Fundam. appl. Nematol., 1997, **20** (5), 409-424

Forum article

CONSTRUCTION AND INTEGRATION OF LARGE CHARACTER SETS FOR NEMATODE MORPHO-ANATOMICAL DATA

Jim DIEDERICH *, Renaud FORTUNER ** and Jack MILTON *

** Department of Mathematics, University of California, Davis, CA 95616, USA and ** 4, rue des Jardins, 17130 Montendre, France.*

Accepted for publication 16 December 1996.

Summary - The problems encountered in designing a large sized character set for a morpho-anatomical database are discussed. A new definition of the concept of taxonomic character is proposed where traditional characters are decomposed into a biological structure, a property of this structure and a state or value of this property. Properties are always taken from a short list of basic properties. Concepts to aid with the design of a character set are discussed together with specific guidelines for using these concepts.

Résumé – ***Construction et intégration de grands ensembles de caractères pour les données morpho-anatomiques des nématodes*** – Les problèmes rencontrés lors de la conception d'un grand ensemble de caractères pour une base de données morpho-anatomiques sont discutés. Une nouvelle définition du concept de caractère taxinomique est proposée selon laquelle les caractères traditionnels sont décomposés en une structure biologique, une propriété de cette structure et un état ou valeur de cette propriété. Les propriétés doivent toujours être prises dans une courte liste de "propriétés fondamentales". Les concepts qui aideront à créer un ensemble de caractères sont discutés, de même que des directives spécifiques permettant de traduire ces concepts dans la pratique.

Key-words : basic properties, biological characters, biological databases, data modeling, morpho-anatomy, schema design.

While computerized and on-line data are more available than ever, great care has to be taken in structuring data properly to avoid the typical problems of early binding, that is, paying for initial poor design by making future use and modifications more difficult and expensive.

There has been considerable discussion about the kinds of characters that are needed for identification or systematics, but far fewer articles have been published on the format that should be used for storing these characters in databases. Existing formats are often proposed with a particular application, or type of application, in mind, which may make it difficult to use the same data for a different application. The situation is even worse if we look beyond strict systematics/identification data. For example, nomenclatural data are often disconnected from descriptions, geographical distribution are often disconnected from museum collection records, and all of these data are often disconnected from bibliographic references. Other kinds of data such as physiological, ecological and biochemical data seem to exist each in a world of its own.

There exist various projects for "bringing together" existing or new databases such as "Species 2000", an initiative from International Union of Biological Sciences (Anon., 1996). More generally, there is considerable discussion in the biological community about general databases, standardized databases, etc., which was recently reviewed during a symposium (Fortuner, 1993*a*) and will not be reviewed again here.

In nematology, only limited character sets exist for morpho-anatomical characters, often in the shape of tables or compendiums, usually addressing only a few dozen characters for small numbers of species, typically at the genus level. While some efforts have made it possible to share data in terms of creating standards for a common format, this does not solve the problem of the logical expression of the characters, and some important questions and problems have not been addressed.

This paper discusses an organization of morpho-anatomical data that would ease the construction of large character sets, and enforcement of the standards we propose here should make integration of small databases perhaps not easy, but at least feasible.

** Correspondant du Muséum National d'Histoire Naturelle, Paris, France.

ISSN 1164-5571/97/05 $ 7.00 /

Goals

Kinds and use of data

Systematics uses many other types of data besides morpho-anatomical characters: ecological characters such as host-parasite relationships, physiological characters (*e.g.*, amphimixis *vs* parthenogenesis), biochemical characters (molecular systematics), and more. Creating a single database with a wide variety of character types would allow a more straightforward use of these data in particular applications. However, we believe that this could be detrimental in the long term when the data of a particular kind would be needed in other areas of research. Each type of character requires a specific database structure, and this article concerns morpho-anatomical data only. However, it will be shown in the discussion below that a similar structure can be used for other types of data, which should make it easier to define links between data. The morpho-anatomical data considered here belong to both traditional categories of quantitative and qualitative data, although the distinction is far from clear. Here, the terms "quantitative" and "qualitative" are used with the meaning of quantified (real numbers, integers) or descriptive (textual) data, respectively.

In building a database, it is important to keep in mind that various applications will use the data in many different ways. Consequently, it would be counterproductive to tailor the data to one purpose, such as identification *via* dichotomy, over other possible uses, such as identification using similarity or Bayesian probabilities (Horvitz, 1993). Also, the data should be useful as well for the whole of systematics, including alpha-taxonomy, classification, etc., and for disciplines other than systematics, including physiology, ecology and molecular biology. Therefore, whatever the source of the data, either extracted from published descriptions or supplied by biologists, it will be important to represent the data as faithfully as possible and, instead of first tailoring the data for a particular purpose such as identification, homology representation, ordination methods, cladistics, or any other application, we should aim at providing data in a format that can subsequently be used for different applications. Note that the concepts of state-based relationships (Diederich & Milton, 1991; Diederich, in press) and multiple sets of states (Diederich, in press) aid in achieving these goals. The actual uses of the data is outside the scope of the present article, but this topic will be briefly addressed in the discussion section below.

Expressiveness and uniformity

In the construction of large character sets, two main goals seem to conflict: expressiveness *vs*. uniformity of the data. Existing taxonomic characters are usually expressive, but far from uniform, as each author has placed his or her own stamp on the taxonomic descriptions found in the printed literature, which have been created over the decades without predefined or agreed-upon standards. DELTA (DEscriptive Language for TAxonomy), proposed by Dallwitz (1980), is widely accepted as a global standard for taxonomic descriptions, in particular by TDWG (Taxonomic Databases Working Group), which develops standards for taxonomic databases (Bisby, 1994). However, DELTA does not address the problem of uniformity of data and gives complete freedom to each author for using any form of character. Uniformity in DELTA refers only to the coding method, not to the characters themselves. Other proposals for data representation have been made, but much more can be done to achieve uniformity. One of the reasons for this situation is that making the data uniform generally makes the data less faithful to what was represented in the descriptions. Naturally, the reason for having uniform representation of the data is that it makes the data easier to manipulate. In the electronic world the more expressive the representation the more complex the software is that is needed to support as well as use it. Our general approach to handling these conflicting goals is to view traditional characters in terms of their constituent parts and to formulate guidelines for constructing characters based on several new concepts.

The concept of taxonomic character

The word "character" has many different meanings in taxonomy (Colless, 1985), but the general idea is: a character is a characteristic that can be used to differentiate, classify, or identify taxa. Taxonomic descriptions usually record characters, but a few non-differentiating characteristics may also be recorded for descriptive purposes.

With respect to the nature of the words found in character names, there are two main ways the word "character" is used in taxonomic literature. One represents a general concept such as bulb shape or stylet length, *i.e.*, an organ name (bulb, stylet) and an abstract concept describing this organ (shape, length). The domain of possible qualitative states, such as (round, oval, square), or the range of possible quantitative values, such as (2.5-8.5 µm) is kept separate from the character and is called "character states" or "character values", respectively. The other combines the above abstract character and the character state or value into a single unit, confusingly also called "character", as in "bulb round" or "stylet 5.5 µm". In this case, the concepts "shape" and "length" are implicit.

