain ranges or summary scores such as mean or mode.

Representing structural contextualisation (abstraction level 3). — Structural specialization is primarily recognised by the context in which the structure is recorded. The context is defined by a structural path in which a series of structures are linked together. For example stem leaves can be represented by the structural path *stem* → *leaf*. Further specialisation can be recorded using spatial modifiers (see: Representing modifiers and relative statements above).

Representing quantitative vs. qualitative (abstraction level 4). — Both quantitative and qualitative data are gathered by taxonomists. However, while the decomposition into description elements is easily applied to quantitative scores, often with qualitative statements the associated property is less readily discernible and not explicitly recorded. To accommodate the distinction between the two kinds of statement the Prometheus description model includes two forms of description element: quantitative and qualitative (Fig. 1B & C).

An example of a quantitative statement is "leaf length 5 cm". In order to record this statement in a description element it is necessary to specify a structure (leaf), an explicit property (Length), a value (an individual number: 5) and the appropriate defined unit (cm) (Fig. 1B). For quantitative statements relating to the property "number", no units are required. The current model provides a minimal list of named quantitative properties {Angle, Density, Diameter, Height, Length, Number and Width}. This list can be expanded to allow further quantitative properties as needed (e.g., Colour, as defined by RGB values.). Detailed ontologies defining measurement concepts (e.g., units and dimensions, etc.) for the biological domain are being developed by others such as the Science Environment for Ecological Knowledge project (SEEK: http://seek.ecoinformatics.org) and could ultimately be used to constrain and define the terms used in Prometheus quantitative description elements.

In natural language expressions of qualitative statements there appears to be no consistency in the inclusion in the statement of the property being described. So for example, the shape of a leaf would as often be described in a statement like "leaf shape oval" as it would be by the statement "leaves oval" in which the property *shape* is only implied. It appears that it is often assumed that the property being implied, and any associated alternatives that may be included in other descriptions, can be unambiguously deduced by the reader. For human readers this may well be the case, but no such assumption can be made on the part of a computer. It is therefore imperative that the property being described be included in the statement if it is to be automatically processed. However, it is apparent that to force such a requirement on users each and every time they wish to construct a statement would be overly burdensome. The Prometheus description model has therefore been constructed in such a way that a simple statement such as "leaves oval" can be represented directly in a qualitative description element composed of a defined structure (leaf) and a defined qualitative state (oval) (Fig. 1C). Note that no explicit property is specified in the description element, however it is a requirement of the system that when a qualitative state is being defined, it is associated with a property, for example the state oval would be associated with the property shape, and the description element just described will accordingly have an implied property of shape as a result of this association. The association between qualitative state and a property is a requirement imposed by the Prometheus ontology and will be discussed in more detail later in the paper.

Arguably it should be possible to describe all physi-

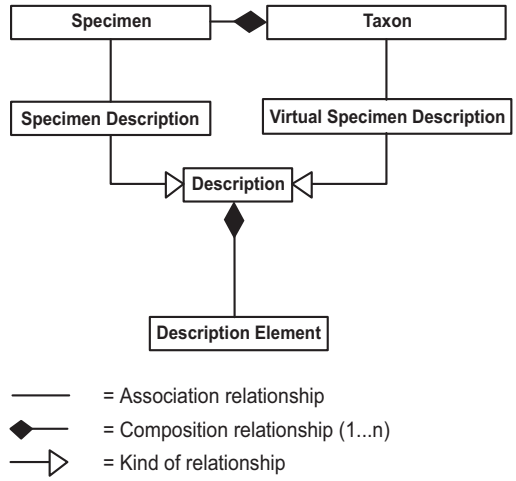

Fig. 3. A description is composed of description elements. A specimen can have multiple specimen descriptions. A taxon can be circumscribed by its component specimens (Pullan & al., 2000), and can be described by virtual specimen descriptions.

cal data quantitatively and, as stated earlier, Prometheus encourages quantitative description wherever practicable. However, often this is neither reasonable nor useful for taxonomists, who take the expedient approach of categorising continuous quantitative variation into a set of more easily handled discrete qualitative states with a concomitant increase in abstraction and loss of precision. Often there is only a vague description of the circumscription of these states; usually because the detail required to describe such states in absolute quantitative terms would often be prohibitive. For example, leaf shape is usually described in terms of discrete states such as linear or lanceolate, although in reality leaf shape is a continuum that could be numerically recorded through application of appropriate mathematics. Moreover, some properties such as texture defy numerical description and there is no option but to handle them qualitatively. For these reasons a concrete description element can be either quantitative or qualitative, even though assigning a qualitative score requires a degree of abstraction.

