A Methodology for Composing Well-Defined Character Descriptions.

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

[1]*School of Computing, Napier University, Edinburgh, EH10 5DT, U.K. {j.kennedy, t.paterson, g.russell, a.cannon}@napier.ac.uk*
[2]*Royal Botanic Garden, Edinburgh, EH3 5LR, U.K. {m.pullan, k.armstrong, m.watson}@rbge.org.uk*

Taxonomy has been described as "the science of documenting biodiversity", which involves collecting, naming, describing, identifying and classifying specimens of organisms (Keogh, 1995). Descriptions are the fundamental information units used in the process of constructing classifications and communicating taxonomic concepts. The quality of stored description data is limited by the lack of a formal model and methodology for composing specimen descriptions, and by the absence of an agreed defined terminology. This impedes the communication, interpretation and reuse of original descriptions. This paper describes a novel approach to composing and recording taxonomic descriptions of botanical specimens. An underlying model for creating character descriptions is presented together with a process for creating an ontology of defined terms, which will be used to compose these description elements. It is hoped that these developments will facilitate the unambiguous interpretation of descriptions and enhance the taxonomic process.

KEYWORDS.- character, taxonomy, description, database, model, ontology

INTRODUCTION

The Problems Inherent in Current Taxonomic Descriptions.- The general procedure for a taxonomic revision, modelled in Fig. 1, results in the description of individual plant specimens in terms of identifiable 'characters'[1] (typically morphological, but also physiological, molecular etc.) and the sorting of these specimens into taxonomic groups based upon character similarities. A specimen description is therefore composed of 'character data'. Higher-level descriptions for taxa (species, family, genus etc.) can then be composed by summation and abstraction of the shared characteristics of member specimens. Taxonomic practice, however, is highly individualistic and there is no common agreed procedure or format for the creation and representation of these descriptions. Furthermore, the detail and format of descriptions varies according to their usage (for example in monographs, geographical flora or field identification guides).

As a consequence of the way in which descriptions have historically been collected and recorded, the interpretation, integration and re-use of character data may be problematic. Typically neither the 'character' concepts, nor the terms used to describe these 'characters' have been rigorously or consistently defined, and subsequent interpretation of descriptions may therefore be ambiguous, and descriptions from disparate sources cannot be compared or reused with any confidence of accuracy. Indeed,

[1] Problems with the definition of 'character' are explored in the following section, 'character' here means in general any statement on a feature of an organism

descriptions may have no formal structure being simply a general textual description, without decomposition into separate character states. Historical descriptions are therefore of limited value to a modern taxonomist, who must often re-examine a specimen in order to interpret the original description or redescribe the specimen themselves. Within descriptions it may not even be possible to distinguish direct observations from extrapolations (Watson, 1971). Nor are the original thought processes and methodologies underlying the summarization of description data available for evaluation. Furthermore, much of the original descriptive work may have been lost or discarded, as common practice has been to record only the 'characters' of interest for a given study. It is often impossible for subsequent readers of a description to distinguish between absence of a feature in the material being described and mere failure to seek or comment on it. More detailed original descriptions, typically recorded onto paper 'proformas', are not composed in a format suitable for reuse, and are often discarded or forgotten, causing a significant loss of potentially useful information (Diederich & al., 2000).

The Concept of 'Character':- Whilst 'character' data are the basic building blocks of descriptions, there is little consensus on what the term ‘character’ actually means. This leads to problems interpreting taxonomic descriptions. A ‘character’ might be defined in general terms as ‘a statement on a feature of the organism’ although many different specific definitions for 'character' have been proposed (Davis & al., 1963; Blackwelder, 1967; Wiley, 1981; Colless, 1985; Stuessy, 1990; Fristup, 1992; Bailey, 1999). Some taxonomists consider that to be a 'character', a feature must be useful for differentiating between specimens.

During the taxonomic process (Fig. 1) a taxonomist seeks to identify comparable structures amongst specimens, where variation between those structures allows the phenetic classification of specimens. The 'character' concept is derived by partitioning observed variation into characters and character states i.e. the combination of a structure, the aspect of the structure being described and its possible states (Wilkinson, 1995; Cannon & McDonald, 2001). For example, the 'character' *leaf shape* may be recognized and in a group of specimens the 'states' *obovate*, *ovate* and *lanceolate* observed. The description of that group would then read ‘leaves obovate, ovate or lanceolate’.

Given the wide number of uses for character data across different branches of taxonomy, it is not possible to develop a universal definition for the term ‘character’. In fact 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) distinguished two contrasting usages of the term ‘character’ in published descriptions. In one usage the character is a general concept, e.g. *leaf shape*, which is separated from the score, e.g. the state *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 can be a value or a ‘character state’. In the other usage the structure and the score are combined, e.g. 'leaves obovate', in which case the property (*shape*) is implicit. In order to allow taxonomists to be explicit about every aspect of a character statement, Diederich proposed the decomposition of 'character' into the separate elements structure, property and score, where score can be a value or a state. This useful definition forms the basis of the Prometheus model to represent characters.

A serious problem with 'characters' in practical terms is that the selection and definition of character data is ambiguous. Whereas an individual taxonomist's concept of what constitutes a particular character may be adequate for their taxonomic purpose, it is necessary for them to be able to decompose and define this concept in order to adequately communicate the basis of their classifications (Davis, 1963). Whilst the selection of characters for taxonomic descriptions is of critical importance for the validity of the work, it is clearly also necessary to ensure that the characters chosen are unambiguous and easy to interpret in order to allow other researchers to understand and extend or reuse the original work.

In summary, it is apparent that the lack of an unambiguous, universal model and methodology for composing character descriptions impedes the communication and reuse of taxonomic specimen descriptions.

