A Universal Character Model and Ontology of Defined Terms for Taxonomic Description

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

Abstract. Taxonomists classify biological specimens into groups (taxa) on the basis of similarities between their observed features ('characters'). The description of these 'characters' is therefore central to taxonomy, but there is currently no agreed model, defined terminology nor methodology for composing these descriptions. This lack of a common conceptual model, together with the individualistic working practices of taxonomists, means that descriptions are not composed consistently, and are not easy to interpret and re-use, nor are datasets comparable. The purpose of the Prometheus II project is to improve the interpretation and comparison of plant descriptions. To this end we propose a new conceptual model for unambiguously representing character descriptions, and have developed a controlled vocabulary as an ontology of defined terms, which will be used to describe specimens according to our character model.

1 Introduction: Problems With The Quality Of Descriptive Data

Taxonomy is the branch of biology concerned with the classification of organisms into an ordered hierarchical system of groups (taxa) reflecting their natural relationships and similarities. The central taxonomic process, establishing relatedness and classifying organisms, is based upon the identification and description of variation between comparable structures on different specimens, with the critical taxonomic skill being the identification of such 'characters' that prove useful for classification. Integral to this process is the ability both to define the 'character' concepts used, and to describe the observed 'character states' precisely.

Whilst character data are the basic building blocks of descriptive data, there is little consensus amongst taxonomists on what the term ‘character’ actually means, making the interpretation of taxonomic descriptions problematic, nor is there an agreed terminology with which to compose descriptions. Specimen descriptions represent a huge potential data resource, not just for future taxonomic revisions, analyses and the creation of identification keys *etc*., but for other biological disciplines such as biodiversity and ecological studies. However, these uses require the meaningful integration of data from different description sets, which, in the absence of both an agreed character model and particularly a shared descriptive terminology, is currently not possible.

E.Rahm (Ed.): DILS 2004, LNBI 2994, pp63-78, 2004.
Springer-Verlag Berlin.

A character concept is derived during the taxonomic process by partitioning observed variation into characters and 'character states' i.e. the combination of a structure, the aspect of the structure being described (its property) and its possible states [1,2]. For example, a 'character' *leaf shape* may be recognized and, in a group of specimens, the 'states' *obovate*, *ovate* and *oval* observed. The description of that group of specimens would then read ‘leaves obovate, ovate or oval’.

A ‘character’ might be defined in general terms as ‘a statement on a feature of the organism’ although many different specific definitions have been proposed [3]. Diederich *et al.* [4] propose a useful universal definition of 'character', decomposed into structure, property and score, which enables taxonomists to be explicit about every aspect of a character statement. However, Diederich did recognize that in many usages of 'character', the 'property' is not explicitly recorded.

Taxonomic descriptions are composed of descriptions of character states for an individual specimen or a group (i.e. a taxon, such as a particular species, genus *etc*.). Traditionally descriptions are recorded in semi-formal natural language, and several electronic description formats and applications have been developed to allow the storage and analysis of data [5-7]. However, these formats have been developed to support the flexible use of character concepts and terminology. Flexibility implies a lack of standardisation in the use of character concepts, and in the absence of a well-defined character model and an agreed terminology, descriptions are generally only consistent within a single data set. Consequently, taxonomists cannot communicate the basis of their work adequately [8], nor meaningfully integrate data from various sources.

To date it has not been possible to achieve universal definitions for characters or the terminology used to describe 'characters' (see for example the experiences of TDWG who have attempted to standardize the terminology for botanical descriptions [9]). This is not surprising given the wide variation in structures and characters across the whole taxonomic range (e.g. comparing algae with flowering plants). Furthermore, descriptive terminology is domain specific, with the same word having differing meanings in different taxonomic groups (homonyms), or different words being used in various taxonomic fields to describe the same concept (synonyms).

The Prometheus project [10] aims to improve the methodology for taxonomic description and provide tools for recording data more rigorously. Taxonomists recognize problems with current working practice, and several authors have suggested that there should be a standard approach to taxonomic description [4,9,11], however, they are concerned that an improved methodology should not restrict the expressiveness of their descriptions. Prometheus aims to provide an integrated suite of tools for developing descriptive ontologies, automating the generation of proformas (description templates, detailing the 'characters' to be scored for a specimen), and providing interfaces for entering and storing descriptions to a database that will form a repository of compatible descriptive data. A prototype ontology has been developed which defines and constrains terms necessary for describing flowering plants (angiosperms).

2. Representation of Characters as 'Description Elements'

We propose a new data model for character description, which facilitates the standardization and integration of data. 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 creation of a defined terminology with which to compose actual descriptions is addressed in section §3.

To allow taxonomists to be explicit about every aspect of a character statement we have developed Diederich's definition of 'character' [4], which he decomposed into structure, property and score (Fig 1.). This composition of character is represented as a Description Element (DE), in which a character description is created by recording the defined structure, the defined property and the observed score (Fig 1.). However, we note that taxonomists record both quantitative and qualitative data and that whilst this decomposition is readily applicable for quantitatively measured characters with a real score (such as the properties *length, width, height etc.*), many qualitative statements record the 'state' of a structure as the score, and often the associated 'property' is less readily discernable, and typically not explicitly recorded. To accommodate this, the model requires two kinds of DEs: Quantitative and Qualitative, where Qualitative DEs do not require explicit association of a property with the structure/state combination (Fig 1.). *Specimen Descriptions* are composed of the set of DEs for that specimen.