Representing variation. — It is common practice for a description to indicate variation within and between the entities being described. Two circumstances can be recognised in which it is necessary to record variation:

1. Recording variation within a single specimen. — There are various nuances of this depending upon whether the description is concrete or abstract and whether the information expressed is qualitative or quantitative.

a. Concrete quantitative variation in a specimen is represented in a description by recording multiple concrete description elements for the same property of the same structure, each with a single value. This would allow for multiple measurements of leaf length from a single specimen to be recorded.

b. Abstract continuous quantitative variation in a specimen is represented in a description by creating a single description element in which two values are recorded. The two values represent the extremes of the continuum being expressed. For example, the maximum and minimum leaf length observed on a specimen could be recorded in this manner.

c. Abstract discontinuous quantitative variation is represented in a description by recording multiple abstract description elements for the same property of the same structure. When such a situation is encountered in a description, the information expressed is interpreted as being the "or" of the description elements. This allows statements such as there are 3, 5 or 7 petals on a flower, or even more complicated statements such as when there are 3–5 or 7 petals on a flower, because each description element may contain an expression of a continuous range as described in point "b" above.

d. Abstract qualitative variation within a single structure is represented by creating a single abstract qualitative description element with multiple states scored. The multiple scores are interpreted as the "and" of the selected states. For example, a mottled leaf may be recorded as being green and yellow simultaneously. An important point to note here is the contrast with the equivalent situation in a quantitative description element in which multiple scores in a single element are interpreted as a range. Even though taxonomists often record qualitative ranges in natural language (e.g., leaves ovate to lanceolate), such ranges are explicitly excluded from the Prometheus description model because it is not possible to unambiguously interpret such ranges. This is perhaps best illustrated by consideration of an example. The following text is a quote taken from the definition of the DELTA format (Dallwitz & Paine, 2004) and describes the use of ranges in terms of colour of a structure in which three alternative colours are defined: red, coded as 1; black, coded as 2; and yellow, coded as 3. The dash in the coding indicates that a range is being represented ... "the attributes 2,1–3 and 2,1–2–3 are not equivalent: the former denotes colours between red and yellow (red, orange, and yellow, but not black), while the latter denotes red, black, yellow, and their intermediates" (Dallwitz & Paine, 2004).

It is evident from the above passage that the author of the interpretation of the coding considers orange to be an intermediate between red and yellow. However orange is not explicitly coded in the character definition and the intermediate orange is merely an inference based upon the understanding of colour of the interpreter. What is more, the second range in the example specifies the range red-black-yellow. Where should orange be placed in this order—between red and black, or between black and yellow or indeed even though the interpretation indicated that there are un-stated intermediates, should orange now be considered outside of the range? The only solution to this problem would be to impose a globally accepted ordering for colour; however, it is unlikely that such an ordering can be universally agreed upon for most qualitative data. Such information can be better expressed by explicitly scoring the actual intermediates that are considered to exist using the mechanism described in point "c" above and not leaving it up to the reader (machine or human) to guess at the implied intermediates.

e. Abstract qualitative variation of a structure within a particular specimen is represented by constructing multiple qualitative description elements for the same property of the same structure. This is interpreted as the "or" of the states in the respective description elements.

2. Recording variation between specimens within the same taxon. — A description of a taxon is

intended to provide the user with a mental picture of what a specimen belonging to that taxon should look like. Such a taxon description can be modelled as a description of a *virtual specimen*. A virtual specimen is a highly abstract concept, containing within it a summation of all the kinds of specimen variation described above. However, it is not uncommon to find disjunctions in this variation such that different forms of the taxon can be discerned. Often this is handled in the taxonomic classification so that the disjunctions are partitioned into subordinate taxa (e.g., a species may be divided into a number of subspecies or varieties) and separate descriptions are provided for each of the subordinate taxa. When this is not the case, the disjunction in variation is handled by allowing a taxon description to be composed of multiple virtual specimens.

ESTABLISHING A CONTROLLED VOCABULARY

The model as presented thus far provides a means of representing taxonomic descriptive information in a form that allows consistent parsing of descriptive statements. This is not, however, sufficient to ensure comparability of information across datasets when qualitative descriptions have been captured. Under these circumstances, in order to allow comparability, a standard conceptual framework must be established that constrains and controls qualitative term usage. In other words an *ontology* is required. Ontologies are designed specifically for the purpose of knowledge-sharing and reuse, and in the broadest sense can be defined as "an explicit specification of a conceptualization" (Gruber, 1993). Under this definition there can be many forms of ontology ranging from a simple set of definitions for a formal controlled vocabulary to a complex system of definitions and relationships representing knowledge based on the objects, concepts, and other entities that are assumed to exist in some area of interest and that can be used to reason over the knowledge domain being described (Genesereth & Nilsson, 1987). By specifying a standard controlled vocabulary for the description of specimens and taxa, Prometheus aims to prevent semantic heterogeneity between descriptions that have been composed solely with terms from a common ontology. The ontological aspects of Prometheus specify the definition and controlled usage of descriptive terms. All the components of the underlying data model are represented by defined terms in the ontology, including defined structure terms and defined state terms. Instances of these defined terms in the ontology are used to compose description elements. The use of terms from the ontology is constrained by the relationships between terms asserted in the ontology, as shown in Fig. 4. A demonstration ontology has been created (http://www.prometheusdb.org/resources.htm), which defines the terms and constrains the use of these terms for constructing general angiosperm descriptions.

Definition of terms. — In the model defined terms are represented by the classes *defined structure* and *defined state* and consist of a single word or short phrase, a text definition and may include images to assist in the interpretation of the definition. These definitions are intended for human use only and exist as an aid to the user when selecting the terms to be included in a description. The primary role of definitions is therefore within the user interface where consistency of term selection has to be established. Within an interface based on this form of ontology it is intended that the user will be presented with appropriate definitions whenever faced with a term selection decision. This reinforcement of definition will promote consistency of term selection within a description set for a project, across datasets between projects and between different users.

