On the representation of observational data used for classification and identification of natural objects

Jacques LE RENARD, Noël CONRUYT

Muséum National d'Histoire Naturelle, CNRS D-0699
Laboratoire de Biologie des Invertébrés Marins et Malacologie, CNRS D-0699
55, rue de Buffon, 75005 Paris France.

Abstract: Starting from an analysis of the qualities of descriptions and of the observational mechanisms, this paper illustrates the interest of using a "naturally" structured representation of observational data. A particular attention is given to the formalisation of descriptive concepts and their corresponding representation from a biological point of view. The main goal is to build classification (class definition) and identification systems that take into account diversity, inter-dependancy and variability of observed characters and to handle as well as possible incomplete information.

1. The central role of descriptions in natural sciences

The so called observational sciences have to deal with the ability to analyse the reality of things, in a word to describe. The description activity is so straightforward that we may wonder why we are analysing it. Yet every one knows that there exists some good descriptions and some bad ones, and that their use raises a lot of problems.

Why do we need descriptions? What are their qualities ? Here are questions for which we should bring clear answers, before searching which computer solutions are likely to facilitate and improve the activity of describing that occur for example when classifying or identifying living creatures and other natural objects.

1.1 Goals of a description

The description of different entities that compose our world appeared in the early antiquity as the fundamental way to increase knowledge. To "learn" what is an animal, a plant, a rock, etc., one needs to observe, but in addition to make a cognitive representation (for oneself) or a written representation (for others). The transmission of knowledge implies the notion of description.

A scientific description is an objective abstraction. It is an abstraction because it allows to free oneself from real observations that gave substance to it; and objective because it does not admit interpretation. Ideally, there is no distortion but a simple transcription "as identical" of concrete features of the observed subject into characters or characteristics that are represented. Traditionally the representation is made under textual form, using pictures to illustrate it; today, we can also use a lot of media that allow more power and flexibility.

One proceeds to describe, in the first place, in order to increase the number of particular (individual) descriptions, and next to learn about Nature at a more general level and better understand it. Qualities expected from descriptions are derived from this double objective.

1.2 Qualities of a description

We saw that the essential quality of a description is its objectivity, a perfect description is at the same time true and complete. Any method that is aimed at describing more easily must therefore allow to cover all observable features and to express them exactly, without ambiguity. In this way, the information content of the description is maximised. Ideally, a

perfect description should permit to rebuild exactly the primitive object; practically, a description is satisfactory when it gives a "good estimate" of this object, especially concerning its specificities. This implies to take into account not only the descriptive characters, but also the different links (topologic, relational, of dependancy etc.) that exist between these characters, because these links carry information too.

Aditional qualities can be mentioned such as clarity and concision, as for all scientific writings. Some authors give attention to the elegance of the text. It is rare to mention understandability as a quality that makes description more easily understood by someone who is not a specialist of the domain. This means that one should use a less technical vocabulary, with the possible counterpart of loosing accuracy and conciseness : so there is some compromise to be found, still waiting for a solution that allows to adapt the "level" of the description to the user. However, it is not sufficient for a description to be excellent in itself : it needs moreover to allow comparisons with other descriptions.

1.3 Qualities of descriptions

With the classification or identification goal in mind, our main concern is to compare description one another. When these descriptions have been written by the same author, they follow generally a common framework, and it is easier to compare them because homologous characters are located in corresponding parts of the texts. But when the authors are different, they could follow heterogeneous observation "methods"; comparisons between descriptions are then more difficult to achieve.

The notion of homology is very important; it allows to ensure that only comparable characters are compared. It is based on the fact that every biological object has an organization plan which is found identical in the other objects of the same kind. Recognizing and taking into account this general constitution plan (*bauplan* in German) allows a natural structuration of descriptions, following what we will call a *descriptive model*.

Remark : The above considerations virtually concern all natural subject descriptions. Although, both for classification and identification, each specialist restricts his studies to a more particular domain, like a zoological or botanical group, and/or a geographic area, and/or an ecosystem etc. In the following paper, it is a delimited domain instead of a "universal system" still not reachable at the present time that we have in mind.

2. Representing descriptive data

We start from the point that we can understand only what we can model, and that it is better to adjust the model to the reality than the reverse. We will study in more details what are the elements that constitute a description, and how they are arranged together by the descriptor (i.e. the person, generally the specialist, who makes the description, and not a described character which should be called ... a descriptum). We will deduce from this how descriptive models must be conceived, according to the above quality constraints.

2.1 Natural Structuration

As an example, let us take a particular domain : farm animals. In such a domain, anyone is a "specialist". Let us see how the specialist will classify and identify these animals.

First observation : all these animals have four limbs, two anterior, two posterior. The anterior limbs are either legs, or wings for poultries. So we learn that there are two principal categories, that the specialist will immediately name Mammals and Birds (with capital letters because we are in a scientific area). Next, among the Mammals, the cat and the dog of the farm are distinguished from others because they eat meat. Here are two other categories : Carnivorous for them, Herbivorous for the others. Among the Carnivorous, there is the Cat

that looks like the Tiger, and the Dog that looks like the Wolf : Felids and Canids etc. Thus, a true hierarchy (taxonomy) of categories (or classes, in a broad meaning) can be constructed, starting from the most general (Animals) to the most particular (Species like Cat, Dog, Horse, etc.). This "systematic" is founded on a hierarchy of discriminant characters (nature of anterior limbs, of food, etc.) that are more or less accessible; for example, on what criteria can the distinction between Felids and Canids be based ?