The Aims of the Prometheus II Project:- With modern information storage systems it should be possible to store and reuse original and detailed descriptive data in a more consistent and accurate manner. However, the reuse of data forces constraints on the manner in which data is collected, described and stored. Several authors have suggested that there should be a standard approach to taxonomic descriptions (Watson, 1971; Deiderich & al., 1997; Silvarajan, 1991; TDWG, 2000; Hagedorn, 2002). 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.

Current electronic description formats used by taxonomists (e.g. DELTA, Lucid and NEXUS, Dallwitz 1980; Dallwitz & Paine 1986; CBIT 2003; Madison & al., 1997) have been designed to allow flexibility, and therefore allow taxonomists to record character states however they choose (including combining of more than one observable characteristic within one character). These formats give very limited support for determining defined characters and for representing character hierarchies. Of the electronic description formats, DELTA is the most fully featured for use in taxonomic revisions, Lucid being designed for creating keys and NEXUS being a general extensible file format for any systematic information. Flexibility, however, implies a lack of rigour in the use of character concepts and the absence of an agreed terminology means that descriptions are generally only consistent within a single data set which therefore precludes communication and reusability, and comparisons between datasets.

It is the aim of the Prometheus II project (Prometheus 2003a) to improve current taxonomic working practice, by exploring techniques and providing tools for more rigorous data collection and recording. Prometheus proposes the use of a formal character data model together with a defined ontology of descriptive terms in order to represent 'character' descriptions of specimens in a consistent manner. By encouraging the use and reuse of well-defined terms and concepts, descriptions should become less ambiguous and more compatible, providing the opportunity for useful comparisons between descriptions.

METHODOLOGY AND MODELLING

The Prometheus II Description Process:- As summarized in Fig. 2 and elaborated in subsequent sections, the overall approach of the Prometheus II project involves:

1. Compilation of lists of defined terms for the structures, states and properties to be used in specimen descriptions, drawn from botanical glossaries etc.
2. Creation of a defined term ontology for a circumscribed taxonomic domain (here angiosperms). As described in following sections this entails:
    a. The selection and import of all the defined terms necessary to describe angiosperm specimens, i.e. terms with their preferred definitions.
    b. The creation of a hierarchy of structural terms, representing all the possible structural relationships in a given aniosperm specimen.
    c. The designation of Structural Types, Regions and Generic Structures.
    d. The grouping of state terms into sets of terms which are used to describe alternative aspects of the same feature.
    e. The linking of these state sets to the structures that they are used to describe.
    f. The export of this ontology in machine-readable form (e.g. XML).
3. The computer assisted creation of electronic proformas for the description of a group of specimens. Guided by the angiosperm ontology the user will specify which structural elements they wish to describe, and for each of these elements, which descriptive concepts they wish to describe or measure. The descriptive concept ('character') is defined by selecting the set of state terms available for description or by selecting an appropriate quantitative measure.
4. The scoring of real specimens using a proforma template to save character descriptions in a consistent structured format within the Prometheus Database.

.

The Use of Defined Terms in Taxonomic Descriptions:- The Prometheus approach emphasizes the definition of terms. Terms are simply the bare words found in a standard description, for example '*leaf*' or '*pubescent*'. Definitions might simply be textual (drawn from standard botanical glossaries) but could be expanded to include digital images and representative examples. By constraining descriptions to be composed using only defined terms, explicit statements about features can be unambiguously and reproducibly interpreted.

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. Ideally a single definition of a term will be used in Prometheus descriptions, but as a term will always be used in context of its definition it is possible to allow for homonyms (i.e. words with alternative definitions). Similarly it will be possible to define synonymous terms which share a preexisting definition of another term. This would allow taxonomists to use their preferred term, whilst explicitly declaring a shared definition between synonyms. Such a mechanism should make it explicit that there is no subtle difference in meaning between alternative synonymous terms. The creation of ontologies restricting term usage and defining relationships between terms will further add to the contextual definition of terms.

Prometheus will provide a variety of different classes of 'defined terms', necessary to capture the different elements of the data model. The central types of term are structures (e.g. *leaf*), quantitative properties (e.g. *length*), qualitative states (e.g. *pubescent*), units (e.g. *cm*) and a variety of types of

modifiers necessary to refine the character descriptions (e.g. *rarely*, *before*, *above*). The relationships between these term types are discussed further in the context of ontology creation (see below).

The Prometheus II Character Description Model.- Prometheus proposes a new data model for character descriptions which seeks to address the problems described in the introductory sections. The model is intended mainly for recording the information collected for new descriptions, but might also be used to record an interpretation of an existing description.

The Prometheus model developes Diederich's definition of 'character' (Diederich & al., 1997) which he decomposed into structure, property and score enabling taxonomists to be explicit about every aspect of a character statement. We hold this definition to be true for quantitatively measured characters, but observe that for characters described by qualitative state terms it is often difficult to extract the precise property that is being described by the state. In many cases a recognizable 'property' is too general to be useful (e.g. *amplexicaul*, *appressed* and *alternate* could all be said to describe different aspects of the *arrangement* of leaves) in other cases the property is a complex combination of structure and state with no simple property label (e.g. *didynamous*: with 4 stamens, in 2 pairs of 2 different lengths). Therefore our model distinguishes between quantitative characters (combinations of structure, property and value) and qualitative characters (combinations of structure and state).

A statement composed using defined terms, and describing a particular feature (character) of an organism in either form (quantitative or qualitative) is referred to as a Description Element. A Specimen Description will be composed of the set of description elements describing features of that specimen.