When taxonomists use the word "character", it is not clear whether they refer to one or the other mean-

ing. Taxonomic descriptions also may include both forms as in "tail length=58 µm" and "pharyngeal bulb elongated"; the first example clearly separates the value (58 µm) from the character, the second does not. However, both forms are treated as single units.

THE LIMITATIONS OF TRADITIONAL CHARACTERS

This lack of uniformity between (organ, state) and (organ, property), as well as the use of any characters ranging from simple to highly complex, may require the user to determine which form is to be used in accessing the data as well as require much more software coding in the application programs.

When a character is combined with a state into a single unit, flexibility is reduced. Making "isthmus thin" a unit and "stylet slender" another unit, for example, makes it more difficult to state that thin is synonymous with slender, so that synonymy is independent of the structures having this property. While it may be obvious to a nematologist using the system that "lips rounded" and "lips hemispherical" mean the same (because nematodes are transparent and a three-dimensional structure such as the hemispherical lip region is seen under the microscope as a two-dimensional outline), how would the software detect and use this information properly in a similarity computation if rounded and hemispherical are not somehow indicated as synonymous, independent of the structures they describe? It would also be more difficult to program the system to determine that lips had the property "shape" using "lips rounded" as a unit. Thus, building a knowledge base in which structures and properties are combined, rather than independent, would be more difficult. The situation is even more problematic with complex characters.

In the case of simple characters such as "lips rounded", the organ and the state are easy to recognize, but this might not always be the case. A single character may be a very complex object, grouping several biological structures, concepts, and states, *e.g.*, "cuticle of female and juvenile with outer layer, if present, thin, membranous, closely adpressed" (Raski & Luc, 1987) which refers in fact to two structures- the two cuticular sheets composing the cuticle- and several concepts: presence or absence of the external sheet, relative thickness of the two sheets, and their position relative to one another.

Besides description and systematic analyses, one important use of morpho-anatomical characters is differentiation of taxa. This requires that characters be described in the same manner for all the taxa considered at a particular level, for example, for all the species in a genus: species X has rounded bulb, species Y has square bulb, or, for the genera in a family: genus W has deirid present, genus Z has deirid absent. Comparisons for identification purposes often follow a dichotomous format: if an organism has a round bulb, it may belong to species X but not to species Y. The way traditional characters are recorded corresponds well to this traditional use.

However, there are many cases when dichotomous comparison may not be the best approach (Fortuner, 1993*b*), and other comparison methods should be used for which the traditional recording of characters is no longer suitable. For example, it would be very difficult to build a routine to calculate a coefficient of similarity between a specimen with "cuticle of female and juvenile with outer layer, if present, thin, membranous, closely adpressed" and a species with "cuticle of female and juvenile with outer layer, if present, thin, membranous, *not* closely adpressed".

In sum, managing disparately described traditional characters would be difficult, if not impossible, because they appear as an agglomerate of several types of information (one or several organ names and states or values), because important information is often missing (the abstract concept being described), because they are linked to particular taxa as described by particular authors, and because they do not share a common format or have any natural arrangement. Traditional characters are like collections of dots in a painting by Seurat. Together they make a picture, but Seurat himself would have had a hard time finding where he put all the red dots in *Dimanche d'été à la Grande-Jatte*.

A NEW APPROACH FOR THE MANAGEMENT OF CHARACTERS

We propose managing characters by formally viewing traditional characters in terms of three constituent parts, then employing very strict guidelines in the context of several new concepts.

The first element is called a structure, representing any part of the organism, from the whole organism itself down to cell organites and molecules. We choose to use the term structure because of the morphological setting, though others in computer science and biology have used the terms entity or object. The second part we call a property, *i.e.*, the concept or aspect of this structure that is being described. Others have used the terms attribute, trait, quality, aspect, etc. The third part consists of states (descriptive) or values (numerical) as found in the various taxa. For example, the character "lips rounded" is composed of a structure, Lips, a property (implied), shape, and a character state, rounded. Such decompositions can be found in other studies (*e.g.*, Lebbe, 1991), but here we propose to enforce strictly the separation between structures, properties, and states according to several guidelines that will be discussed below. By contrast, Lebbe (1991) decomposed what he called "descripteurs" (which are the traditional characters) into an

entity he called "subject" (similar to our structures) and an attribute he called "quality" (similar to our properties), but without enforcing rules or guidelines. One of his descripteurs "number of teeth of the pod" is decomposed into a subject, "pod", and a quality, "number of teeth", which includes a biological structure. We would decompose this character into two biological structures, "pod/teeth", and a property, "number".

Biological structures are arranged in a natural hierarchy of systems, organs, tissues, cells, and cell organites, each subdivided as finely as needed. This hierarchical organization is familiar to biologists in concept, though actually formulating an organization can be problematic.

Once the biological structure that is being described is identified, it becomes possible to infer what property is being described by considering the related character states. In "lips rounded" for example, the obvious property is shape, in "outer layer thin" it is thickness, and in "body 500 µm long" it is length. However, it is not always easy to determine what the property is, and in some cases it does not exist as a concept in the field. Still it is possible to handle this situation in a straightforward manner, as we shall indicate.

The NEMISYS character set

This article is based on work in the NEMISYS (NEmatode Identification SYStem) Project, an effort to create a morpho-anatomical database for over 4000 plant-parasitic species, with perhaps an additional 6000 species of other types to be added later. Nematode taxonomic descriptions are quite complex since they include anatomical characteristics in addition to external and internal morphological ones, and nematode systematics is based on a very wide range of characters. Addressing problems in this very complex domain should yield concepts and principles which would prove to be helpful to others with similar interests. This has recently led us to launch the GENISYS (General Identification SYStem) Project, which is understood as an expansion of NEMISYS.

The structure of the character set we constructed for the nematode order Tylenchida includes 272 biological structures with 797 properties, which would represent well over 5000 traditional characters if states were included, and would be larger still but for various methods we use to consolidate characters. This set is already considerably larger than we had expected it would be when we began to construct it, and it is the largest set of characters for a biological database that we have encountered. Size considerations may have some implications for the care needed in this type of endeavor in general and for building biodiversity databases. So far, this character set has been populated with only a few dozen test taxa.

We have no reason to believe that character sets for other kinds of organisms would be significantly smaller. Nematodes are often qualified as being a "difficult" group compared to other organisms, because both internal and external organs are used in identification of the transparent nematodes instead of only the external organs in many other groups. However, we indicated that a morpho-anatomical database should not be constructed for identification alone but should include all existing organs that will be needed in other applications. Compared to the entire set of biological structures, from systems to cells and organites, of, *e.g.*, *Homo sapiens sapiens*, nematodes are actually rather simple animals.