Synonyms and homonyms. — An examination of a range of botanical glossaries revealed that it is not uncommon to find terms that have multiple concepts embedded within a single definition. For example "excurrent" is defined by Lawrence (1951) as "extending beyond the margin or tip, as a midrib developing into a mucro or awn; or, descriptive of the habit of a plant with a continuous un-branched axis, as *Picea* or *Abies*. The opposite of deliquescent". Wherever this was detected during the construction of the demonstration ontology, the definition was split into its constituent concepts. In

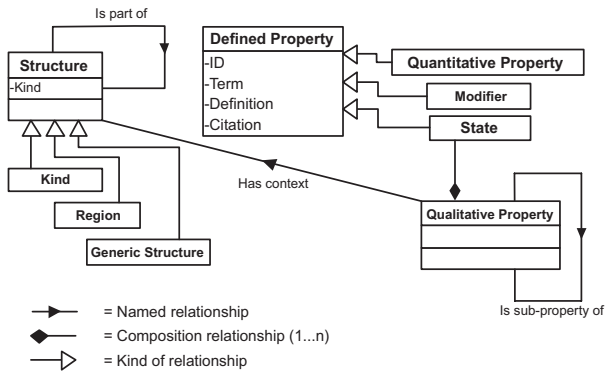

Fig. 4. All terms in the ontology are types of defined term. Structure terms are organized by the optional compositional relationship 'is-part-of'. The specialized structure terms, kinds, do not participate in this relationship, whilst further specialized structure terms (regions and generic structures) are permitted to form part-of relationships with any other structure term. State terms are aggregated into usage groups, which can be represented as a hierarchy of qualitative properties, with contextualization to permitted structure terms.

effect this is the creation of homonyms—multiple terms with the same name but differing definition. In order to ensure that such homonyms can be distinguished by the system, separate defined terms with unique system identifiers are created.

The issue of term synonymy was considered in detail during the development of the demonstration ontology. It was concluded that true synonyms, i.e., two different words associated with exactly the same concept in all contexts are very rare and may not actually exist. Different terms are invented at different times to cope with differing situations and this is reflected in subtle differences in their definitions. Under some circumstances it may be possible to ignore the differences in definition and yet in other circumstances recognising the subtleties may be key to the correct interpretation of descriptions (Shetler, 1975). It is common practise to handle synonymy by creating mappings between terms and a direct analogy can be drawn here between synonymy in the context of descriptive terms and synonymy as it relates to taxon concepts. For many years it was considered adequate to handle taxonomic synonymy by assuming a single classification of names in which all interpretations of a name can be handled. Recent thinking has, however, indicated that this is inadequate and that the simultaneous recognition of multiple overlapping classifications is required if data associated with taxa are to be handled correctly (Berendsohn, 1995, 1997; Pullan & al., 2000; Gradstein & al., 2001; Berendsohn & al., 2003). The same problem applies to descriptive terms and in order to handle term synonymy correctly, multiple overlapping classifications of terms would have to be constructed. However, compared with taxonomic synonymy the network of interpretative frameworks for descriptive terms is vast, and it would be impractical to consider building representations of those frameworks to cater for all eventualities. An alternative solution to this problem is that instead of requiring a formal representation of synonym relationships to be created and stored within the system, allow users to informally and transiently imply such relationships when constructing a query. For example, if a user considers scarlet and vermillion to be synonymous, this can be specified in a query, or when integrating data, by indicating that records matching one or the other should be returned or merged.

RELATIONSHIPS BETWEEN STRUCTURAL ENTITIES IN THE PROMETHEUS ONTOLOGY

Part-of relationships. — *Part-of* relationships allow a representation of the potential structural composition of any organism to be covered by the scope of the ontology. For example, the scope of the demonstration ontology was restricted to only those structures that appear on angiosperms. The network of part-of relationships in the structural ontology does not, therefore, represent any specific organism form. Rather it represents a set of potential structural relationships that may or may not exist in any given specimen. During the process of creating a description, those parts of the ontology that correspond to the form of the particular specimen/taxon being described are flagged as being present.

The structural ontology is rooted at the defined structure "entire plant", and all other structures are related back to this defined structure through a sequence of part-of relationships linking the intermediary defined structures in the chain (e.g., an ovule is part-of an ovary, which is part-of a flower, which is part-of an inflorescence which is part-of the entire plant). This chain of defined structures is termed the structural path and is used to identify and distinguish between different structural contexts for the same defined structure (e.g., stomata on stems compared with stomata on leaves). Therefore, when the structural component of a description element is created by reference to a defined structure in the ontology, the reference is in fact made to the structural path and not just to the terminal defined structure in the path.

Generic structures and regions. — During the construction of the ontology, it became apparent that it was not possible to efficiently or exhaustively represent every conceivable combination of structures that may occur. In the main, the difficulty arose with certain types of structures that can appear in many different structural contexts. For example hairs and pores potentially can be found on any structure. Explicitly recording part-of relationships for all such generic structures would be impractical. The problem was circumvented by adding a subclass of defined structure called generic structure to the model. The use of the generic structures is constrained so that they can only participate in part-of relationships by being appended to structure paths drawn from the ontology during the process of building a description template and not during the process of ontology construction.