The model requires two kinds of description element: qualitative description elements to describe a structure in terms of a defined qualitative state (e.g. *obovate*), and quantitative description elements to describe a feature in terms of a value and a defined unit. (e.g. *length: 5 cm*). Arguably it should be possible to describe all physical data quantitatively, and Prometheus would encourage quantitative description where practicable. This would permit the direct comparability of description elements' data. However, often this is neither reasonable nor useful and taxonomists tend to assign qualitative states by breaking up continuous quantitative variation and complex character properties into more easily handled discrete states. 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 (Hickey, 1973).

Quantitative Description Elements.- In order to record the statement '*leaf length 5 cm*' a quantitative description element is composed using a defined structure (*leaf*), an explicit defined property (*length*), a value (*5*) and the appropriate defined unit (*cm*). The property term is associated with one or more values (to express ranges), which are individual numbers (e.g. *5*) and must be associated with a defined unit. For quantitative statements that do not have units, for example number of petals, '*count*' is defined as a unit. Clearly there is a finite list of defined quantitative properties which can be described by these elements, which might minimally consist of the set {*Angle, Density, Diameter, Height, Length, Number, Width*} and be expanded to allow further defined quantitative properties as needed (e.g. *Colour*, as defined by RedGreenBlue or RHS[2] value etc.).

[2] Royal Horticultural Society Colour Chart

Qualitative Description Elements.- In order to correctly record a statement such as '*leaves obovate*' a qualitative description element is composed with a defined structure (*leaf*) and a defined qualitative state (*obovate*). Note that no explicit property is specified for qualitative scores. Attempts were made to categorize state terms by the type of qualitative property described (e.g. {*Arrangement, Colour, Dehiscence, Development, Form, Fusion, Habit, Life Cycle, Orientation, Persistence, Presence, Sex, Shape, Smell, Symmetry, Texture*}). However, for many state terms such divisions prove arbitrary or contentious, as the state can reflect aspects of several properties. Taxonomists view the actual qualitative property of a 'character' as a gestalt of property in context of the structure being described. This observation was used to group state terms into sets in the ontology (see below). These usage groups therefore circumscribe the implicit 'property' that is being described in the qualitative description element.

Ontologies for Taxonomic Descriptions.- Collecting rigorously consistent data for storage in a database is dependant upon the specification of the underlying data model, and ontologies are increasingly used for constraining and defining the capture of meaningful data. An ontology is a formal specification for structuring knowledge about a domain (i.e. botanical description in this case). Ontologies can range from the highly-informal (specified purely in natural language) to rigorously-formal (in a language with formal semantics, theorems and proofs). An ontology can comprise of concepts, relations between concepts, functions defined on concepts, axioms and instances. A formal ontology would be overly complex and restrictive for defining a description vocabulary, but a restricted and structured form of natural language for use in taxonomy could be created as a semi-formal ontology. The ontology will, however, have a degree of sophistication as not only will terms be defined and restricted (as concepts) but there will be a variety of relationships between concepts beyond simple subsumption ('Is-A'), (e.g. 'Part-Of', Type-Of').

The Prometheus Ontology will assist in the organization, reuse and communication of knowledge within its domain, but will not support computational inference beyond establishing compatibility for comparisons because the actual definitions belonging to terms are purely textual.

The relationships and classes within the Prometheus Ontology are shown diagrammatically in Fig. 3. There are two primary classes of concepts: *defined structure terms* and *defined state terms* as subclasses of *Defined Terms*.

Defined Structure Terms.- Structure Terms can be related to form a structural hierarchy, using multiple 'Part-Of' relationships, to form a network hierarchy where each structure can be related as a (possible) subpart of multiple superstructures (e.g. **B**, **C** and **D** are parts of **A**; **B** is a also a part of **E**). The 'Part-Of' relations should express every possible structure/substructure relationship in the taxonomic domain of the ontology. For a given real specimen, the 'Part-Of' relationship is optional (in that structure **A** *might have* subparts **B**, **C** and **D**, and structure **E** *might have* subpart **B**). This hierarchy is most simply viewed as a tree hierarchy with multiple possible instances of a given structure (**B** can be a part of both **A** and **E**, hence in the tree there are two instances of **B** together with all of **B**'s potential subparts). The entire 'Part-Of' tree is rooted on the structure 'entire plant or specimen'. An illustration of the 'Part-Of' relationship with actual structures is given in the Results section (Fig. 4).

Within the ontology we have subclassed 'Generic Structures' (e.g. *hairs*) and 'Regions' (e.g. *base*) as specialized types of Structure Terms which are not included in the 'Part-Of' hierarchy as they would occur repeatedly as parts of multiple, or even potentially all, structures. These terms will be added to the structure hierarchy at the stage of proforma creation, as part of the process of delimiting an actual Proforma Ontology for use in recording a set of descriptions. However, conceptually Generic Structures and Regions are equivalent to Structure Terms when used in descriptions.

The creation of the structure hierarchy allows any structure to be referred to in the context of its superstructures. Each node in the structure tree is uniquely identifiable by virtue of its 'path' (e.g. **B** that is part of **A** [path **A.B**], opposed to **B** that is part of **E** [path **E.B**]). This allows descriptions to be composed using structures that are unambiguously defined both by definition and in terms of their relationships to other structures. This extends our character description model, in that a Description Element can be composed of *structure-in-context* (using a path defined in the ontology), property and score, rather than simply a defined structure, property and score.