Problems in the design of character sets

Given the representation of character introduced here, we now examine some of the problems that can arise in creating a set of characters. The problems we identify below are not problems simply because they violate some arbitrary standard that we have in mind. Indeed, they present very practical difficulties in designing, using, and integrating databases. Some of these problems are immediately clear to the designer. However, some remain hidden until later in the lifecycle of the database. While our concept of character helps eliminate some of these difficulties, additional concepts are required, together with the guidelines defined below.

The representation of characters is not as straightforward as it might initially seem, even at the structural level, and even with the concept of character we propose. For any living entity there can be many kinds of representations, corresponding to multiple points of view. One way to look at structures is based on containment, or structures that contain substructures, *e.g.*, the dorylaimid oesophageal bulb contains glandular and muscular cells. Another is regional, such as structures in the head, tail, or mid-body. Another is functional such as the digestive system, the reproductive system, the nervous system, etc. Still another is by physiological function, such as contractibility, which would include the stylet and vulva muscles in addition to the somatic muscles. When a biologist is asked to define the list of structures for a particular group of organisms (at the phylum or order level), he or she should be asked to arrange these structures according to the best known representation: the plan of organization of the organisms according to systems. However, the other points of view should also be accommodated, so that, *e.g.*, the digestive system muscles (and the properties attached to these structures) should appear in both the digestive system and

Table 1. *A standard decomposition.*

Structure	Substructure	Property	States
Body			
	Lateral fields		
	Deirid		
	Annuli		
		Width	
		Orientation	
			symmetrical
			retrorse

the muscular system *via* an interface for viewing characters.

Even within a single logical representation, the decomposition of traditional character can create some problems. For example, let's consider three structures that are present on the Body: Deirid, Lateral fields and Annuli, each with its own set of properties, including width and orientation of the annuli. A straightforward decomposition is shown in Table 1*, where properties have been indicated only for annuli. However, there are some criconematid species in which the orientation of the annuli in the anterior part of the body is different from that in the posterior part of the body. There are many possible ways of handling this decomposition, and the choice could certainly affect how well a system works and how easy subsequent integration of existing systems might be. Table 2 shows some possible decompositions.

We can easily find grounds to criticize or support any of these possible decompositions. With Table 2A we have created two additional substructures out of "body". Though this seems reasonable, it causes one to duplicate some substructures and properties under both the "anterior part of the body" and the "posterior part of the body," as we see with "lateral fields" or "width". While this duplication is not a significant problem for displaying the substructures, it would present some difficulties in accessing and storing the data, as discussed in the following paragraphs. Note that the decomposition in Table 2A has an advantage in that a substructure such as "deirid", which is present in the "anterior part of the body" can be properly placed and not duplicated.

With the decomposition in Table 2B we get the same problem of duplication of substructure or property for the annuli. That is, either they would have to be duplicated under both "annuli in the anterior part" and "annuli in the posterior part", or a separate superstructure "annuli" would have to be created. In either case, the solution is awkward in terms of managing the character set as well as in accessing and storing the data. Furthermore, "deirid" cannot be handled as well as it is in Table 2A, and if we simply place it under "body" as a substructure it would not be properly differentiated as belonging to the anterior part of the body.

The next alternative, Table 2C, certainly solves the problem of duplication of the property "width" encountered in Table 2A and Table 2B, but it brings to light another one. The property is no longer "orientation", but it is two properties, "orientation in the anterior part" and "orientation in the posterior part". This would prove quite problematic in accessing the database. If one accessed the data using the condition "annuli, orientation=retrorse", it is likely that data would not be properly retrieved. Either the system would not know that "orientation in the anterior part" was indeed subsumed by the name "orientation" or it would not know that it really needed to access the data using "annuli, orientation in the anterior part=retrorse" and "annuli, orientation in the posterior part=retrorse". Moreover, even if there were a character processor to modify the condition "annuli, orientation=retrorse" to handle this situation it would be more difficult to build than it should be, as will be indicated below.

It seems that Table 2D is the proper approach since it requires no changes in the structure of the character set, thus avoiding the problems of Table 2A-C, except for "deirid" placement, and it only requires adding one state to the set of states. However, with a condition like "annuli, orientation=retrorse", should data be retrieved that had states "symmetrical anterior and retrorse posterior" and how would the system know to do this? Certainly this can be handled, but it needs to be done as systematically and as seamlessly as possible to avoid having to deal with individual cases in different characters. While Table 2D is the easiest to implement, it is not clear that queries will be handled any better than with the other alternatives.

There are other subtleties involved in the decompositions in Table 2A-C. For species that have the same annulus orientation all along the body (which happens to be the case for the majority of nematode species), we would have to record the data twice, once for

* In the Tables the structures and substructures are always capitalized, with substructures below and to the right. A property is shown in lower case below and to the right of its (sub)structure, with its states listed in a column below and to the right of the property. Any structure except Body can be a substructure of another structure and any substructure can have substructures under it. Finally, note that the examples in the various Tables do not reflect the entire hierarchy of nematode structures, but only those structures that are relevant to the various examples. This explains the differences in the decompositions presented, *e.g.*, in Tables 5B and 8B.

Table 2. Alternative decompositions.

Structure	Substructure	Property	States
A Body	Anterior part of the body		
	Lateral fields		
	Deirid		
	Annuli		
		Width	
		Orientation	
			symmetrical
			retrorse
	Posterior part of the body		
	Lateral fields		
	Annuli		
		Width	
		Orientation	
			symmetrical
			retrorse
B Body	Lateral fields		
	Deirid		
	Annuli in the anterior part of the body		
		Width	
		Orientation	
			symmetrical
			retrorse
	Lateral fields		
	Annuli in the posterior part of the body		
		Width	
		Orientation	
			symmetrical
			retrorse
C Body	Lateral fields		
	Deirid		
	Annuli		
		Width	
		Orientation in the anterior part of the body	
		Orientation in the Posterior part of the body	symmetrical retrorse
D Body	Lateral fields		
	Deirid		
	Annuli		
		Width	symmetrical retrorse
		Orientation	symmetrical anterior and retrorse posterior

the anterior end and once for the posterior end, and be able to deduce from this that it was the same in both ends when doing a similarity computation, for example. This also explains why we have listed "retrorse" and "symmetrical" with both the anterior and posterior parts even though only one applies to each part in the case of most species.

The main point that this series of examples demonstrates is there are many seemingly reasonable decompositions, and given these alternatives it could be quite difficult to maintain a consistent and uniform decomposition that could be exploited easily by the system to access and manipulate the data properly. Also, the implications of the choices made are not always obvious, and the algorithms that would have to be built into the system to handle all of the different choices can create serious difficulties. Furthermore, this clearly illustrates the difficulties that would exist in trying to integrate two or more of even the smallest of character sets.

Biological character design

The concept of character that we have provided is in and of itself insufficient for creating a good set of characters. One can still formulate a poor character set while adhering to this approach of character representation. Mechanisms are needed to aid in maintaining and possibly enforcing the ideal and in maintaining other principles that emerge from the guidelines we offer. Such mechanisms are directly analogous to the existing tools and principles that aid in designing any good relational database. Several concepts, including basic properties and name extensions were introduced by Diederich (in press). That paper introduces these concepts formally, and here we focus on how these ideas can be properly used, even if the concepts are not supported in systems used by biologists to represent characters and build character databases. Our primary aim is to provide assistance in the use of these concepts to create a more consistent and uniform set of characters. We first briefly describe these concepts before discussing their use in creating character sets.