A similar situation is encountered when it is required to specify a region of a structure (e.g., a leaf may be divided into basal and apical regions). As with generic structures, it is neither practicable nor desirable to specify all the potential regions of structures. A further subclass of defined structure, region, has therefore been added to the model. Again, as with generic structures, regions can only be added to structure paths during the process of building a description template.

Kind-of relationships. — A third class of defined structure is required to handle summary terms. A summary term such as "berry" refers to a structure but also

carries with it some implicit state information in that it is a fruit, which is always fleshy and indehiscent with seeds submerged in pulp. A *summary structure* is included in the ontology as a defined structure, related to a parent structure through a kind-of relationship (in Paterson & al., 2004 this relationship was referred to as a *type-of* relationship but has subsequently been renamed so as to avoid confusion with taxonomic types). So, for example, in the demonstration ontology, a berry is defined as being fleshy and indehiscent with seeds submerged in pulp and is related to the parent structure fruit to indicate that a berry is a kind-of fruit. As such, summary structures are not directly included in the structural path but exist merely as an attribute of their parent structure. When selected to form part of a description, a summary structure inherits the structure path and substructures of its parent. Future development may include the automatic association of states and/or structures with a kind-of relationship.

State relationships in the ontology. — States as represented by terms found in botanical descriptions typically fall into three categories:

• Basic state terms: These are general terms used to describe structures under a wide range of circumstances without any obvious restriction as to which structures they can be applied. They are represented by defined states and form the largest group of terms in the ontology. As discussed below, basic terms are grouped together according to similarity of use.

• Specific structural state terms: These are similar to basic terms but have a clearly identified structural context (e.g., terms describing stamen arrangement such as monadelphous and diadelphous, which can only be applied to flowers). As with basic terms, these are divided into state groups, and the structural context is defined by linking the state group to one or more defined structures using an applies-to relationship in the ontology.

• Enumerative state terms & presence/absence state terms: A large number of commonly used terms merely express the presence/absence of a structure (e.g., stipulate: possessing stipules), or enumerate a structure (e.g., biovulate: containing 2 ovules). Within the Prometheus description model explicit mechanisms for representing this information already exist. For example, the information represented by the term biovulate can be represented by the quantitative description elements <ovules→present> & <ovules→2→number >. Similarly, the information in the term stipulate can be represented by the qualitative description element <stipule→present>. Presence/absence and enumerative terms are, therefore, considered redundant and have been excluded from the state ontology.

Grouping states. — Grouping of states is required to allow the construction of meaningful comparative statements. For example, in order to contrast the leaf shape in two specimens it is necessary to be able to recognise which elements of two descriptions refer to leaf shape. In essence a classification of states is required. Initial attempts at constructing such a classification aimed to create state groupings using the non-hierarchical concept of property as described by Diederich (1997); of which shape would be an example. However, although a list of properties similar to those identified by Diederich exists for use in quantitative description elements, for a number of reasons this was found to be an unworkable approach when considered in a qualitative context. When reading or writing a natural language text, the process of categorising states into properties appears to occur in a subconscious manner using some sense of the natural affinity of states. Often texts are written without making explicit to which property reference is being made, and it is the job of the reader to interpret where commonalities lie between descriptions. To compound the problem, it is apparent that state terms are not conceived in a property-oriented manner. Rather they are created on an ad-hoc basis for different purposes, at different times. When a new descriptive term is invented, the inventor does not worry about the property into which the term fits. There is, therefore, no single pathway or mechanism for the conception and evolution of state terms, and it is difficult to define a consistent and non-arbitrary mechanism for assigning states to properties. For example, while "red" is clearly a state of colour, it is not clear where a term such as "keeled" naturally fits. It could be classified either as a shape or as an arrangement depending on the preference of the classifier. It was, therefore, concluded that there is no single classification of property into which all state terms will easily fit. Retrospectively placing states into such an arbitrarily selected arrangement yielded unsatisfactory results in which users found it difficult to locate the states for which they were searching and the property suggested for many state terms is seen to be contentious.

Nevertheless a grouping mechanism is required if any comparison of descriptions is to be performed. An approach based upon the idea of "natural affinity" described above is being adopted and tested. Under this scheme a hierarchical arrangement of *state groups* will be created as an aggregation of the states in their subordinate groups. The hierarchy is rooted at the special state group called "property", and which effectively contains all states in all state groups. The number of levels in the state group hierarchy will be unlimited. Within each group the members are considered to be possible non-exclusive alternatives, but some groups may contain only one member. It is intended that such an hierarchical arrangement of groups of states will be minimally contentious and easy to navigate.

Ascending the hierarchy of state groups represents an increase in the degree generalisation of concept. All the members of child groups are specialisations of their parent and the parent group represents a more general "property" than its constituent children. A consequence of this model is that when making a comparison between qualitative descriptions, knowledge of the level of generalization at which the states were grouped will inform the comparison process. An example of how this might work is shown in Fig. 5. Such a scheme constrains state term usage in order to allow description comparison, but it is flexible in allowing the user to select at which level of generalisation they wish to work.