A further specialized subclass of Structure Terms is *Types*. Structure Terms can be defined as a 'Type-Of' another structure if they are examples of the supertype that always have a number of descriptive states true for each instance of that supertype structure. For example, a *berry* is a 'Type-Of' *fruit*: it is a fruit that is always fleshy, indehiscent and has seeds submerged in pulp. Types reflect an awkward apparent blurring between states and structures when describing specimens, in that a berry is clearly a structure in itself, but it is also a collection of states for a particular structure (fruit). Stuctures that are defined as 'Types-Of' other structures are excluded from participation in structural hierarchy; they can only be used in a description as an attribute of their supertype, so that in a description a fruit can be recorded as being of type berry. At description time taxonomists often informally conceptualize structures with collections of states using 'types of' relationships, confusing the boundary between description instances and ontology concepts (for example, referring to green hairy leaves as a 'type of' leaf). In order to prevent an explosion in these type definitions we have defined a set of rules for types:

1. Structure Terms can be defined as a 'Type-Of' another Structure Term.
2. This excludes a Type from participation in the structure/substructure hierarchy.
3. A Type must share the defining features of the parent, but have several (i.e nore than one) states scorable that are always true for that Type of the Structure.
4. A set of Types should be exclusive (an instance of a *fruit* can be either a *berry* or a *nut*, not both).
5. Types are identified by a single noun/name (*berry*) rather than an adjective plus supertype name (*male flower*).
6. There are no hierarchies of Type (representing *nutlet* as a type of *nut,* which is a type of *fruit* is overly complex; *nutlet* and *nut* are both types of *fruit*).

Defined State Terms.- Unlike Structure Terms, State Terms have not been divided into subclasses in the ontology, nor are dependencies between states represented. This expert knowledge will be contributed by the individual taxonomist who is creating proformas or scoring specimens. However, states have been organized into sets (called State Groups) of terms which tend to be used to describe alternative aspects of the same concept (such as 'Texture' or 'Leaf Outline'). These State Groups are linked to

the structures in the ontology that they can be used to describe. For some of the State Groups (for example 'Texture') it is not sensible to restrict usage to a set of structures as the states can be applicable to an extensive range of structures (or indeed potentially all structures). State Groups are given a name reflecting the descriptive concept (e.g. 'Texture') some names include an indication of the usage context ('Leaf Outline') and in other cases it is difficult to express the concept with a simple name. The set of states, and possibly the structures to which they are applicable, can be considered to circumscribe the 'character property' for which the states are the possible scores.

Ideally these State Groups would be composed of the exclusive alternative states for the descriptive concept of an actual structure. However, the extent of such exclusive state groups proved difficult to define and it was not possible to be certain that a given structure could never be described by more than one state in a group. All state groups are therefore considered as potentially 'multistate' for a particular instance (i.e. a description element is allowed two or more states scored). Furthermore, terms can be members of more than one State Group (although this might more properly be represented by creating homonyms with different definitions for inclusion in each group).

Because taxonomists tend to think about sets of 'similar' or related states there was an attempt to decompose State Groups into subgroups describing similar states. This proved highly problematic as some subgroups aggregated terms with 'similar' definitions and others aggregated states that seemed to describe a similar aspect of a state/property, but with very different definitions. (These latter subgroups might more properly form State Groups in their own right). Although taxonomists liked these subgroups they have been given no ontological weight, but will be used to guide the interface display of the terms within more intuitive sets.

It is possible to argue that the addition of new state terms to an existing set of states could alter the contextual meaning of all the states in that group. However, in order to maintain the compatibility of descriptions recorded with old and new versions of the state group it is important to declare that the definitions of the terms are not altered by this process, and that meaning is unambiguously captured in the textual definition of a state. This is particularly relevant if a user chooses to make only a limited number of states available within a proforma to describe a given structure: to allow compatibility with other taxonomists' descriptions Prometheus will have to treat each usage of the state (applied to that particular structural context) as equivalent.

Expandable Ontologies.- At this stage the aim is not to produce a complete ontology for the description of even a limited taxonomic domain (angiosperms), but to create an expandable ontology which can be augmented with the addition of more state and structure terms as required. The addition of these terms will not alter the meaning of existing terms nor the interpretation of proformas or descriptions composed with earlier versions of the ontology. To this end inclusion of new structures and 'Part-Of' reationships does not alter the existing hierarchy and paths, but creates new additional context paths (e.g. to insert a new structure **X** into the structure hierarchy **C** part of **B** part of **A**, so that **X** is a part of **B** and **C** is a part of **X**, we retain the path **A.B.C**, and introduce a new path **A.B.X.C,** so that **C** now has two possible structural contexts, the original one still being valid).

Similarly in the early stages of ontology development it is to be expected that users will find the need at the time of proforma specification to add a new defined state term to the ontology. Alterna-

tively, a user might desire to add an existing state to another usage group, or expand the list of structures that are linked to a state group. In either case these changes can be achieved by simply extending the ontology, without altering the interpretation of previous proformas and descriptions.

In the future it is conceivable that the ontology could be further constrained by adding more expressive relationships. For example these might include: Temporal or Developmental Relationships ('becomes'); Evolutionary Relationship ('homologous with'); Functional Equivalence ('Leaf' with 'Phylloclade'). More complex relationships such as dependencies could be added as functions or rules (e.g. state based dependencies and exclusivity: 'if tepals present, then sepals and petals not', 'if glabrous, then hairs not present').

RESULTS

The Process of Ontology Creation.- Whilst there is consensus amongst taxonomists that a more formal description methodology would be beneficial, so far it has not been possible to agree a standard description format or terminology (e.g. TDWG 2000). Furthermore, analysis of some suggested rigid approaches (e.g. Diederich & al., 2000) by Prometheus II taxonomists suggest they are limiting and inevitably lose some of the expressiveness of traditional descriptions.