Concepts for character set creation

In examining numerous realizations of early versions of a set of nematological characters, which exhibited all of the problems presented above, we discovered that by enforcing a very strict separation between structure and property, most of the properties used in early character sets belonged in fact to a short list of properties, the basic properties. This helped eliminate many of the problems that previously affected the list of characters. Every structure could then be described by a few properties, almost always taken from these basic properties. We have not previ-

ously seen consideration of such a strict separation between structures and properties in creating characters.

Each set of basic properties, such as the set seen in Table 3, is associated with a type of data, morpho-anatomical data in this example, rather than a particular category of taxa such as nematodes. Nothing in the set in Table 3 indicates that these basic properties are tied to any particular group of organisms. They could be used for descriptive data about fish, birds, various insects, plants, and more. It is very likely that there is another set of appropriate basic properties for physiological databases, again independent of the group of organisms, as well as another set of basic properties for ecological databases, biochemical databases, and so on.

Table 3. *Basic properties for morpho-anatomical data (* : relational properties).*

Appearance	Dimension	Placement/ Location	Quantity
Posture	Length	Position relative to*	Presence
Shape	Height	Distance to*	Quantity
Kind	Width	Orientation	Number
Texture	Diameter	Angle	
Arrangement	Depth		
Symmetry	Ratio of* size		

As seen in Table 3, the basic properties for morpho-anatomical data are grouped into four broad categories: Appearance, Dimensions, Quantities, and Placement. Basic properties within a category tend to have the same characteristics. For example, every basic property has a specified default range taken from the set (binary, discrete, continuous) and scale taken from (nominal, ordinal, interval, ratio) (Zar, 1996).

Some basic properties come with a predefined set of states as well, automatically specified in the definition. For example, the basic property "presence" obviously has the states (present, absent). Characters themselves have properties, which are data about data, or metadata for short. For each property, metadata such as range and scale are also included in the definition.

In some descriptions the authors provide actual measurements for structures while others may simply state that the structure is "long" or "short", using a qualitative value for an intrinsically quantitative character. These states can be considered as fuzzy states, *i.e.*, they are not actual measurements, but suggest a possible range of measurements for the species described, it being understood that an expert will know the meaning. In creating a set of characters this can be a problem since we would need two properties for each measurement. Thus "length" could be used for numerical values and "fuzzy length" for fuzzy states. This is quite artificial, though, and increases the number of characters considerably. Consequently, we bundle the fuzzy states into each measurement's basic property. Thus, "length" comes with the fuzzy states (very short, short, intermediate, long, very long) and "width" has the fuzzy states (very narrow, narrow, intermediate, wide, very wide). The records for storing data would contain fields for fuzzy states as well as measurements. The interaction of the measurements and their fuzzy states is the subject of further work. Suffice it to say that with the addition of fuzzy states it is easier to construct and manage the set of characters and to acquire the dataset itself. Quantities and their fuzzy states are handled in a similar fashion with fuzzy states (a few, several, many, about a dozen, etc.) included with basic properties that are quantities.

Naming is a tricky business in any kind of project, far more of a problem than many realize or appreciate. Avoiding this problem explains why defining "stylet straight" as a character is so desirable, since it may be difficult to determine what the property is, if any. In our initial character set creation, a variety of artificial properties were used such as "nature", "aspect", "type", etc. Complications began to mount when a structure had more than one such character. To simplify this, we selected one generic property, "kind", and we allowed multiple sets of states to be listed within this property (multiple sets of states are also allowed for other properties such as "shape"). An example is the structure "lateral field lines" and the property "kind" for which there could be two sets of states, *e.g.*, (indistinct, faint, distinct) and (smooth, wavy). Thus a single property can incorporate multiple sets of states. One could argue that there should be two properties, "distinctness" and "smoothness". However, turning states into properties, *i.e.*, "smooth" → "smoothness", may not be the best approach since one could alternatively make the property "waviness" or any other property derived from one of the states, which is clearly a less uniform way of doing this.

It should be noted that some basic properties, called "relational", cannot stand alone but require a more complex name when used to form a character. These properties are indicated by an asterisk in Table 3. For example, "distance to" does not mean anything in itself since it indicates how far the structure being described is from another structure or a landmark. A landmark is a characteristic point of the organism that is used as a reference point for describing a structure. Structures often serve as landmarks. A landmark is not a character or a property, but its name often

appears in the name of traditional characters. Thus, in the character "hemizonid, distance to the excretory pore" the structure "Excretory pore" is used as a landmark relative to the "hemizonid".

Guidelines for character set creation

In this section we present several guidelines in decomposing characters consistent with the definition we have given. One could call them rules rather than guidelines as long as it is understood that rules always have exceptions. The spirit of the guidelines is taken in this way, that is, any violation of a guideline should be considered a very serious matter. We assume that basic characters with their properties, including name extensions, fuzzy states, and the like are available. (They facilitate this process considerably, but these guidelines alone should aid in creating a more uniform and consistent character set for anyone creating a set of characters, independent of whether or not their system explicitly supports basic properties.)

General guidelines

The first and most fundamental guideline for creating a character is:

Guideline 1. *Follow the ideal of the decomposition of a character into three parts: a structure that is part of a specimen such as "body", a property that is the abstract concept that is being described, such as "shape", and a state or value such as "round".*

In general, structures should not contain descriptive or qualitative terms. Properties should not contain structural or state-oriented terms. States should not contain structural or property-oriented terms. Exceptions should be well understood and uniformly applied.

Guideline 1 is easily stated but not always so easily followed. It can play an important role in detecting when a character might have been poorly formulated. In Table 4A, the property "tip shape" is not a pure property, but a combination of the structure "tail tip" or "tip" and the property "shape". Allowing properties to be logical combinations of structures and properties results in a free-for-all in creating a set of characters and makes it much harder to integrate databases and to manage, modify, and use a set of characters effectively. Table 4B gives a decomposition in agreement with Guideline 1. In Table 4C, structure names appear in states. Again, this can affect integration, but there is another immediate practical problem. Suppose we simply wanted to indicate that "posterior" and "below" are synonyms. In any set of states that used "posterior to <structure name>", you would have to indicate synonymy with "below <structure name>", and this would have to be done for each structure name, rather than simply indicating the synonymy between "posterior" and "after", independent of the structure. Note that the same problem arises with "tip shape" in Table 4A if one wished to indicate synonymy between "tail tip" and "tail end". An additional problem arises for instance with a char-