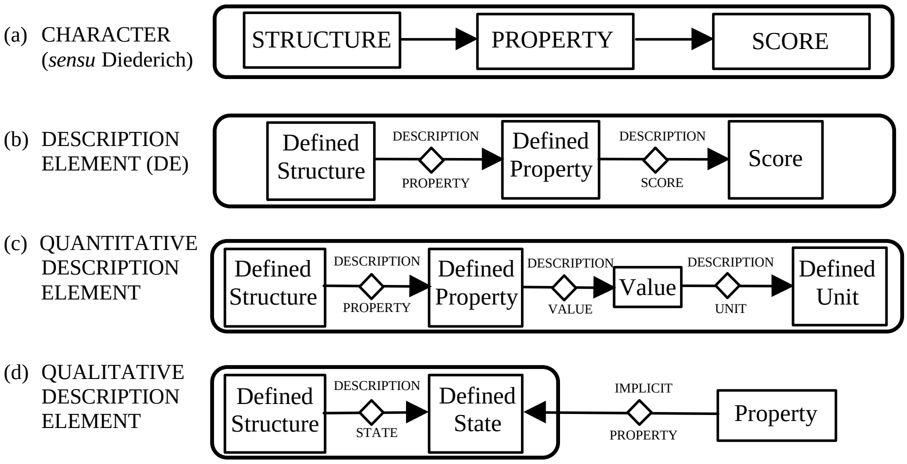


Fig. 1. Diederich's decomposition of 'character' into structure, property and score is represented as a Description Element, which uses defined terms to capture each element of a character statement. Quantitative DEs include a numeric value as score and require a defined unit (e.g. Leaf Length 5 mm); Qualitative DEs do not explicitly record a property, but states describe an implicit property (e.g. Leaf Oval *(OutlineShape)*). The recording of multiple values or states within a DE is discussed later (§2.3).

2.1 Quantitative Description Elements.

In order to record the 'character' statement ‘leaf length 5 cm’ a quantitative DE is composed specifying a defined structure (*leaf*), an explicit defined property (*length*), a value (an individual number: *5*) and the appropriate defined unit (*cm*). For quantitative statements that do not have units, e.g. *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 {*Angle, Density, Diameter, Height, Length, Number, Width*} and be expanded to allow further defined quantitative properties as needed (e.g. *Colour*, as defined by RGB values *etc.*). Detailed ontologies defining measurement concepts (i.e. units and dimensions *etc*.) for the biological domain are being developed by others (e.g. [12]) and could ultimately be used to constrain and define the terms used in Prometheus Quantitative DEs.

2.2 Qualitative Description Elements.

In order to correctly record a statement such as *‘leaves oval’* a qualitative DE is composed with a defined structure (*leaf*) and a defined qualitative state (*oval*). Note that no explicit property is specified for qualitative scores, although a state is associated with an implicit property, which might be defined by grouping states into 'usage groups' (see section §3.2).

Arguably it should be possible to describe all physical data quantitatively, and Prometheus would encourage quantitative description where practicable to permit the direct comparability of DE data. However, often this is neither reasonable nor useful to taxonomists, who assign qualitative states by categorising continuous quantitative variation or represent complex character properties with more easily handled discrete states. 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.

2.3 Representing Concrete and Abstract Data

When describing a specimen, character data may be an accurate record of the state of an actual individual structure, or may represent an average or representative state for the collection of such structures on the specimen. Taxonomists use both types of data but often do not distinguish between them. In order to distinguish between the former and latter case, a DE can be explicitly recorded as Concrete or Abstract. Some taxonomic work will represent variety by recording a large number of instances of concrete DEs for a structure/state, whereas other work will express variety as a collection or range of abstract DEs. The types of analyses that can be performed on description data will depend on whether the data is real (concrete) or summary (abstract). Taxon descriptions are by definition abstractions as they are a summary of the specimens in the taxon and will only be composed of abstract DEs.

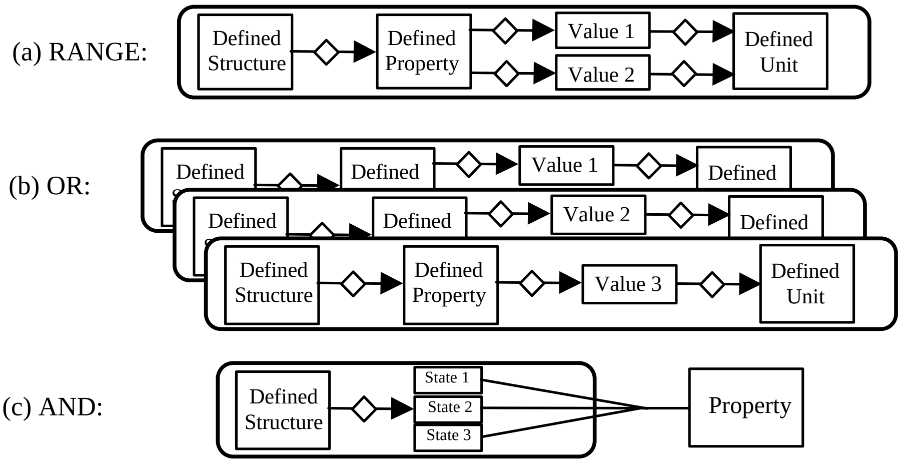


Fig. 2. Capturing Variability in Description Elements. (a) Quantitative ranges are captured by storing two values in a quantitative DE (e.g. Leaf Length 5 to 10 mm). (b) Multiple DEs describing the same structure and property capture multiple single-valued alternative states (OR) (e.g. Petal Number 4 or 8 or 12). (c) Multiple states of a single property can be saved in a single qualitative DE (AND) (e.g. Leaf Brown and Green and Yellow *(Colour)*).