Developing the demonstration ontology. — During its development, there was much discussion about the appropriate level and scope of the ontology. On one hand, it would be possible to attempt to create an ontology that allows anybody to create and compare a description of any organism. On the other hand an ontology could be constructed with a specific user group in mind that would satisfy their data capture and information exchange requirements. Problems are evident with either approach. To build an all encompassing ontology that could capture all things that anyone would want to record about any organism would be a project of such magnitude that it would be impossible to achieve in any reasonable time span. Moreover, imposition of such a scheme upon such a broad domain would undoubtedly be contentious. On the other hand allowing individual interest groups to each develop their own ontologies in a more bottom-up fashion is problematic because data exchange is only possible between groups that subscribe to the same ontology, and little will have been gained with respect to the challenge of enabling meaningful information exchange across the whole of taxonomy. A middle way is therefore required in which the intended user group is broad enough for the ontology to have an effective influence on information exchange and yet is not so broad that capturing the concepts involved becomes impracticable. The most pragmatic approach to this problem came out of the concept of biologically "natural groups". These are defined as being groups in which the tendency to compare entities within the group is greater than the tendency to compare entities outside the group. This is usually determined by the frequency of occurrence of common structures on individuals in each group. Examples of natural groups as we see them would be: bacteria, fish, birds, amphibians, fungi, gymnosperms and angiosperms. For the purposes of constructing a prototype ontology the natural group, angiosperms, was selected for this study; it was of appropriate size and is of interest to a sufficiently broad community.

In establishing a vocabulary, rather than develop a de-novo set of definitions, a literature-based approach was adopted and terms were compiled from standard botanical references. The term list was then edited to

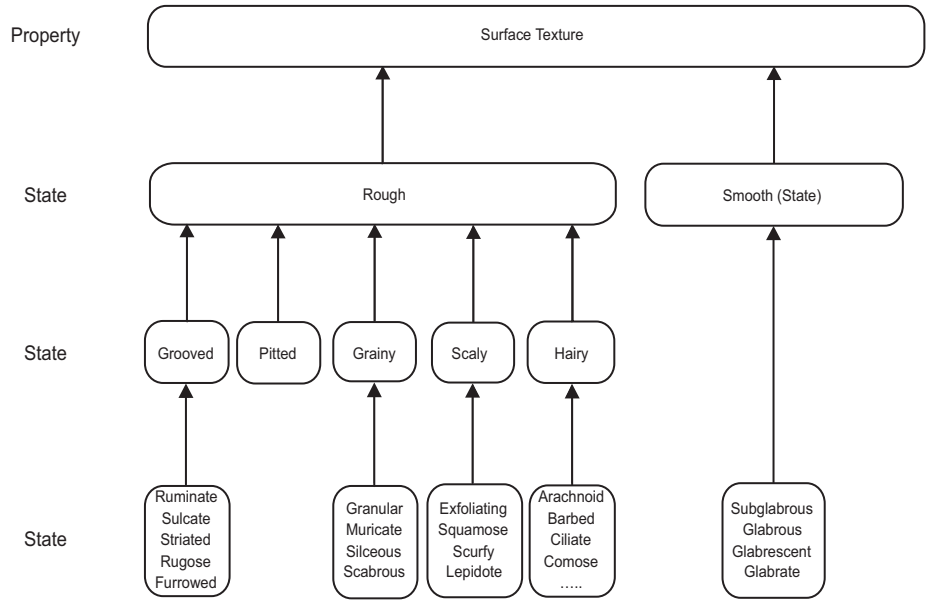

Fig. 5. A hierarchical arrangement of properties and states in which descending the hierarchy represents an increase in specialisation. When comparing two descriptions of a specimen based on the same ontology, the descriptions have to be compared at the least level of specialisation common to both descriptions. For example, using the arrangement above, consider the situation where one description states that a structure has the surface texture rough, and the other describes the same structure having a grooved surface texture. The only conclusion that can be drawn from this is that both descriptions consider the structure to have a rough surface texture.

check for redundancy. In instances where multiple records of the same term from different authors were found, a preferred definition was flagged for use. The clearest and most concise definitions were chosen in each case.

USING THE ONTOLOGY-BUILDING PRO-FORMAS AND CREATING DESCRIPTIONS

The ontology acts as a pool of defined terms that are available for use in building descriptions, together with relationships between the terms which constrain their use. However, for most projects, the ontology will be too broad in scope and a scoring interface will typically only require a small subset of the available terms. Allowing individual projects to define a restricted view of the ontology in which only a subset of terms are visible, solves this problem. These subsets can be referred to during the description-building process by means of a *pro-forma*. In essence, constructing a pro-forma defines the subset of terms to be used in a particular project.

As well as providing a mechanism for scoring the states of structures, specifying the pro-forma view of the ontology also has an important role in mediating the expression of the various forms of variation identified earlier. The most complex aspect of pro-forma specification is the construction of compound entities required to express variation in the form of a structure within a specimen. This is achieved by an extension of the kind-of relationship described for the structural ontology. Within this process the user is required to make multiple representations of the same structure with differing combinations of states; a process we have termed cloning. For example, in a set of specimens it may be recognised that there are two "kinds" of leaves—green hairy leaves and yellow glabrous leaves. This will require that two clones of leaf be constructed in the pro-forma; one will have the colour scored as green and the other scored as yellow. The clone with colour scored as green will also be scored as hairy. The other clone will be scored as being glabrous. Once this has been achieved, scoring is then simply a matter of recording the presence of the appropriate structural forms for each specimen described. Once flagged as being present, additional scores can be associated with the selected clone. In effect, cloning allows descriptions of individual structures to be produced, which can then be joined together to describe an entire specimen.