Table 4. *Use of Guideline 1.*

Structure	Sub-structure	Property	States
A Tail			
		Shape	
			filiform
			conoid
			cylindroid
			broadly rounded
		Tip shape	
			filiform
			conoid
			cylindroid
			broadly rounded
B Tail			
		Shape	
			filiform
			conoid
			cylindroid
			broadly rounded
	Tip	Shape	
			filiform
			conoid
			cylindroid
			broadly rounded
C Excretory pore			
		Position	
			anterior to median bulb
			at median bulb
			posterior to median bulb
D Excretory pore			
		Position relative to - median bulb, nerve ring	
			anterior=before= in front of at=just at posterior=after =behind

acter such as "in species X, the median bulb is anterior to the excretory pore". How would the system know that this is the same as "excretory pore posterior to median bulb"? A better approach would be to separate the positional terms like "anterior" from the structure name, and use the general fact that "A posterior to B" is equivalent to "B anterior to A" for any structures A and B. The actual property would be "position relative to #" in which the position of the structure being described (structure A) is given in relation to that of another structure or landmark (structure B), indicated by the # sign. The person entering data would have the possibility of replacing # by any structure from the list of structures defined for a particular group of organisms. The proper decomposition for Table 4C is shown in Table 4D using name extensions for a relational basic property. (A name extension is a form of data that semantically serves as a modifier for either a structure or property name; name extensions are indicated within {} after the structure name.) Note that the states form a standard set of states with a standard set of synonyms that are part of the basic property itself, saving the designer time in creating the character set. If structure names were allowed in the state names, this would be more complicated, as each set of states would have to be hand tailored with special handling required by any application programs that use the character set.

Basic properties play a key role in following Guideline 1, either as a supporting tool in the design system or as a conceptual device for forming characters, which leads to the next guideline.

> **Guideline 2.** *Whenever a character is created, its property should be selected from among the list of basic properties.*

The way we view creating characters is that, once a structure is properly identified using our guidelines, its properties will be directly selected from the set of basic properties. Ideally, if this guideline were followed for a given structure, it is unlikely that at the property level there would be any problems of the type discussed in the previous section. Table 5A is an example of "property explosion" that comes from ignoring this guideline and creating many different characters out of what are in fact different views of the same general character. In addition, the qualifier "only" is contained in the name of some properties, making the semantics more complex. One would need to add the property "presence elsewhere than on tail" to handle cases where areolations are found on body but not on tail; "areolations" are in fact substructures of the lateral fields. The lateral fields generally run along the whole body. They are often composed of lines, and the spaces between the lines are called bands, which can have transverse striae, called areolations. The solution is shown in Table 5B, where name

Table 5. *Use of Guideline 2.*

	Structure	Substructure	Property	States
A	Areolations		Presence on whole body	
				present
				absent
			Presence on neck only	
				present
				absent
			Presence on tail only	
				present
				absent
			Presence at vulva	
				present
				absent
			Situation on outer bands only	
				present
				absent
			Situation on all bands	
				present
				absent
B	Lateral fields - (on body, on neck, at vulva, on tail)			
		Lines		
			Number	
		Bands - (outer, inner) Areolations		
			Presence	
				present
				absent

extensions have been used for positional qualification of the Lateral fields and their Bands.

There is a further problem in Table 5A, as "situation" has been used instead of "presence", a typical kind of inconsistency in character sets. Related to the latter problem, some basic properties are used in certain biological groups under different names. For example, botanists call phyllotaxis or foliation the arrangement of leaves along the stem. These terms

need not be used to create new, ad-hoc properties, but can be entered as synonyms of the existing basic property "arrangement". However, there will be times when it is necessary to add new basic properties. For example, "color" is not a basic property in Table 3 because nematodes are colorless, but it will need to be added when the system is extended to other biological groups such as birds. The definition of basic property given by Diederich (in press) gives a rationale for standards to follow in creating new basic properties. Table 5B solves all of these problems by a judicial use of basic properties and name extensions.

Guidelines for non-relational properties

Naturally there are exceptions to Guideline 1, and they will be discussed within the remaining guidelines. Guideline 2 is most important in dealing with non-relational basic properties, *i.e.*, those basic properties that do not inherently relate characters. This type of property is sufficiently important to warrant its own guideline.

Guideline 3. *If a basic property is non-relational, then generally speaking it should be used as is, without modification.*

Exceptions are:

- synonyms may be added to the name of the character to clarify its meaning. For example, if "presence" is the basic property used, and in some instances it is referred to as "situation", then "situation" could be added as a synonym.

- any non-structural or non-descriptive term may be added to the basic property name, often best added as a name extension, whenever there is a possibility of additional name extensions. For example, in Table 6A, "length along the axis" clarifies the meaning of the "length" of the stylet, and since the length might be measured in other ways, such as along the dorsal edge or from tip to tip, "along the axis" is best added as a name extension (Table 6B). Similarly, the various levels along the nematode body where the diameter can be measured (Table 6C, another example of property explosion) are best added as name extensions (Table 6D)

- if a structural reference is used as a name extension in a non-relational property, as in "diameter - (at the vulva, at mid-body, ...)" (Table 6D), the referenced structure should be selected from existing structures in the list of characters. In addition, the basic property should not represent a property of the referenced structure.

To clarify this last point, the guideline rules out properties such as "length of the stylet" since "length" is a property of the referenced structure "stylet". However, in the case of "body, diameter at the vulva" (Table 6C, D), "diameter" represents a property of the "body" and not of the referenced structure "vulva".

Table 6. *Use of Guideline 3.*

Structure	Sub-structure	Property	States
A Stylet	-		
	-	Length along the axis	
B Stylet	-		
		Length - {along the axis}	
C Body			
		Diameter at mid-body	
		Diameter at stylet	
		Diameter at vulva	
		Diameter at anus	
D Body		Diameter - {at mid-body, at stylet, at vulva, at anus}	

Note that whenever a basic property is used to form a character, we call it an instantiation of a basic property, and the instance (of the property) can then be qualified within the character. However, if it is qualified then it is likely that one of the guidelines is being violated and the situation should be examined and understood, that is,

Guideline 4. *Whenever there are modifications to an instantiated basic property, that should signal possible problems with the character.*

Non-relational properties that contain structural or state-oriented terms should be avoided, as they generally violate Guideline 4. For example, "tail, tip shape" (Table 4A) violates the Guideline 1 and uses the property "tip shape" that includes a structure "tip". "Tip" is a substructure of the "tail" and "shape" is its property. Clearly Guidelines 2 and 3 have been violated since "tip shape" is not a basic property and "shape" is a non-relational basic property which has been modified to create a character. These suggest problems exist with the character and should be carefully considered. The structural term "tip" should be represented as a substructure of "tail" to form the character "tail tip, length" (Table 4B).

Guidelines for relational properties

As discussed above, basic properties such as "distance to" and "ratio of" will necessarily reference other characters or other structures. The latter often represent landmarks or structures used as landmarks.

Guideline 5. *Relational properties should reference other characters* via *name extensions.*

Some relational properties are rather straightforward to handle. For example, the traditional character "distance from excretory pore to anterior end" is easily represented by "excretory pore, distance to - (anterior end)". At the opposite end of the spectrum, ratios can be complex, and the traditional "ratio o" in nematology refers to the ratio of two properties and involves no less than three structures: it represents the distance between opening of the dorsal oesophageal gland and stylet base divided by the stylet length.