2.4 Recording Variability in Description Elements

It is common practice for a description to record a range of values for a given measurement (e.g. *length 5-10mm*). The frequency with which recording ranges is necessary makes it sensible for our implemented model to allow a quantitative DE to explicitly record a pair of score values to capture a range (Fig.2a). However, an observed range is not necessarily a continuum, for example if the flowers on a specimen may be observed to have 3, 5 or 7 petals. In this case the 'range' can be represented by recording multiple alternative (OR'ed) DEs for that property (Fig. 2b). Only abstract DEs will ever express ranges or alternatives, as concrete DEs record actual measurements for a single, real structure.

Taxonomists also currently record ranges in qualitative states, for example *leaves round to ovate*. It is not possible to unambiguously interpret such a description, as there is no representation of intermediates in the categorized continuum of the character states. Therefore any interpretation of a range is subjective. If it is not possible to record a range quantitatively it is necessary to represent the range of possible qualitative states by defining states that break up the continuum of variation without leaving significant gaps, and to record the existence of multiple alternative (OR'ed) DEs for that property (similar to Fig.2b). For example, a specimen may have leaves with apices that range between *acute* and *acuminate*. Ideally this would be recorded quantitatively as a range of angles (e.g. 10-80^o). However, this range could also either be defined by two qualitative states which encompass a wide range themselves (leaves with an apex angle of <50^o are *acute*; with an apex angle of >50^o *acuminate*), or a user could define a large set of states that describe all the possible intermediate angles. The leaf apex is then described with a set of DEs that explicitly describe all the possible variations.

Whilst (abstract) quantitative DEs can record ranges, or alternative values, for a quantitative property, a single quantitative DE can only have one measurement for one property (e.g Leaf: Length: 5 mm; without considering measurement accuracy here). On the other hand, because qualitative states are not simple quantitative measures and are not necessarily mutually exclusive, qualitative descriptions may include multiple states ('AND' Fig. 2c) (e.g. Leaf: Brown and Green and Yellow). A qualitative DE can therefore record multiple (AND) states for a given score, whereas alternative (OR) states are recorded in multiple linked DEs for separate instances of a structure, with each DE describing the same implicit property (e.g Leaf: Brown OR Leaf: Yellow OR Leaf: Green).

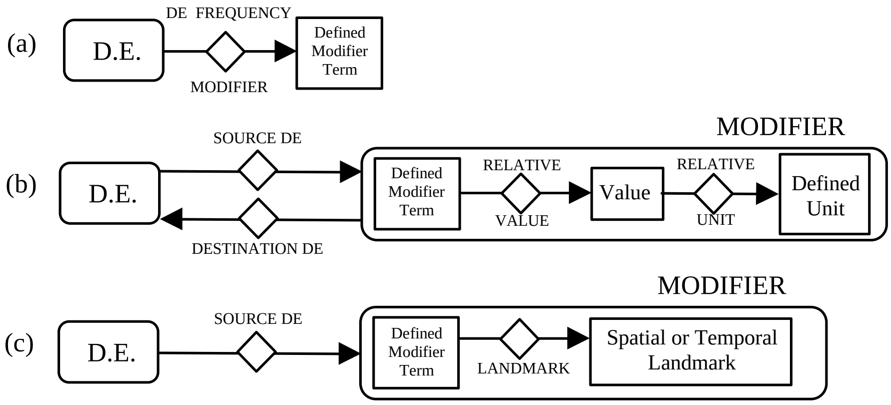


Fig. 3. (a) Description Elements (DEs) can be modified by a simple frequency modifier term. (b) Two DEs can be related by Relative, Spatial or Temporal Modifiers (relative modifiers may also include a value and unit). (c) DEs may also be modified by landmark statements.

2.5 Modifiers of Description Elements

To allow rich and flexible description using our model it is necessary to allow additional information to be associated with DEs, this is achieved by using 'modifiers'. The simplest modifiers are *frequency modifiers*, which are simple defined terms that can be added to an abstract DE to indicate relative occurrence {e.g. *mostly, often, usually, sometimes, rarely*} (Fig.3a).