Descriptions of separate specimens are distinguished by using the pro-forma as a template and generating a separate score sheet for each specimen. Descriptions recorded using score sheets based on the same pro-forma are automatically comparable. However, if an hierarchical model of state groups has been employed, some user interaction may be required in order to determine areas of commonality between descriptions when they are based on different pro-formas. Nevertheless, all descriptions based on pro-formas derived from the same parent ontology are guaranteed to possess a degree of compatibility.

DISCUSSION AND CONCLUSIONS

The Prometheus description model was conceived as a natural extension of the Prometheus taxonomic model (Pullan & al., 2000). Just as in the development of the Prometheus taxonomic model, one of the main aims of the current project was to produce a model of taxonomic descriptions that moves away from the taxon-oriented approach prevalent in existing description models, towards a more objective specimen-oriented approach. This change in emphasis has been achieved by building a model based upon a study of taxonomic working practices rather than the usual approach of building a model based upon taxonomic outputs (Cannon & McDonald, 2001). As a result, it is believed that this is the first model that fully recognises and models the layers of abstraction required to formulate a taxon description. At the specimen level, individual observations about individual structures can be concretely recorded. These can then be abstracted into a summary of individual structure scores from a single specimen. A final phase of abstraction can then be performed that summarises scores from a group of specimens to create a taxon description, although even when this level of abstraction is reached the link back to specimens can be maintained by modelling a taxon description as one or more virtual specimens.

There are a number of possible advantages to this approach. Firstly, it is hoped that this method of design will result in systems that are more readily incorporated into the everyday work of taxonomists and therefore are more likely to be widely adopted by the taxonomic community. Secondly, because the model presented here will allow data from all stages of the description process to be recorded, the data sets generated using this model should provide a more enduring legacy than systems that only allow data to be recorded from the most abstract level of the process, i.e., the final taxon description. Undoubtedly, it is far easier to understand highly subjective taxon descriptions if the raw data upon which they are based are also readily available. Moreover, the raw data have the most potential for reuse; it is possible to create numerous descriptions at various levels of abstraction that can be tailored to suit particular purposes from a single set of raw data. Nevertheless the Prometheus descrip-

tion model will allow a user simply to record abstract taxon descriptions without reference to the specimens from which these data were obtained. However, if users wish to maximize both the consistency of interpretation of their descriptions by reducing the level of subjectivity within them, and maximise the potential for the reuse of their data, they should be encouraged to capture concrete specimen-level data.

Although we have adopted a new approach to modelling descriptive data, there are nevertheless many commonalities between the Prometheus description model and existing models. This has occurred for two reasons. Firstly, there was a desire to build upon existing work and draw the best elements from current models into this one. Secondly, because all the models are operating within a common domain, it is inevitable that there will be many similarities between them. It is believed that the model presented here adequately handles all the features provided by existing models as well as incorporating novel features. The remainder of this discussion will therefore focus on the new features of the model rather than attempting to provide comprehensive comparison of all the features common to existing models.

As discussed earlier, when developing a data model for the representation of descriptive data, one of the primary concerns was that of ensuring consistent parsing of descriptive information. As a result the Prometheus descriptive data model was based upon the design proposed by Diederich (1997), which was seen to have several key advantages over the DELTA format. Consequently there is the greatest number of similarities between the Prometheus description model and that of Diederich. However, the Prometheus description model differs from that of Diederich in a number of significant ways.

1. Diederich's model does not provide any mechanism for the definition of terms. It appears that the primary focus of the Diederich model was limited to ensuring the consistent representation and parsing of descriptive data rather than the consistent interpretation of the data by either humans or computers. Paradoxically, although there are problems with the consistent parsing of DELTA data, DELTA does provide means for including definitions of characters and states. However, definitions are not compulsory and tend only to be provided when DELTA sets are intended to be used within identification systems where the need for consistent identification of structures and states is obvious (Pankhurst, 1991). In a sense the Prometheus description model includes a hybrid of the approaches adopted by Diederich and DELTA.