Before continuing with more guidelines, let's examine how name extensions may be used to handle problems encountered in the structural decomposition. We have already seen name extensions used in Table 4D. The problem shown in Table 2 can be solved (Table 7) effectively by adding the name extensions "-anterior part" and "-posterior part" to the structure "body".

***Table** 7. A proper decomposition for Table 2.*

Structure	Substructure	Property	States
Body - {anterior part, posterior part}			
	Lateral fields Deirids Annuli		
		Width Orientation	
			symmetrical retrorse

While Table 7 seems almost identical to Table 2A, there are some significant differences. First, the natural decomposition is maintained with no duplication of properties or substructures. Second, name extensions can be added whenever they are needed in character sets that support them, without changing the structure of the set, the one clear advantage of Table 2D. Third, and perhaps most important, it is relatively easy to create a single algorithm to detect the existence of name extensions and either to enforce them or not, as a condition in accessing the data. The choice can be left to the user or can be set by default. For example, the condition "body, annuli, orientation=retrorse" with no condition on the name extension would retrieve all species that had "orientation =retrorse", regardless if the retrorse annuli were located in the anterior part or not. On the other hand, the added condition "and body.name extension=anterior part" would retrieve only those candidates for which the extension were stored with the data. While the use of name extensions would require a subsystem to process a list of characters prior to accessing the data, so would all of the other alternatives in Table 2. However, in this case it would be easier and would be a more systematic approach to doing so. Name extensions allow a bit more expressiveness while promoting a uniformity and consistency of expression. Also note that the data is only stored once for those species that have the same annulus orientation in both ends of the body, *i.e.*, the name extension data field is left blank.

Note that one might be tempted to add the name extensions to the substructure "annuli" rather than to "body". However, to be correct semantically one would have to add "- anterior part of the body" as the name extension, for otherwise "- anterior part" would appear to be the anterior part of the "annuli", which is clearly not intended. This leads to the guideline:

Guideline 6. *A name extension used with a structure should qualify only that structure and avoid references to other structures.*

Name extensions are useful whenever there are similar substructures with the same properties. For example, some dorylaimid species may have multiple supplements that are numbered 1, 2, 3, etc. One could then have a single structure "supplements" with name extensions 1, 2, 3, etc. This example and the previous example of positional qualifiers, *i.e.*, anterior part, leads to the guideline:

Guideline 7. *Use name extensions for repeating multiple identical substructures and for positional qualifications of substructures.*

The degree to which one violates the condition "multiple identical substructures" may indicate whether name extensions should be used or not. For example, if a nematode has six lip sectors that are not identical in shape, then it may not be obvious that name extensions should be used. Contrast this with the case of the lateral lip sectors where the right and the left sector are the same except for position. In this case a name extension would be appropriate. However, one would not want to have "lip sectors" as a structure with name extensions "-subdorsal left", "-subdorsal right", "-lateral left", "-lateral right", "-subventral left", "-subventral right". There are several reasons for this. Clearly, these are not obviously substructures of a single logical grouping whereas "subdorsal sectors", "lateral sectors", and "subventral sectors" are three logical groupings, each having its own set of positional name extensions. Also, a judgment should be made on whether the substructures are generally processed separately or not and whether making a distinction is normally important. This leads to the next guideline.

Guideline 8. *Use name extensions for repeated substructures of a logical grouping that are not identical,*

where distinctions are the exception rather than the rule.

Note that the name extensions "on body, on neck, at vulva, on tail" in Table 5B could have been attached to both "lines" and "bands". This obvious duplication would be unnecessary and suggests an additional guideline:

Guideline 9. *Put the name extensions as high up in the hierarchy of structures as possible.*

Note that we do not say that allowing alternate views of the list of characters is not desirable, as they should be supported by software that displays the character set. For example, all structures where areolations appear could be shown, but these alternate views should be built on top of as uniform and consistent a set of characters as possible.

The example in Table 8A can be handled with bands and ridges as substructures of lateral fields. We would also support decomposing the states into separate state lists {low, high} and {separated, contiguous}. But is easy to overlook the fact that certain states are indeed fuzzy states for certain measurements. We have seen this repeatedly, and it is illustrated by this example, where {high, low} are really fuzzy values for the property "height" rather than for some other property such as "kind" (Table 8B). One could argue that "low, contiguous" is a valid state and is best not decomposed. One simply needs to keep in mind that there are tradeoffs with respect to queries and synonyms as pointed out before. Also, it would be easier to build a "summary character" (Diederich & Milton, 1993) out of properly decomposed basic properties and simple states then to decompose a complex state, once it is entered as such in a database. Because it is all too easy to overlook fuzzy states we propose two guidelines for states.

Guideline 10. *Each set of states should indicate a single type of information reflected by the name of the property.*

Guideline 11. *Each set of states should be examined as potential fuzzy states and included with the appropriate instance of a basic property.*

Table 8C is best handled with two properties: number and kind (Table 8D). Guideline 10 is also a matter of judgment in terms of the level of granularity of the states. Table 8C represents two kinds of information in the state: the number of branches and the kind of branches, from two branches equally developed to one branch reduced to a post-uterine sac, just as Table 8A had two kinds of information about bands and ridges. Table 8D presents a decomposition in agreement with the above guidelines.

There are certainly other concerns in building a set of characters. In particular, there is a tendency to embed information in structure names, property names, and states that would be best represented explicitly. Some of these concerns have been presented in regard to state-based relationships (Diederich & Milton, 1991). While a set of guidelines could be helpful in this regard as well, up to this point

Table 8. *Use of Guideline 10 and 11.*

Structure	Sub-structure	Property	States
A Lateral fields			
		Kind	
			low bands, contiguous; high ridges, contiguous; high ridges, separated
B Lateral fields			
	Bands		
		Height	
			low
			high
		Arrange-ment	
			contiguous
			separated
	Ridges		
		Height	
			low
			high
		Arrange-ment	
			contiguous
			separated
C Genital branches			
		Number	
			2, equal 2, post reduced 1 + PUS 1
D Genital branches - {anterior, posterior}			
		Number Kind	
			developed reduced reduced to a PUS

we can say that following the guidelines we have presented will alert the designer to potential problems that can be analyzed and resolved.

These guidelines are based on a retrospective examination of what we did in creating our nematode character set, along with some of the ideas that lead to the development of basic properties and how they are defined. They are designed to help with the creation of a new character set for a particular biological group, or with adding new structures to an existing character set. Refinements or additions may be needed in the future, as there will always be situations that we have not yet encountered, but we believe that following these guidelines will yield much more uniform and consistent character sets for the increasingly challenging applications that demand such data.

Discussion

QUALITATIVE *VS* QUANTITATIVE CHARACTERS

The present work was done without always adhering to the traditional classifications of characters into, *e.g.*, quantitative *vs* qualitative, or discrete *vs* continuous. It is true that each of our basic properties has default range and scale (Zar, 1996), which are entered as metadata and can be used to place each piece of data into a particular category, if needed. On the other hand, the concept of fuzzy states means that a basic property such as length, with the default range and scale that makes it a typical quantitative character, can also be represented in a qualitative manner.