Other modifiers are required to relate two DEs in order to capture one state in relation to another, (e.g. *leaf length* in comparison to *leaf width*). These modifiers therefore link source and destination DEs, and have associated defined terms and possibly values (Fig. 3b). We distinguish three types of these modifiers, with associated sets of defined modifier terms: *Relative:* {*greater-than, less-than, equal-to, ratio, not-equal-to, less-than-or-equal-to, greater-than-or-equal-to*}; *Spatial:* {*at, above, below, between*}; *Temporal:* {*after, before, while*}. Relative modifiers allow undefined scores to be related (e.g. *leaf length* 'less than' *leaf width*) or with an associated value (e.g. length is twice width: *length* 'ratio: 2' *width*). Spatial Modifiers allow measurements to be more accurately defined (e.g. *trunk diameter* 'at' *branch*. In order to allow the flexibility of natural language descriptions these modifiers can also relate a DE to a 'landmark statement', for example *trunk diameter* 'at' <breast height> (Fig.3c). Temporal modifiers allow the time of year, or sequential order of events to be recorded, and can again use 'temporal statements', e.g. *flowers* 'in' <spring>; *fruit colour* 'before' *fruit colour* (i.e. to describe unripe before ripe).

Whilst these modifiers allow storage of rich data in a less ambiguous form, reflecting the current style of natural language/free text description, actually processing and analysing some of this data in comparisons might prove highly complex, especially if landmark statements are included, which are essentially free text.

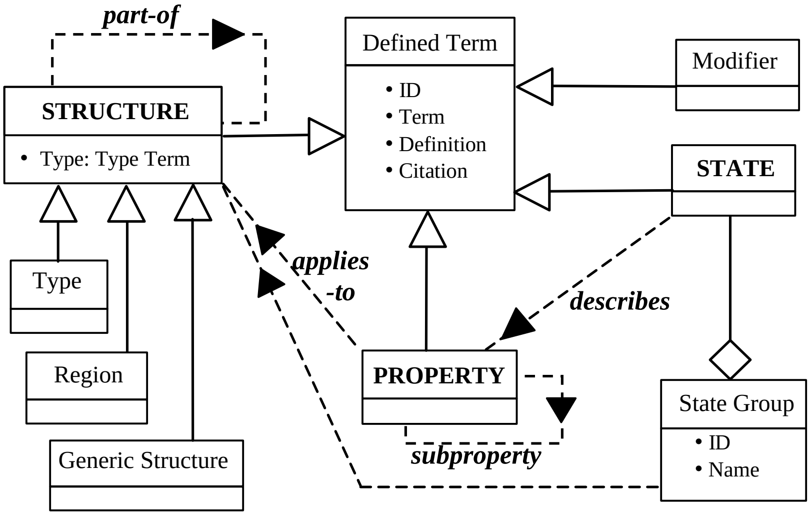


Fig. 4. Concepts and relationships in the descriptive term ontology. All terms are specializations of Defined Term. Structures can be 'Part-Of' other structures recursively, and may have attribute: Type (itself a specialized Structure). States are composed into groups, which may be restricted to ('applies-to') certain structures. Therefore these state groups may represent 'de facto' properties, which may include a structural context. Alternatively states can be considered to describe a given property, which may be applicable to only certain structures.

3 Using an Ontology to Specify a Defined Terminology

Taxonomists expressed concerns that using an ontology to define and constrain term usage in specimen descriptions might restrict the flexibility and expressiveness of current natural language description (which is, however, not machine processable). However, a minimal requirement for enhancing the interoperability of specimen descriptions is the consistent use of a set of defined terms. Capturing definitions and information about how terms can be used in relation to each other represents the creation of a semi-formal ontology, i.e. a constrained and structured form of natural language. Ontologies can be used to mediate both structural and semantic data integration by generating a unified view of local ontologies (as a mediated schema), and developing a common global ontology integrating concepts amongst data providers [13,14]. By specifying a standard controlled vocabulary for specimen description Prometheus will prevent semantic heterogeneity between descriptions that have been composed solely with defined terms from this common ontology. The relationships and classes within the Prometheus ontology are shown diagrammatically in Fig. 4. Instances of the primary classes of concepts: defined Structure, State and Property are used in Description Elements to create character descriptions according to our character model.

3.1 Structure Terms in the Ontology

The creation of a consensual ontology defining structural terms for the limited domain of flowering plants was chosen as a realistic initial goal for the project, particularly when restricted to the inclusion of macroscopic anatomical-morphological features found in traditional specimen descriptions. This ontology might subsequently be expanded to include further structural terms (e.g. microscopic and subcellular terms).

The ontology requires definition of the structure terms necessary to describe angiosperms. However, it is important that a description captures not only the structure being described, but its structural context and composition i.e. what it is *part of*, and what structures are *part of it*. These potential relationships between structures in the ontology are captured with a 'Part-Of' relationship (Fig. 4).

Rather than creating a universal structural 'map' of an idealized angiosperm, the taxonomists required that Part-Of relations in the ontology reflect the variety of possible structural compositions found across the 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 proforma or description will that particular structural context be affirmed. The Part-Of hierarchy therefore forms a Directed Acyclic Graph, but can be viewed more intuitively by the taxonomists as a branched tree with multiple instances of some structures (Fig. 5a,b). For example, *androecium* appears as part of five structures in the hierarchy and each of these structures can have two structural contexts, dependent upon whether florets are present, giving ten possible context paths that can be chosen for *androecium* (Fig.5c). Each path uniquely identifies the 'node' in the structure hierarchy, and allows a description to unambiguously specify the described structure both as a defined structure term and by its relationships to other defined structures. Defining terms in such an ontology, which specifies structural relationships between parts, therefore improves our character model, allowing specification of a defined structure in its context.