2. The concept of the "basic property" is central to the Diederich model. However, in developing the Prometheus ontology it was found that the relatively flat model of property adopted by Diederich was incapable of satisfactorily handling the vagaries of qualitative properties. Requiring states to be grouped into properties actually requires a classification of states to be performed, and just as with a taxonomic classification there are many layers of abstraction present in the concept of property. Adopting a hierarchical approach to property will allow this multi-layered classification of property to be more accurately modelled. Using the Prometheus description model a single state may belong to numerous properties at different levels of abstraction. We believe that this produces a more natural and less arbitrary classification of property than was possible with the flat model as evidenced by the fact that even between publications relating to the Diederich model from the same year, there are significant variations in the semantic categorisation of the basic properties. For example in Diederich & al. (1997) the basic property "presence" is placed in the semantic category "quantity" where as in an earlier paper (Diederich, 1997) it was placed in "appearance". Furthermore in using a more hierarchical approach we have not found it necessary to create a catch-all property, such as the property "kind" used by Diederich (1997), in order to handle awkward states that otherwise defy classification in the flatter model. In contrast, the concept of property is almost completely absent from DELTA, although in some cases it may be inferred from the character comments often appended to a DELTA character. For example, it is common to see states such as "leaf <colour>" appearing in a DELTA character set—the text in angle brackets often only being included as a means of generating understandable natural language descriptions from the encoded data rather than being perceived as contributing to the overall semantics of the dataset. Failure to consistently group states into properties can also lead to problems in which inappropriate combinations of state are included in a single character. Although these issues do not create problems within individual DELTA datasets, they make it difficult to ensure comparability between datasets.

3. The use of pro-formas derived from the base ontology is a unique feature of the Prometheus approach (Cannon & al., 2004). It provides the user with the flexibility to work with the ontology in a manner that suits them whilst retaining control over the data structure and term usage. Because of this flexibility a range of possible working styles can be envisaged ranging from a DELTA-style approach in which pro-forma level "kind-of" relationships are used to create complex characters based on the forms of structure exhibited by the specimens under study, to a more purist approach, in which each specimen is individually scored using basic states, allowing subsequent analysis of the raw data in order to extract the structural forms that have been observed. This approach

has the advantage of not introducing additional bias into the scoring process, which may be introduced when complex characters are created on an a priori basis.

4. Although the important semantic difference between recording concrete and abstract data has long been recognised (e.g., Thiele, 1993), to our knowledge the Prometheus description model is currently the only model in which this difference is incorporated. Neither the DELTA model nor that of Diederich is capable of handling concrete data. For example, neither could represent multiple measurements of the same structure from a single specimen.

5. The Prometheus description model is the first taxonomic description system that incorporates an explicitly expressed ontology. It could be argued that both DELTA and the Diederich model do incorporate some form of ontology by including mechanisms for elaborating characters and creating relationships between them. Explicit relationships between characters are expressed in DELTA using character dependencies. A similar but more fine-grained concept of state-based dependencies is employed in the Diederich model to the same effect. However, the fact that the majority of the relationships within these models are not explicitly expressed means that they lack the semantic clarity that a true ontology is capable of delivering.

The Prometheus ontology functions as the framework from which taxon descriptions can be created, providing a standard set of defined terms and controlling valid combinations of terms from different categories, (i.e., states, properties and structures). Inputing data using the ontology means that an author has subscribed to a particular ontology, ensuring semantic consistency between datasets.

The demonstration ontology was, as far as possible, constructed in an impartial manner. Nevertheless, elements of the ontology are bound to reflect the personal preferences of the builder. Hopefully the pro-forma based approach will ameliorate many of these problems, yet it is still to be expected that individual taxonomists may want to construct their own ontologies. In order to minimise data heterogeneity the construction of alternative ontologies is to be discouraged, and we expect that pragmatic reasons may achieve this. Although a pro-forma can be built in one day, the ontology upon which it is based takes considerably longer to construct requiring the collection and collation and classification of terms and definitions. Assuming that the ontology-based approach is widely adopted and alternative ontologies do eventually appear, it is hoped that this will occur in an open manner in which the most frequently used ontologies will become adopted as community standards and the less frequently used ontologies will fall by the wayside. The use of multiple alternate ontologies will, however, require a manual mapping of the terms and definitions between ontologies in order to allow meaningful data integration.

Detailed botanical ontologies are being developed by other groups, particularly the Plant Ontology Consortium (POC, 2002). POC are constructing highly detailed anatomical and "trait" ontologies, initially for three scientifically well-characterized model species (*Oryza*, *Zea* and *Arabidopsis*). However, the POC ontologies are intended to define genetically-based traits and specific mutations rather than more generally applicable taxonomic characters. In many respects the level of detail specified in these ontologies goes beyond that required for taxonomic description, and being species-specific the ontologies are inappropriate for taxonomy.

There is a similar representation of structures according to POC's anatomical ontologies and the Prometheus ontology, with POC also recognizing the importance of defined terms and relating these hierarchically using a central part-of relationship. The POC ontologies, however, also incorporate an "instance-of" relationship, which is somewhat analogous the kind-of relationships for structures found in the Prometheus ontology. However, structures related with the "instance-of" relationship in the POC model can fully participate in structure hierarchies. This has not been included in the Prometheus description model as it was found that incorporation of such relationships into the structural hierarchy made the ontology overly complex and difficult to navigate. Furthermore, it was found that such relationships start to become meaningless when considered over a large taxonomic range. POC ontologies also include an additional "derived-from" relationship, which expresses developmental information currently not represented in the Prometheus description model.

In summary, it is believed that the specialization of the Prometheus description ontology into individual pro-forma sub-ontologies is a novel means for facilitating the collection of compatible description data. It is also believed that capturing the rich semantic content expressed in the ontology (for example, the ontologically defined context of a structure via its path) allows not only efficient and consistent knowledge sharing and re-use but will also allow rigorous representation and analysis of taxonomic concepts.