A minimal requirement for enhancing the interpretability and comparability of specimen descriptions is the provision and consistent (constrained) use of a set of defined terms, with agreed, unambiguous meaning. It is relatively straightforward to compile a list of terms with definitions from standard botanical works, but agreeing a single definition for a given term can be contentious. In order to circumvent this problem we have decided to build an initial ontology containing defined terms that are applicable to only a limited domain of taxonomy (in this case, angiosperms). Typically ontologies seek to capture consensual knowledge in a domain, however, in this case an ontology is being developed to promote consensus within the taxonomic community. We expect that this ontology could be expanded to include terms necessary to describe similar taxonomic groups (e.g. Gymnosperms), but that separate ontologies might be required for more distant taxa (e.g. fungi, algae). Descriptions compiled using terms from the same ontology would be inherently compatible, but there would be no automatic compatibility when comparing descriptions composed using distinct ontologies without first performing a mapping of term equivalencies between the ontologies. This reflects the reality of there being less value in comparing taxonomic descriptions across widely separated phylogenetic taxa. It is hoped that by demonstrating the benefits of using an ontology that specifies a constrained description terminology an increasing number of taxonomists would adopt the ontology and contribute to its maintenance and expansion to include terms specific to their taxonomic domain.

The angiosperm ontology was created and recorded in a standard relational database with simple Java tools which allowed the taxonomists to enter terms and definitions, define relationships between terms and view the resultant ontology. Instances of specimen descriptions (i.e. sets of description elements) will be stored within the same overall database schema.

The Structural Ontology.- Taxonomists initialy expressed concerns regarding the specification of an ontology that defines and constrains the terms used in specimen descriptions. It was argued that this could restrict the flexibility and expressiveness of current natural language description, which is, how-

ever, not machine processable. Our first goal, the creation of a structural ontology for angiosperms was considered to be one of the less contentious areas of botany, especially as we are currently limiting ourselves to consideration of macroscopic anatomical-morphological features found in traditional specimen descriptions. This might subsequently be expanded to include further structural terms (e.g. microscopic and subcellular terms).

The taxonomists felt that it was important not to create an idealized, universal structural map for angiosperm description, but to reflect the variety of possible structural compositions found across the whole taxon. This requirement is met by allowing a given structure to be defined as potentially part of several other structures. Only when one of these contexts is chosen and used in an actual description will that particular structural context be affirmed. For example, *androecium* appears as part of five strucutures the hierarchy (*androgynophore*, *androphore*, *column*, *flower* or *gynostemium)*. Each of these five structures could belong to a variable number of superstructures, in this case a *flower* can be part of *inflorescence* or *florets*, so in total there are ten possible context paths that a proforma creator has available for *androecium*.

Because of this expansion of possible path numbers when the hierarchy was composed in this fashion it was decided that very frequently occurring structures (i.e. 'Generic Structures' e.g. *hairs*, *pores*) and Regions of structures (e.g. *base*, *apex*) would not be explicitly specified in the ontology structure hierarchy, but added at the time of creating a Proforma Ontology, only where they are of interest for the specimen descriptions.

The State Term Ontology.- The taxonomists perceived the selection of the defined state terms for the ontology to be even more problematic than selection of structure terms. It was felt that individual taxonomists used personal preferred state terms and had an individual perception of state definitions and relationships between states. It is the aim of Prometheus to adequately define state terms to ameliorate these individualistic working practices. Another objection was the perception that prepopulating the ontology with state terms is counter to taxonomic practice, where a taxonomist creates his concepts of extant characters only by examining the specimens. Creation of a defined term list might imply predefinition and restriction of allowed character states to capture specimen descriptions, a perceived criticism of existing electronic description formats. However, in the Prometheus model the ontology state terms are not character definitions, but are part of the vocabulary used to compose 'character' descriptions (i.e. description elements) at the time of specimen description.

As the state term lists were being compiled, it became apparent that a large number of commonly used terms merely expressed the presence/absence of a structure (e.g. *stipulate:* possessing stipules), or enumerated a structure (e.g. *biovulate*: containing 2 ovules). A central aim of Prometheus is to create more explicit, quantitative descriptions, to improve compatibility. As such there are explicit mechanisms to record presence or absence, and to count structures and use of such state terms is discouraged. This becomes problematic where the state descriptors both imply presence of a structure, and the state of that structure (e.g. *tomentose*: densely covered in short hairs). To support current taxonomic practice some such terms are currently included in the ontology, particularly where the implied structure is a region or generic structure. Ideally the information expressed would also be captured explicitly in Prometheus model terms, perhaps by encoding translation rules for these terms in the ontology, or possibly

by introducing an explicit mechanism at proforma creation time. Taxonomists should be made aware that if they use such terms they might lose detail and comparability in their descriptions.

A Prototype Ontology.- Our first draft ontology for angiosperm description (biased towards the specific needs of our local taxonomists) contains 24 Region Terms, 46 Generic Structure Terms, 269 Structure Terms and 695 Qualitative State Terms. 126 Of the Structure Terms are defined as Types of other Structure Terms. This leaves 143 Structure Terms that are part of the structure hierarchy for which there are 170 optional 'Part Of' relationships described. Only 19 Structure Terms are currently described as potentially part of more than one superstructure.

The State Terms are distributed between 72 State Groups that reflect their usage context, with between 2 and 79 members of each group. 38 State Terms are members of more than one (typically 2) group. 17 Of the groups have their constituent states organized into subgroups for presentation purposes.

A print out of the state groups and the structure hierarchy represented as an expanded tree, and the whole ontology in XML format can be viewed online (Prometheus, 2003b). Each of the 536 structure nodes in the tree is identifiable by its unique path; of these 331 are leaf nodes (with no substructures).

CONCLUSIONS

The taxonomic process (as shown in Fig. 1) relies on the accurate recording and reinterpretation of character description data. The Prometheus II project aims to improve the quality and interpretability of data storage, which should enhance many of these taxonomic procedures.