There are certain anatomical structures (e.g. *hairs*, *pores*; referred to here as 'Generic Structures') that might potentially be part of many if not most other structures. Similarly any structure can be subdivided into 'Regions' (e.g. *base*, *apex*). If Part-Of relationships were explicitly recorded for these structures there would be an unmanageable explosion of possible context paths in the ontology. For clarity we decided that Regions and Generic Structures represent specialized types of structures that are not explicitly included in the hierarchy, e.g. the region *base* has not been added to every structure in the ontology. Instead it is the responsibility of a user of the ontology to explicitly specify where these structures should be added to an instance of the ontology for use in descriptions (e.g. adding *hairs* to *leaves* and *petals*, and *apex* to a *leaf* when defining a proforma ontology, see §3.3).

A further specialized subclass of structure terms is 'Types'. Structure terms can be defined as a 'Type-Of' another structure term if they are examples of the supertype that always have a number (more than one) of descriptive states true for each instance of that supertype structure. Types reflect an awkward apparent blurring between states and structures when describing specimens, e.g. a *berry* is clearly a structure in itself, but it is also a collection of states for a particular structure (a *fruit* that is always fleshy, indehiscent and has seeds submerged in pulp). We exclude types from the structural hierarchy and treat them as an attribute of their supertype, so that in a description a fruit can be recorded as being of type berry.

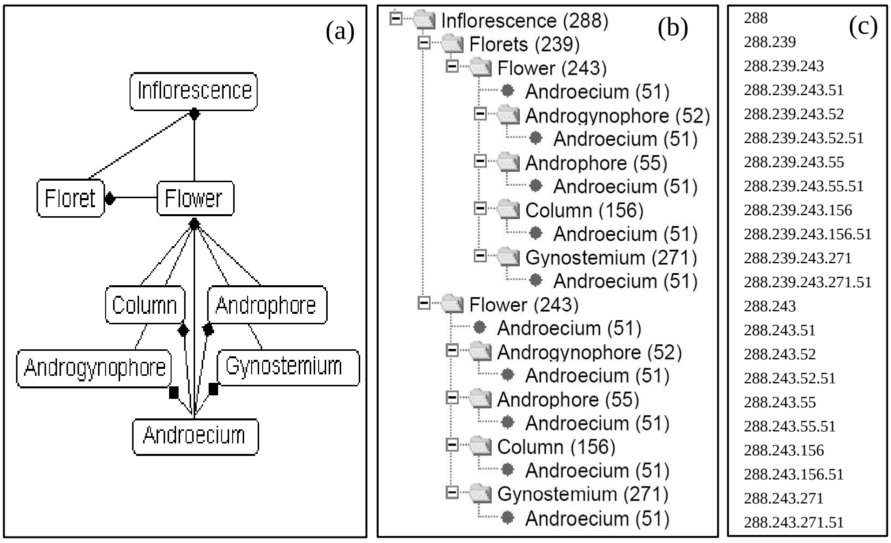


Fig. 5. Representing the 'Part-Of' hierarchy as (a) a Directed Acyclic Graph (b) a Tree Graph (c) a dot separated string of structure IDs detailing the hierarchical path.

3.2 State Terms in the Ontology

Taxonomists perceived the selection of allowed state terms for the ontology to be more problematic than specifying allowed structures. One objection was that individual taxonomists use their own personal preferred state terms and have an individual perception of state definitions and how they relate to other states. It is the aim of Prometheus to adequately define state terms to ameliorate these individualistic working practices. Another objection to prepopulating the ontology with state terms was that this is counter to taxonomic practice where taxonomists create their concepts of extant characters only by examining the specimens. Creation of a defined term list might imply predefinition and restriction of allowed character states, a criticism of other description formats. However, in the Prometheus model defined state terms are not character definitions, but are part of the vocabulary used to compose character descriptions at the time of specimen description.

As discussed previously (§2.2) it is difficult to define in a consistent and non-arbitrary fashion the underlying 'properties' associated with qualitative DEs and state terms. This is because a taxonomist's interpretation of a state can include aspects of several properties. Initial attempts to categorize state terms in terms of the qualitative property described (e.g. {*Arrangement, Colour, Shape, Texture etc.*}) suggested that for many state terms such divisions were arbitrary and contentious. For example, whilst 'red' is clearly a state of *colour*, is 'keeled' a *shape* or an *arrangement* of petals?

However, taxonomists can intuitively organize state terms into sets that are used to describe alternative aspects of the same feature (i.e. usage groups). The composition of such a set of states (a 'State Group') could be considered to circumscribe an implicit, *de facto* 'property', which in some cases is qualified by a given structural context. This reflects taxonomists' apparent conceptualization of the qualitative property of a 'character' as a gestalt of property in context of the structure being described.

Once created, analysis of these state groups reveals a meaningful 'property' that each group describes, and any particular structural context represented by the extent of the group. Thus a hierarchy of properties and subproperties can be created, which allows all state groupings to be defined in terms of described property and a possible structural context. For example we can distinguish different subclasses of *Arrangement*: e.g. *Architecture, Form, Position etc.* Furthermore a state group may be defined by a structural context of a particular property, for example all the states describing 'leaf' architecture. If expressed explicitly these properties would allow qualitative descriptions to be represented in a similar fashion to quantitative descriptions, with all states being the score of an underlying qualitative property, analogous to a value being the score of a quantitative property.