Development of a novel description methodology and data model can only be validated by providing tools to create, explore and use defined ontologies for specimen description, allowing taxonomists to record descriptions compliant with this constrained format. An angiosperm ontology for the description of one taxonomic dataset has been developed and is currently being extended for the description of further test datasets. Providing tools which allow data entry using only a con-

trolled defined terminology enforces semantic homogeneity, and will aid future integration of any database created using these tools.

ACKNOWLEDGEMENTS

The Prometheus II project was funded by the BBSRC/ESPRC grants BIO14354 and 95/BIO14353.

LITERATURE CITED

Berendsohn, W. G. 1995. The concept of "potential taxa" in databases. *Taxon* 22: 207–212.

Berendsohn, W. G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283–309.

Berendsohn, W. G., Döring, M., Geoffroy, M., Glück, K., Güntsch, A., Hahn, A., Jahn, R., Kusper, W.-H., Li, J., Röpert, D. & Specht, F. 2003. *MoReTax: Handling Factual Information Linked to Taxonomic Concepts in Biology*. Bundesamt für Naturschutz, Bonn. [*Schriftenreihe Vegetationskunde*: 39.]

Cannon, A., Kennedy, J., Paterson, T. & Watson, M. 2004. Ontology-driven automated generation of data entry interfaces to Databases. Pp. 150–164 in: Williams, M. H. & MacKinnon, L. M. (eds.), *Key Technologies for Data Management, 21st British National Conference on Databases, BNCOD 21, Edinburgh, U.K., July 7–9, 2004, Proceedings. Lecture Notes in Computer Science 3112*. Springer, New York.

Cannon, A. & McDonald, S. M. 2001. *Prometheus II — Qualitative Research Case Study: Capturing and Relating Character Concepts in Plant Taxonomy.* www.prometheusdb.org/resources.html; Prometheus: www.prometheusdb.org

Colless, D. H. 1985. On "character" and related terms. *Syst. Zool.* 34: 229–233.

CSIRO 2001. *BIOLINK, CSIRO*, Australia. URL: http://www.biolink.csiro.au/index.html

Dallwitz, M. J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–46.

Dallwitz, M. J. & Paine, T. A. 2004. *Definition of the Delta Format.* URL: http://delta-intkey.com/www/standard.htm

Davis, P. H. & Heywood, V. H. 1963. *Principles of Angiosperm Taxonomy.* Oliver and Boyd, Edinburgh.

Diederich, J. 1997. Basic properties for biological databases: character development and support. *Math. Computer Model.* 25: 109–127.

Diederich, J., Fortuner, R. & Milton, J. 1997. Construction and integration of large character sets for nematode morpho-anatomical data. *Fund. Appl. Nematology* 20: 409–424.

Diederich, J., Fortuner, R. & Milton, J. 1998. A general structure for biological databases. Pp. 47–58 in: Bridge, P., Jeffries, P., Morse, D. R. & Scott, P. R. (eds.), *Information Technology, Plant Pathology and Biodiversity.* CAB International, Wallingford, U.K.

Diederich, J., Fortuner, R. & Milton, J. 2000. A uniform representation for the plan of organization of nematodes of the order Tylenchida. *Nematology* 2: 805–822.

Genesereth, M. R. & Nilsson, N. J. 1987. *Logical Foundations of Artificial Intelligence.* Morgan Kaufmann Publishers, San Mateo, California.

Gradstein, S. R., Sauer, M., Braun, W., Koperski, M. & Ludwig, G. 2001. TaxLink, a program for computer-assisted documentation of different circumscriptions of biological taxa. *Taxon* 50: 1075–1084.

Gruber, T. R. 1993. *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.

Lawrence, G. H. M. 1951. *Taxonomy of Vascular Plants.* Macmillan Company, New York.

Maddison, D. R., Swofford, D. L. & Maddison, W. P. 1997. NEXUS: an extensible file format for systematic information. *Syst. Biol.* 46: 590–621.

Pankhurst, R. J. 1991. *Practical Taxonomic Computing.* Cambridge Univ. Press, Cambridge.

Paterson, T., Kennedy, J. B., Pullan, M. R., Cannon, A., Armstrong, K., Watson, M. F., Raguenaud, C., McDonald, S. & Russell, G. 2004. A universal character model and ontology of defined terms for taxonomic description. *Data Integration in the Life Sciences; Lecture Notes in Bioinformatics* 2994: 63–78.

Plant Ontology Consortium (POC). 2002. Conference review: the plant ontology consortium and plant ontologies. *Comparative and Functional Genomics* 3: 137–142; on-line publication: http://www3.interscience.wiley.com/cgi-bin/fulltext/91016047/HTMLSTART

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.

Shetler, S. G. 1975. A generalized descriptive data bank as a basis for computer-assisted identification. Pp. 197–235 in: Pankhurst, R. J. (ed.), *Biological Identification with Computers*. Academic Press, London.

Sokal, R. R. & Sneath, P. H. A. 1963. *Principles of Numerical Taxonomy*. W. H. Freeman and Co., San Francisco.

TDWG-SDD. 2003. *TDWG Working Group: Structure of Descriptive Data.* URL: http://160.45.63.11/Projects/TDWG-SDD/

Thiele, K. 1993. The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* 9: 275–304.