Currently not only is there no agreed methodology for creating a taxonomic description, but there is no agreed data model or standardized terminology with which to record the data. Hence existing storage formats do not support data integration in that they lack a common model and terminology for representing 'character' description data. Whilst the identification of useful distinguishing characteristics for classification remains the critical taxonomic skill, the ability to accurately and unambiguously describe these characters and character states would be greatly assisted by use of a constrained description methodology and language (ontology).

A widely accepted description ontology can only arise by consensus, with individual taxonomists contributing to the ontology in their realm of expertise. The taxonomic range over which an ontology will be useful will depend on how rigorously terms can be defined, how rigorously they are used and indeed whether meaningful comparisons can actually be made across the range. The benefits of the constrained data model and ontology will only be seen if software is provided which supports (indeed requires) the methodology.

Our 'character' description model allows taxonomists to explicitly record each aspect contributing to a potential 'character' as a description element, without requiring that characters are defined in advance of description. The development of an ontology which captures the potential structural relationships between terms allows these description elements to explicitly record not only what defined structure is being described, but also to capture the context of the structure as its 'path' as defined in the ontology.

The angiosperm ontology created in this study can be used as a parent ontology for the specification of individual proforma ontologies, which will contain all the terms necessary to describe a set of

specimens. Creation of a proforma ontology will involve: choosing which structures, in which context (i.e. structure path), are required for description; the addition of relevant generic structures and regions to the structure hierarchy; and the specification of which of the permissable state groups and quantitative properties will be scored for each of the structures in the ontology. Because individual proforma ontologies are derived from the same parent ontology data sets created using separate proforma ontologies will automatically be compatible with each other. Furthermore, the interpretation of descriptions will be unambiguous because all terms used in the description data are unambiguously defined .

To support our proposed methodology we are developing a suite of tools for specifying descriptive ontologies, automating the generation of description proformas, and providing interfaces for entering and storing descriptions to a standard relational database. Such a database will form a repository of compatible descriptive data.

LITERATURE CITED

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

Bailey, J. (ed.) 1999 The Penguin Dictionary of Plant Sciences. Penguin, London.

Blackwelder, R .E. 1967 *Taxonomy: A text and reference book*. John Wiley, New York.

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

CBIT 2003 Lucid is developed by The Centre for Biological Information Technology: University of Queensland, Australia. URLs: www.cpitt.uq.edu.au; www.lucidcentral.com

Colless, D. H. 1985 On 'character' and related terms. *Systematic Zoology* 34: 22--233.

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

Dallwitz, M. J. &. Paine T .A 1986. Users guide to the DELTA system. *CSIRO Division of Entomology Report*. 13: 3--6.

Davis, P. H. & Heywood, V. H. 1963 *Principles of Angiosperm Taxonomy*. Oliver and Boyd, 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: 1--30.

Fristrup, K. 1992 Character: current usages. Pp. 45--51 in: Keller, E. F., Lloyd, E. A. (eds.) *Keywords in evolutionary biology.* Harvard University Press, Cambridge.

Hagedorn G. 2002 Structure of Descriptive Data Conveners Report. TDWG Internatioanl Meeting, Brazil. URL: 160.45.63.11/Projects/TDWG-SDD/Minutes/2002Brazil_report/Presentation.html and www.prometheusdb.org/resources.htm

Hickey, L. J.1973 Classification of the architecture of dicotyledonous leaves. *American Journal of Botany* 60: 17--33.

Keogh, J. S. 1995 The importance of systematics in understanding the biodiversity crisis: the role of biological educators. *Journal of Biol. Educ*. 29: 293--299.

Maddison, D. R., Swofford, D. L. & Maddison, W. P.1997 NEXUS: An extensible file format for systematic information. *Systematic Biology* 46: 590--621.

Prometheus 2003a URL: www.prometheusdb.org

Prometheus 2003b URL:

Sivarajan, V. V. 1991 *Introduction to the Principles of Plant Taxonomy.* Cambridge University Press, Cambridge.

Stuessy, T. F. 1990 *Plant Taxonomy: the Systematic Evaluation of Comparative Data*. Columbia University Press, New York

TDWG 2000 International Working Group on Taxonomic Databases subgroup: Structure of Descriptive Data.: Subgroup session report at the TDWG meeting in Frankfurt (2000)
URL: www.tdwg.org/tdwg2000/sddreport.htm and www.prometheusdb.org/resources.htm

Watson, L. 1971 Basic taxonomic data: the need for organisation over presentation and accumulation. *Taxon* 20: 3--136.

Wiley, E. O. 1981 *Phylogenetics: the Theory and Practice of Phylogenetic Systematics.* John Wiley, New York.