In our ontology groups of states are restricted in their usage to describe certain allowed structures; this is captured by the 'applies to' relationship (Fig.4) between properties (or state groups) and structures. However, some groups of states or properties (e.g. '*Textures*') can apply to such an extensive range of structures that it is not sensible to restrict usage to a subset of structures.

Ideally state groups would be composed of the set of exclusive alternative states for the property of a given structure (i.e. only one state can apply to an individual structure). However, the extent of such exclusive groups proved difficult to define such that a given instance of a structure would never be described by more than one state in a group or property. All state groups are therefore considered as potentially 'multistate' for a given description instance. Where state terms appeared to belong in more than one state group, it was apparent that the contextual meanings of the terms are not identical and in such cases it is necessary to create homonymous terms (with different definitions), which belong to separate state groups.

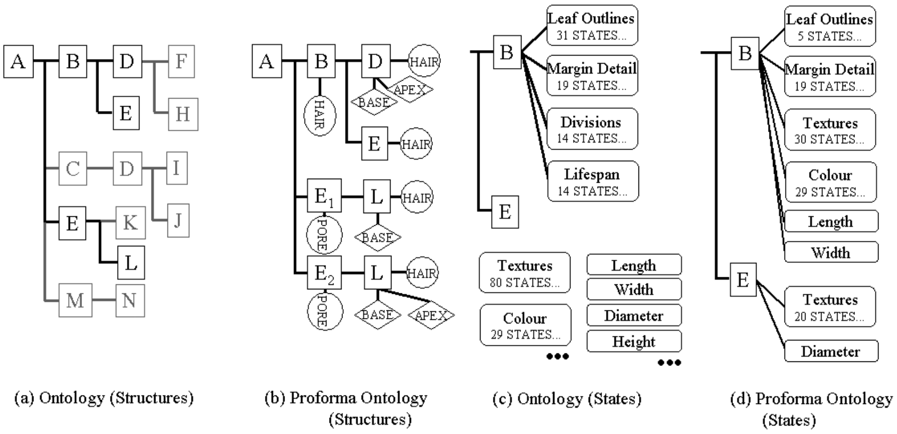
As the state term lists were being compiled, it became apparent that a large number of commonly used terms merely expressed the presence/absence of a structure (e.g. *stipulate*: possessing stipules), or enumerated a structure (e.g. *biovulate*: containing 2 ovules). In order to improve compatibility Prometheus aims to create more explicit, quantitative descriptions. As such there are explicit mechanisms to record presence or absence, and to count structures: therefore use of these types of state terms is discouraged. This becomes problematic where the state descriptors both imply presence of a structure, and the state of that structure, often where the implied structure is a region or generic structure (e.g. *tomentose*: densely covered in short hairs). The interpretation of such a state is contextual, and although *densely* and *short* could be quantitatively defined, it is impossible to define them acceptably for all contexts, thereby making it impossible to define *tomentose* quantitatively. Therefore although we could consider terms such as *tomentose* to have structural and quantitative dependencies we allow their use and assume that they are comparable across descriptions (in terms of the definition), leaving any discrepancy in the exact definition of *tomentose* to be resolved by the taxonomist where necessary.

3.3 Defining Proforma Ontologies

Our ontology aims to include all of the defined structure and state terms necessary to create DEs describing a given angiosperm specimen. Usage of the defined terms is constrained by a number of relationships specified in the ontology. States are grouped into usage groups, which may be linked to the structures which they are allowed to describe; structure types are identified for some structure terms; and a structural composition hierarchy has been specified which details all possible structural contexts. However, generic structures and regions are not yet specified in this hierarchy. In order to create an instance of the ontology that is actually used for a description set (i.e. for a description proforma) the user will select the structures (in context) he wishes to describe and which properties he wishes to describe for these structures. The taxonomist may additionally restrict the list of qualitative states that are available for a given structure. Part of this process involves explicitly adding the regions and generic structures that are to be described to the existing structural hierarchy. These processes, which both extend and restrict the parent ontology, can be seen to create a proforma specific ontology, for a given description set (Fig.6).

Descriptions composed using a specific proforma ontology are automatically compatible with other description data composed using any other proforma ontology derived from the same parent ontology, as the term definitions and structure path contexts are consistent.

A further complication with the specification of proforma ontologies, not considered in detail here, is that it may be necessary to represent any given structure node with more than one copy or clone. This will be necessary to represent multiple versions of a structure, for example when a taxonomist requires to distinguish multiple versions of a leaf that will be described separately (e.g. if there were two identifiable leaf forms present on specimens, some which tended to be small, brown and hairy, and others that were large, pink and glabrous).