Thiele (1993) has argued that "the distinction between qualitative and quantitative data (...) may be more apparent than real". He suggests that shapes can be expressed in terms of numbers and ratios. They could also be represented by a proper transform (see the review on "feature extraction" by Rohlf, 1993). Even "presence" is in fact the property "number" with only two valid values, 0 for absence and 1 or more for presence. This may indicate that our list of basic properties may be shortened even further. It might also be possible to consider at least some of the qualitative properties as "summary characters" for the corresponding quantitative properties. For example, a particular color could be defined as a summary value for specific values of wavelength, grey level and chroma. The decomposition we advocate would make it easy to define such relationships, because of the small number of basic properties, and would make it necessary to define the relationships only once for any number of morpho-anatomical databases, because basic properties are the same in the various biological groups.

REPRESENTATION OF HOMOLOGIES

Homology is the similarity between character states that is due to inheritance from a common ancestor. It differs from convergence, the similarity between character states in unrelated organisms (no common ancestor) and parallelism, the similarity between character states that have evolved independently in related taxa by similar modifications along the same developmental pathways.

An example of parallelism in nematodes is the reduction of the posterior genital branch that occurred in every family (and most genera) of the Tylenchida. Convergence is seen in the three families Belonolaimidae, Dolichodoridae and Criconematidae, which belong to Tylenchida but which are not related at the sub-order level, where, for purely mechanical reasons, the development of a very long stylet is accompanied by the widening of the procorpus and the coiling of the procorpus lumen. Another type of convergence is seen in, *e.g.*, the supplements used in the example introducing Guideline 7 when supplement 1 of a particular species is not homologous to supplement 1 of another species.

Homology, parallelism and convergence are not morpho-anatomical data, they are relationships that exist between morpho-anatomical data. A morpho-anatomical database is primarily intended to store morpho-anatomical data, but relationships can be built on top of these data. For example, the same character "procorpus/lumen, posture, coiled" will be attached to specific taxa in databases for Belonolaimidae, Dolichodoridae and Criconematidae. After these databases are created, a relationship will have to be added for defining the existing convergences. The standardized representation will facilitate this operation since the converging character will always be the same, even when it is recorded in separate databases for each of the families. Supporting this type of view of characters is the subject of future work.

ORGANISM CHARACTER AND TAXON CHARACTER

Jardine (1969) noted that taxa and organisms do not have characters in the same sense. The taxon *Helicotylenchus* does not have four lines in the lateral fields, these lines are seen only in the individuals that belong to this taxon. While this may seem to be a case of splitting setae, it is true that there is an obvious difference in most characters between the data for one specimen (specimen X of *H. dihystera* has stylet length=26.5 µm), and the data for a taxon (the species *H. dihystera* has stylet lengths with mean 25 µm and standard deviation 0.7). In fact, the situation is even more complex: first, some characters have mean and standard deviation also for specimen records, *e.g.*, "annuli, width" and any dimension that refers to structures that exist in multiple copies in a single

organism; second, a species is the union -or the aggregation- of all its populations, each represented by its own mean and standard deviation (or its own set of qualitative states with frequency distributions). The value of a character for a species is the mean of its values in populations of the species, which are the means of the values in sampled specimens for each population, which sometimes are the means of the values in multiple copies of the structure. A proper morpho-anatomical database should allow entering data at three levels: individual specimen, population, and taxon levels, each with mean and standard deviation and with relationships built on top of this data so that population data could be computed out of the individual records of the specimens that form a representative sample of this population and similarly for taxon data out of the population data. The decomposition proposed here (biological structure, basic properties and states or values) would be used, of course, at all levels.

Uses of the characters

Identification and classification

The numerous applications in identification and systematics in general may use coding, weighing, ordering, selecting, testing, etc., of the characters. These various activities are handled differently by the various approaches in identification and systematics, and by the particular applications using each approach, and they will not be discussed here. We only wish to note that consistency and standardization of characters can only make easier any manipulation that needs to be made on the characters.

It should be possible to use the data stored in a general database with the structure advocated here as input to existing applications. For example, several identification applications use DELTA-coded data (Pankhurst, 1993). It should be possible to transform our data into DELTA data (Table 9), which could then be fed into, *e.g.*, Pankhurst's PANDORA system or Dallwitz's INTKEY (Dallwitz, 1993). This transformation would require that an expert selects the characters and states to be coded. Of course, the benefits of name extensions would not be available with DELTA (*e.g.*, n°6 and n°7 in Table 9). These are limitations linked to the DELTA coding itself, but those who wish to use existing identification software could do so. The actual transformation of our characters into DELTA codes could be done in various ways, *e.g.*, using views, and defining the best way to achieve this transformation will be the subject of future work.

Non-morphological data used in taxonomy

Systematics and identification often use other kinds of data. We have not attempted to design a universal system for the representation of any kind of character,

Table 9. *DELTA codes for selected characters in Tables 4B, 4D, and 6A.*

N°	
1	Stylet tip <shape, from pointed to broadly rounded>/
	1. pointed/
	2. narrow rounded/
	3. rounded/
	4. broadly rounded/
2	Excretory pore position relative to median bulb/
	1. anterior <= before=in front of>/
	2. at <= just at >/
	3. posterior <= after=behind>/
3	Excretory pore position relative to nerve ring/
	1. anterior <= before=in front of>/
	2. at <= just at >/
	3. posterior <= after=behind>/
4	Deirid <presence or absence>/
	1. presence/
	2. absence/
5	Body annuli width/
	µm wide/
6	Annuli orientation in anterior part of body/
	1. symmetrical/
	2. retrorse/
7	Annuli orientation in posterior part of body/
	1. symmetrical/
	2. retrorse/

and it may well be that such an enterprise would be doomed to fail from excessive ambition!

However, it is conceivable that the same approach as that used here could be used with other kinds of data to create databases with a structure which is at least related, if not identical, to the one we propose for morpho-anatomical databases. This would make it easier to design application programs using several kinds of data. As already noted, the decomposition of data into entity/property/value is a classical one in computer science. Moreover, "entity" can be defined as "biological structure" also with other kinds of data. For example, physiological functions are carried out by organs, tissues and cells, that is, biological structures. In a physiological database, it might be possible to use the same hierarchical list as the one defined for a morpho-anatomical database, with a different set of basic properties.

Naturally, to have any hope of integrating databases of different types such as morpho-anatomical, physiological, ecological, etc., let alone integrating databases of different species, our focus in this paper, it will be necessary to have a solid foundation of principles as advocated here as a basis for structuring the database.

Uses of morphological data in other fields

Other disciplines may need morpho-anatomical data. Ecology is the first discipline that comes to mind. For example, the biomass of nematodes can be computed from the values of the nematode dimensions stored in a morpho-anatomical database.