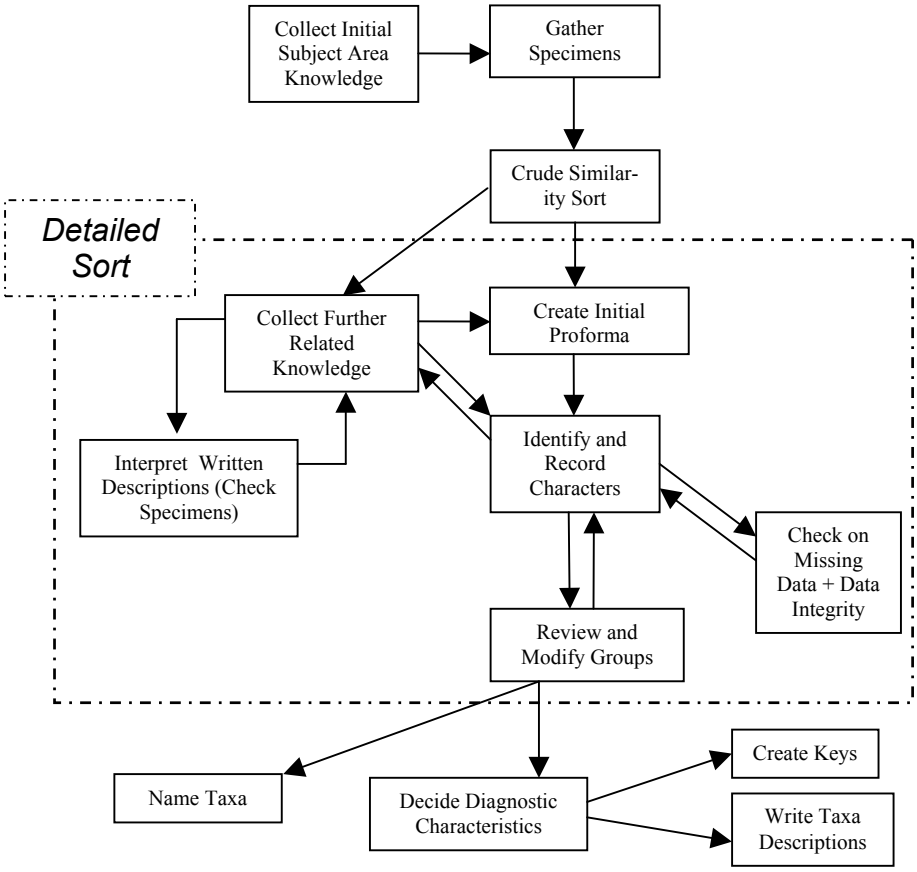
Wilkinson, M. 1995 A comparison of two methods of character construction. *Cladistics* 11: 29--308.

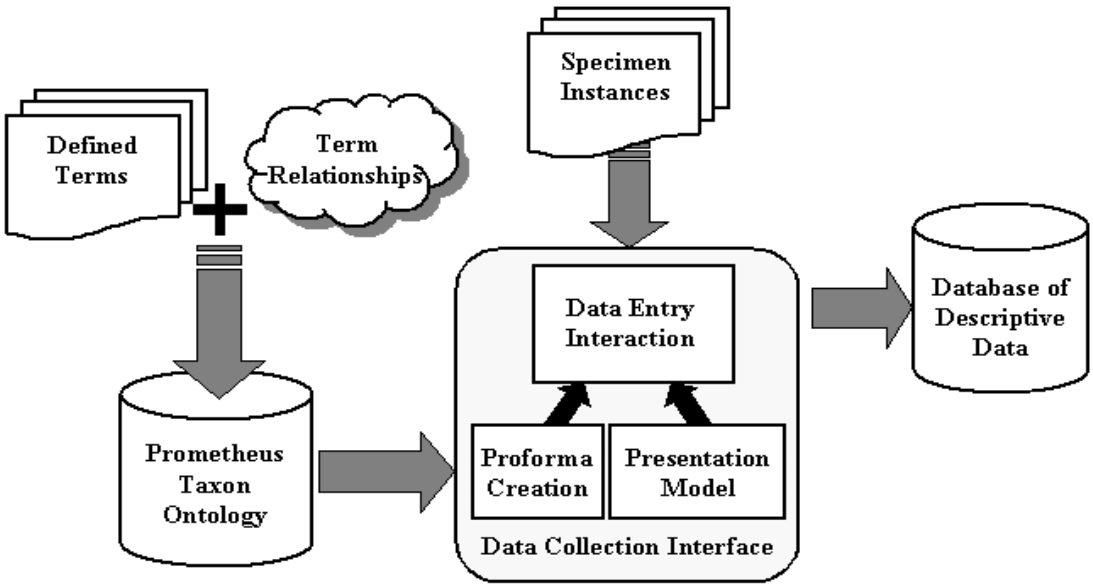
Fig. 1. The Taxonomic Process. An initial crude sorting of specimens into groups based on gross overall similarity is followed by an iterative detailed sort based on individual ‘characters’ which are recorded in a proforma document. Existing literature, research and knowledge informs the selection of appropriate ‘characters’. Final decisions and write-ups occur at the end of the process. A single taxonomist is usually responsible for the whole process.