(a) Ontology (Structures) (b) Proforma Ontology (Structures) (c) Ontology (States) (d) Proforma Ontology (States)

Fig. 6. (a) An ontology specifies all potential *Part-Of* paths in the structure hierarchy. (b) Many structures and paths may be deleted in an instance of a proforma ontology, some structures may be duplicated (*e.g. E*), and appropriate generic structures and regions explicitly included (*hairs, bases etc*). (c) Some qualitative state groups in the ontology are restricted to specific structures (*Divisions, Lifespan etc*), others (*Colours etc.*) and all quantitative properties are applicable anywhere. (d) Proforma ontologies may remove some allowed state groups for a structure, and possibly remove some allowed states from these groups; quantitative properties and unrestricted state groups will be explicitly attached to structures where desired for description.

3.4 Expanding the Ontology

At this stage we do not claim to have produced a complete ontology for the description of angiosperms, but have created an expandable ontology which can be augmented with the addition of more state and structure terms as required, providing that the addition of these terms does not alter the meaning of existing terms nor the interpretation of proformas or descriptions composed with earlier versions of the ontology. To this end new structures cannot be added within an existing structure path in the hierarchy, but a completely new path must be added (e.g. to insert a new structure *X* into the hierarchy [*C part-of B part-of A*: path A.B.C], 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 contexts, the original one still being valid). Similarly, new and existing state terms might be added to new or existing state groups, or new links between state groups and structures expressed.

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 group it is important to declare that the definitions of the terms are not altered by this process, and that the meaning is unambiguously captured in the textual definition of that state.

4 A Prototype Ontology and Future Work

A Java tool was developed which allowed the taxonomists to create and edit a prototype angiosperm ontology by entering defined terms and creating the various relationships shown in Figure 4. The ontology, stored in a relational database, contains over 1000 defined terms (term + definiton + citation). There are 24 Regions, 46 Generic Structures and 269 Structures of which 126 are defined as Types. 160 optional Part-Of relationships organize the 143 remaining Structures into a structural hierarchy, with only 19 Structure Terms 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 group (typically 2). A print out of the structure hierarchy represented as an expanded tree, and the whole ontology in XML format can be viewed [15]. Each of the 536 structure nodes in the tree is identifiable by its path; of these 331 are leaf nodes.

The path of each structure (node) in the structure tree is programmatically calculated and stored in the database. This path represents the identity of each node when included in description elements. We are currently exploring the most efficient way to store this in the database, as an adjacency table of node ID versus parent node ID; as a programmatically parseable string representation of the path (e.g. 'ID1.ID2.ID3'; see Figure 5c); or as an XML fragment representing the path as nested structure terms.

An additional tool is under development that will generate Data Entry Interfaces automatically using the input domain ontology, which can be then be specialized to create project-specific proforma ontologies as described in Section §3.3. This will allow the taxonomists to specify which structures and properties that they wish to describe for a given set of specimens, and create an electronic 'proforma' for data input. Specimen descriptions composed with this interface will be saved to a relational database compliant with our Description Element data model, using terms which are unambiguously defined in our angiosperm ontology. We propose to evaluate these tools and the validity and value of our character data model by using the system to capture real specimen descriptions using our angiosperm ontology. We hope to investigate to what extent it is possible to expand our ontology to describe wider plant taxa, or whether the creation of new ontologies will be necessary to describe disparate plant groups.

5 Discussion

It has been suggested that by committing to a publicly available ontology different data sources can ensure shared meaning and compatibility [16]. However, even if two systems share the same vocabulary there is no guarantee that their data can be integrated unless the ontologies commit to the same underlying conceptualization [17]. The Prometheus conceptual model defines how 'characters' can be represented in a common format, thus allowing description data to be shared between conformant sources and possibly with data sources with schemas that can be mapped to this model. An important distinction between the Prometheus description model, and other electronic description formats for taxonomy, is that the Prometheus methodology does not require that 'characters' are defined before description, but that actual observations can be recorded at description time using an ontology of defined terms.

Ontologies can be used to mediate both structural and semantic data integration by representing a unified view of local ontologies, or by sharing a common ontology amongst data providers [13,14,18]. Here we propose that the consistent representation of characters according to our flexible model will ensure structural and syntactic homogeneity. We propose that the more problematic issue of semantic heterogeneity (including problems of synonymy and homonymy) can be resolved by the use of a single common controlled vocabulary specified as an ontology. Indeed all descriptions created using our common parent description ontology will be compatible. However, if different taxonomic domains require distinct description ontologies, descriptions composed using separate ontologies will not be automatically compatible without an expert mapping of the concepts between ontologies, possibly by mapping to a generic integration ontology. Such mappings are often problematic and inexact, and fail to resolve all semantic conflicts that can result in data loss [13]. The ability to share information with legacy data collected without a well-defined terminology will be severely limited. For these reasons the creation and adoption of description ontologies with as wide a taxonomic range as possible is desirable. However, the current individualistic working practices of taxonomists make acceptance and adoption of an 'imposed' standardized description ontology unlikely. Rather, we hope that by creating and successfully demonstrating the use and benefits of an ontology in one taxonomic domain we will encourage the adoption and bottom up development and expansion of the ontology.