Conclusion

Constructing a large morpho-anatomical database requires solving numerous problems. Redefinition of the concept of character plus using name extensions and basic characters makes it possible seriously to consider constructing a database from published descriptions.

The representation proposed here seems at first very close to proposals made by other authors (*e.g.*, Lebbe, 1991) who also used the classical decomposition into entity/property/value. However, we believe this is the first time in a biological setting that such a strict standardization of entity as a hierarchical list of biological structures and property as a standardized abstract concept has been proposed.

The strictly enforced decomposition of characters, which clearly separates hierarchical biological structures, presents two major advantages. First, this hierarchy corresponds to what is called a plan of organization, which is well known for each main group of organisms and which has been well described by biologists. A specialist will find it easier to describe the hierarchy of systems, organs, tissues, cells, as long as this description is made first, before listing taxonomic characters. Major problems remain (homology, multiple points of view, etc.), but the concept of name extensions and the possibility of having multiple points of view for the same structure go a long way to solving some of them. Also, known homologies or homoplasies could be described as relationships between characters built on top of the character set described here.

Second, the organs, tissues and cells described for a morpho-anatomical database will provide a natural support for recording data in domains other than identification and systematics. For example, many genes are expressed only in specific cells and organs, *i.e.*, in particular structures. The protein produced by the expression of such a gene may have a physiological effect on other structures. In this way, a morpho-anatomical decomposition can be linked to genetic, biochemical, and physiological databases. As another example, a parasite is attracted by its hosts when its sense organs perceive certain compounds released by certain organs of the host, either directly or through the modification of the environment they cause. Here, a link can be defined between the morpho-anatomical structures and ecological (host-parasite relationships, environment) and biochemical data. Enforcing a strict separation between biological structures and properties provides a natural avenue towards interdisciplinarity.

The concept of basic properties also has two major consequences, parallel to those resulting from the concept of biological structures. First, it makes easier the decomposition of traditional characters by providing templates and default definitions. With the list from Table 3, it is obvious that the states "short" to "long" refer to the property "length" rather than an ad-hoc property "lengthening". As soon as a biological structure is defined, from the whole organism to individual cells and cell organites, the character set designer has access to the pre-defined list of properties, and this makes character decomposition a much easier task. Second, the very existence of basic properties is naturally conducive to standardization of characters. Within a group sharing the same plan of organization, it becomes possible to define a set of characters that would allow the description of any conceivable aspect (*i.e.*, basic property) of the structures listed in the plan of organization. Moreover, because basic properties are the same across biological groups, this standardization is not limited to a genus or a phylum, but applies to all living beings. Taxa with different plans of organization are composed of different structures, but any structure, be it from a nematode, a fish, a plant or a human being, has a shape, a length, etc. With the concept of basic properties, finding all the characters concerning shapes is as easy as finding where Mondrian put the red rectangle in *Composition of red and white No. 1*.

With basic properties, it also becomes possible to enforce rules and propose guidelines that will help both the specialist who is creating a character set for a particular biological group and the person who is responsible for extracting published character data and decomposing them according to such a character set. Rule enforcement raises the risk of diminished freedom, and many biologists will chafe at the idea that they may not be allowed to use whichever character they deem necessary for a particular application. Actually, the guidelines proposed here offer a range of possible characters which is positively staggering. While many applications in taxonomy and systematic use less than twenty characters and many published descriptions include only about 50-70 characters, the 272 structures in the latest version of the nematode character set could theoretically be described by twenty properties each, which represents 5440 potential characters (this number would be far greater if states were used to define traditional characters out of these potential characters). This cannot seriously be described as a marked limitation of freedom. In fact, even more characters are available as it would be rela-

tively easy to add new structures as needed, each with a set of basic properties attached, according to the guidelines proposed above.

A recent article in *Science* (Ashburner, 1995) contained some predictions for the future, at least one of which is of great import to taxonomy: "By the year 2000 or so . . . we will also have a complete database of all living organisms, including not only taxonomic data, but also morphological, ecological, biogeographical, and biological data." It seems that this prediction has been viewed with great skepticism in various scientific circles, quite correctly in our view, and one of the main difficulties is finding and describing the vast number of species in remote locations. We believe that the unexpected size of our character set points out yet another serious problem that will become apparent when people try to pull together large volumes of prior work and build formal databases.

If the hidden magnitude of the task holds in other areas too, and we have no reason to believe that it will not, this puts a large premium on being very careful in the early stages of this work and doing the work in ways that ensure that the information will be flexible and useful for many years to come. Not only does imprecise construction of character sets necessarily limit the availability of the information, but new demands placed on the character sets by the increasing needs for different kinds of electronic processing, among other things, may require considerable future effort that could have been avoided by proper construction in the first place.

References

ANON. (1996). Internet index of species launched. *Nature*, 380: 376.

ASHBURNER, M. (1995). Through the glass lightly. *Science*, 267:1609.

BISBY, F. (1994). Global master species databases and biodiversity. *Biol. int.*, 29: 33-40.

COLLESS, D.H. (1985). On "character" and related terms. *Syst. Biol.*, 34: 229-233;

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

DALLWITZ, M.J. (1993). DELTA & INTKEY. *In*: Fortuner, R. (Ed.). *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 287-296.

DIEDERICH, J. (in press). Basic properties for biological databases: character development and support. *J. math. Computer Modell.*

DIEDERICH, J. & MILTON, J. (1991). Creating domain specific metadata for scientific data and knowledge bases. *IEEE Trans. Know. Data Eng.*, 3: 421-434.

DIEDERICH, J. & MILTON, J. (1993). NEMISYS: a computer perspective. *In*: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 165-179.

FORTUNER, R. (1993*a*). *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press, viii + 560 p.

FORTUNER, R. (1993*b*). The NEMISYS solution to problems in nematode identification. *In*: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 137-163.

HORVITZ, E.J. (1993). Automated reasoning for biology and medicine. *In*: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 3-27.

JARDINE, N. (1969). A logical basis for biological classification. *Syst. Biol.*, 18: 37-52.

LEBBE, J. (1991). *Représentation des concepts en biologie et en médecine. Introduction à l'analyse des connaissances et à l'identification assistée par ordinateur*. Thèse de doctorat; Université Pierre et Marie Curie, Paris, xii + 282 + xxiv p.

PANKHURST, R.J. (1993). Principles and problems of identification. *In*: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 229-240.

RASKI, D.J. & LUC, M. (1987). A reappraisal of Tylenchina (Nemata). 10. The superfamily Criconematoidea Taylor, 1936. *Revue Nématol.*, 10: 409-444.

ROHLF, F.J. (1993). Feature extraction in systematic biology. *In*: Fortuner, R. (Ed.), *Advances in computer methods for systematic biology - Artificial intelligence, databases, computer vision*. Baltimore, USA & London, UK, The John Hopkins University Press: 375-392.

THIELE, K. (1993). The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics*, 9: 275-304.

ZAR, J.H. (1996). *Biostatistical analysis., 3rd ed.* Upper Saddle River, NJ, USA, Prentice Hall, x + 662 p.