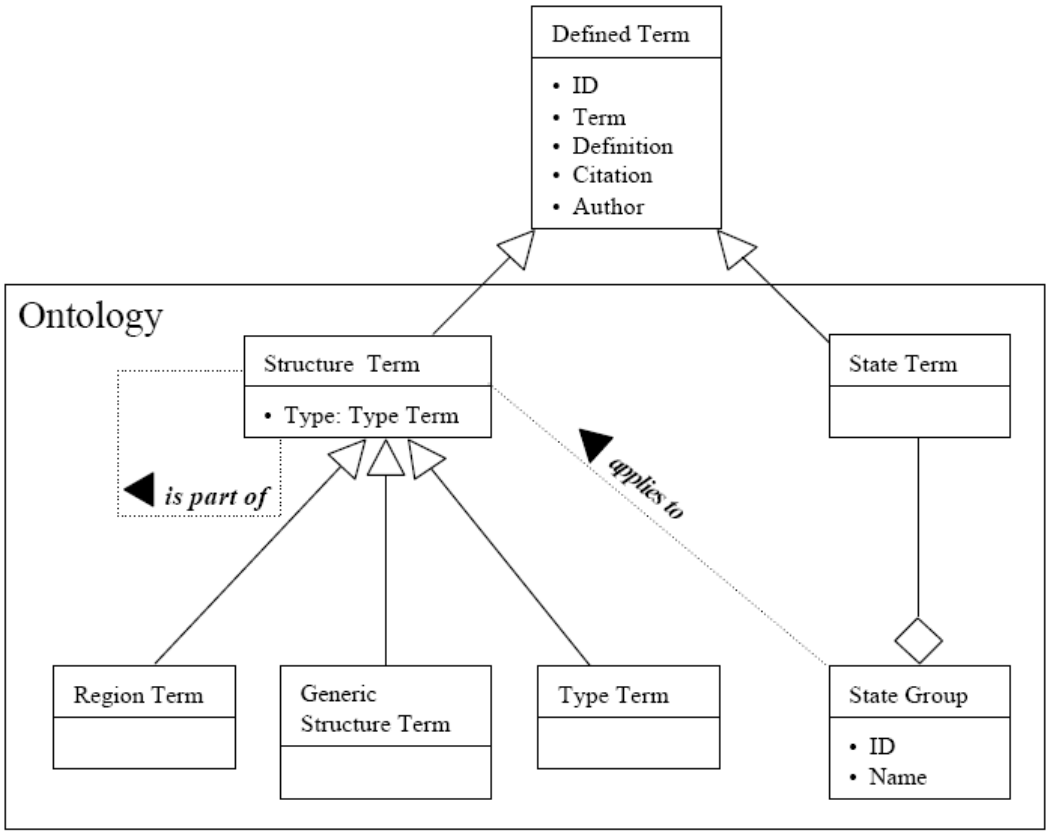
Fig. 2. The Prometheus II Description Process. An ontology of descriptive terms is created, from which users design proforma description templates for specific taxonomic projects. Based upon this proforma and an exisiting presentation model, a user interface is automatically generated facilitating data entry. Stored specimen descriptions are composed of unambiguously defined terms.

Fig. 3. The terms and relationships in the Prometheus Ontology. All Terms are Defined Terms, the main classes being Structure and State Terms. State Terms are aggregated into State Groups, which can be restricted to apply to specific Structure Terms. Region Terms and Generic Structure Terms are specialized types of Structure Terms that cannot participate in 'Part-Of' relationships between Structure Terms. Structural Types can only be used as an attribute of a true Structure Term.

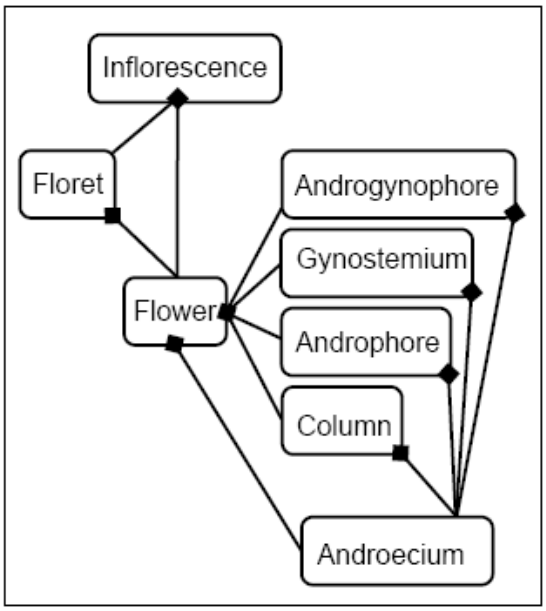
Fig. 4. Representing the 'Part-Of' Hierarchy. (a) As a Network Graph (b) As a Tree Graph. Each node in (b) has a unique identity, which can be represented as the 'path' of the node (e.g. Inflorescence.Floret.Flower.Androecium as opposed to Inflorescence.Flower.Androecium).


Collect Initial Subject Area Knowledge
Gather Specimens
Crude Similar-ity Sort
Detailed Sort
Collect Further Related Knowledge
Create Initial Proforma
Interpret Written Descriptions (Check Specimens)
Identify and Record Characters
Check on Missing Data + Data Integrity
Review and Modify Groups
Name Taxa
Decide Diagnostic Characteristics
Create Keys
Write Taxa Descriptions

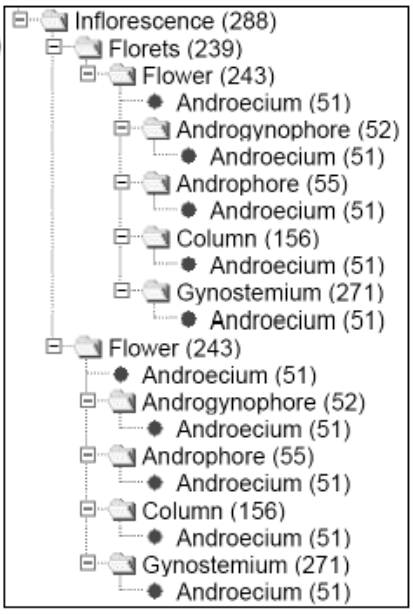

Specimen Instances
Defined Terms
Term Relationships
Data Entry Interaction
Database of Descriptive Data
Prometheus Taxon Ontology
Proforma Creation
Presentation Model
Data Collection Interface


Defined Term
• ID
• Term
• Definition
• Citation
• Author
Ontology
Structure Term
• Type: Type Term
State Term
is part of
applies to
Region Term
Generic
Structure Term
Type Term
State Group
• ID
• Name

(a)


Inflorescence
Floret
Androgynophore
Gynostemium
Flower
Androphore
Column
Androecium

(b)


Inflorescence (288)
Florets (239)
Flower (243)
Androecium (51)
Androgynophore (52)
Androecium (51)
Androphore (55)
Androecium (51)
Column (156)
Androecium (51)
Gynostemium (271)
Androecium (51)
Flower (243)
Androecium (51)
Androgynophore (52)
Androecium (51)
Androphore (55)
Androecium (51)
Column (156)
Androecium (51)
Gynostemium (271)
Androecium (51)