Detailed botanical ontologies are being developed by other groups, particularly the Plant Ontology Consortium (POC) [19-21]. POC are developing highly detailed anatomical and 'trait' ontologies, initially for three scientifically well-characterized model species (rice, maize and *Arabidopsis*). 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 'Is A' relationship, which is somewhat analogous to our Type attribute/relationship for structures, but which can fully participate in 'Part-Of' hierarchies. We found that incorporation of a full 'Type Of' relationship into our structural hierarchy made the ontology overly complex, particularly to non-experts, nor is it easy to agree meaningful 'Type Of' relationships across a large taxonomic range. POC ontologies include an additional 'Derived From' relationship, which expresses developmental information currently not represented in Prometheus.

The POC trait ontologies define genetically-based traits, mutations, phenotypes *etc.* rather than taxonomic 'characters'. Furthermore, there is explicit linkage of traits to the Gene Ontology [20,22], which is inappropriate for the taxonomic domain, where there is typically virtually no genetic information available for specimens.

POC have adopted the Gene Ontology's 'True Path Rule' which asserts that child terms in the relationship hierarchy inherit the meaning of all of their parent terms, thus the definitions of all parent terms on the path of a term must apply to that term [22]. Within Prometheus the hierarchical path of a given structure also has critical importance in determining its absolute structural context. However, the relationships expressed in the Prometheus ontology are only *possible* contexts, an actual structural context is only asserted when a term is used in a description. This distinction allows a flexible ontology that can be used across a wide taxonomic range.

We believe that the specialization of our parent description ontology into individual proforma sub-ontologies is a novel means for facilitating the collection of compatible description data. We also believe that capturing the rich semantic content expressed in our ontology, for example the ontologically defined context of a structure via its path, allows not only efficient and consistent knowledge sharing and reuse but will also allow rigorous representation and analysis of taxonomic concepts.

Development of our 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. We have created an angiosperm ontology for the description of one taxonomic dataset and are extending it for the description of further test datasets. Providing tools that allow data entry using only a controlled defined terminology enforces semantic homogeneity, and will aid future integration of any database created using the tools.

References

1. Wilkinson, M.: A comparison of two methods of character construction. Cladistics 11 (1995) 297–308
2. Cannon, A., McDonald, S. M.: Prometheus II – Qualitative Research Case Study: Capturing and relating character concepts in plant taxonomy (2001)
   URL: www.prometheusdb.org/resources.html
3. Colless, D. H.: On ‘character’ and related terms. Systematic Zoology 34 (1985) 229-233
   Keogh, J.S.: The importance of systematics in understanding the biodiversity crisis: the role of biological educators. Journal of Biol. Educ. 29 (1995) 293 – 299
4. Diederich, J., Fortuner, R., Milton, J.: Construction and integration of large character sets for nematode morpho-anatomical data. Fundamental and Applied Nematology 20 (1997) 409-424
5. DELTA: Dallwitz, M.J.: A general system for coding taxonomic descriptions. Taxon 29 (1980) 41-46
6. NEXUS: Maddison, D.R., Swofford, D.L., Maddison, W.P.: NEXUS: An extensible file format for systematic information. Systematic Biology 46 (1997) 590-621
7. LUCID: Developed by Centre for Biological Information Technology: University of Queensland, Australia. URL: www.cpitt.uq.edu.au, www.lucidcentral.com

8. Davis, P.H., Heywood, V.H.: Principles of Angiosperm Taxonomy. Oliver and Boyd Edinburgh. (1963)
9. TDWG (International Working Group on Taxonomic Databases) URL: www.tdwg.org; Structure of Descriptive Data.: Subgroup Session Report at the TDWG Meeting in Frankfurt (2000) www.tdwg.org/tdwg2000/sddreport.
10. Prometheus URL: www.prometheusdb.org
11. Allkin, R.: Handling Taxonomic Descriptions by Computer. In: Allkin R., Bisby F.A. (eds.): Databases in Systematics. Academic Press London (1984)
12. The Science Environment for Ecological Knowledge: URL: seek.ecoinformatics.org
13. Cui, Z., Jones, D.M., O'Brien, P.: Semantic B2B Integration: Issues in Ontology-based Applications. SIGMOD Record 31 (2002) 43-48
14. Omelayenko, B.: Syntactic-Level Ontology Integration Rules for E-commerce. In: Proceedings of the Fourteenth International FLAIRS Conference (FLAIRS-2001), Key West, FL. (2001) 324-328
15. URL: www.prometheusdb.org/resources.htm
16. W3C: OWL Web Ontology Language Use Cases and Requirements.
URL: http://www.w3.org/TR/webont-req/ (2003)
17. Guarino, N.: Formal Ontology and Information Systems. In: Formal Ontologies in Information Systems. Proceedings of FOIS'98, Trento, Italy. IOS Press, Amsterdam (1998) 3-15
18. Sheth, A.: Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In: Interoperating Geographic Information Systems. M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds.), Kluwer, Academic Publishers, 1998, pp. 5-30.
19. Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., Clark, K., Teytelman, L., Cartinhour,S., Stein, L., McCouch, S.: Conference Review: Gramene: Development and Integration of Trait and Gene Ontologies for Rice. Comparative and Functional Genomics 3 (2002) 132-136
20. The Plant Ontology Consortium: Conference Review: The Plant Ontology Consortium and Plant Ontologies. Comparative and Functional Genomics 3 (2002) 137-142
21. The Plant Ontology Consortium: URL: www.plantontology.org
22. The Gene Ontology Consortium: URL: www.geneontology.org