The specialist is able to recognize things at a glance : he is an expert of his domain. But in order to understand, to know, he needs to analyze the reality more precisely. It is only after having done (or read) various descriptions of Felids, that himself (or another specialist before him) will be able to state the definition of the class named Family Felidae and state that it can be distinguished from Canidae (among other things) by the fact that the posterior teeth bear cutting edges (they are named carnassials), while they are simple molars among Dogs.

We can notice that the "distinction between Canids and Felids" covers two dual approaches. In one hand, from a classificatory point of view, we learn by a generalization process that the character "presence of carnassials" synthetizes (or subsumes) all that has been observed on the different kinds of Felids concerning their posterior teeth. In the other hand, from an identification point of view, we deduce by analyzing the fact that Pussy bears carnassials, that it is a Felid and not a Canid. Whatever the approach, we had to deal with the description of posterior teeth; it is what we will call a *local description*.

2.1.1 Compositional logic

The description of an entity (let say a cat or a dog) is a composition of local descriptions that correspond to everithing that is observable and thus describable.

The description process is not ordered at random, but follows some logic which can be recognized. Whereas a cat or a dog both have a body, a head, four legs and a tail (they inherit all that because they belong to the Mammal class), it would be unconformist to begin the description by the tail; beginning it with the legs would be curious, unless the descriptor is an ant; but the choice between the body and the head is open. In fact, this logic is a matter of specialists who are the only ones to make an agreement on the definition of the most "natural" way for ordering the local descriptions. If it is the head that comes first, then the description will begin by its own characters, like its shape, size, color etc., by its connections with other parts, and then, following in turn a non arbitrary order, will continue with the description of its subparts (eyes, mouth, nose, ears etc.). And so on.

This decomposition into subparts is a basic mechanism; it is repeated as many times as necessary to reach the desirable level of detail (which, remember, depends on the intended use). Thus, we can agree upon an "exploratory tree", where at each node of we have a local description, and where each branch defines a relation between parts and subparts. This tree must provide all situations that may possibly be obesrved, including particular cases and exceptions, without introducing arbitrary limitations. From this point, the tree is generally thicker than necessary for each particular description situation, and some branches can be proved to be irrelevant (without significance).

In particular, when describing, we use an automatic pruning mechanism that makes sense. Thus, when we know that one part is absent, all descriptions on its subparts become irrelevant; likewise, if for example I am describing the farm watch-dog, and it doesn't want to open its mouth, I would greatly prefer not to have to describe its teeth or its tongue. This illustrates a common situation encountered when describing natural objects, where some of the local descriptions are not possible because of the observational situation (part hidden or presently unobservable), or because the specimen to be described is not complete. We can postulate that when a local description is absent, it means that its corresponding part is unknown; conversely, when we state that this part is absent, this brings an information that must be stated explicitly in the description.

This last distinction is important. When I am describing a Cat, if I say it has no tail, then this information leads to the fact that it is a member of the Manx race (cats without tail of Man Island), unless it is an accident and I know that it "had" a tail. On the other hand, if I do not mention the tail, I bring no information; the "value «unknown»" which is often invoked in this case is non-sense, or worse an artificial way to treat as an information what is not. The most convenient way to treat unknown facts in a description is to leave them blank.

2.1.2 "Point of view" logic

It often happens that a description of a natural object might be done at different levels. For example, it will focus on morphology, anatomy, cytology, or again on biochemistry or the genetic map. This is true anyway for each of the observational parts. The information attached to the different points of view are linked by the existing structural relations between these various observational levels.

Practically, the "point of view" logic is very similar to the composition logic. However, it doesn't have as rich a semantic; the fact that one level of analysis cannot be accessible for a given part of the description doesn't imply that this level remains inaccessible when describing its subparts. Another difference can be found when processing classification : a missing subpart will be taken into account whereas a missing point of view is devoid of classificatory signification.

One of the major interests of defining a descriptive model is to preserve the homology of characters even between different levels of observation. Thus, the descriptive model is a way to index knowledge and position it in order to compare it to others; it corresponds somehow to the relational and/or hierarchical structures in data bases.

2.1.3 Specialization logic

Let us come back to the farm animals, supposing that we can make use of a classification about different kinds of breeding farms. If we know nothing about "our" farm, the general model of bred animals contains four limbs, but if we know that it is specialised in aviculture, we can start from a more precise model, animals with two wings, two legs, a beak, feathers, or on the contrary without horns, teeth, etc.

The fact that a more precise concept of our farm is available, at an already abstract level, allows us to restrict the area of domain knowledge, and gives information in advance (without real observation) about some local descriptions. This mechanism, called specialization, is so general that it can appear in a lot of descriptions written by naturalists, in place of true local descriptions. Thus, simply stating that our farm breeds aquatic birds (ducks for instance) partly replaces a description about legs (that are always web-footed) or feathers (always watertight).

The specialisation is a convenient short cut : it allows to fill in "by default" (by inheritance) whole or part of a real local description by a conceptual one. Of course, there is a risk to be imprecise, or moreover to be incorrect. It is thus necessary to complement "manually" the deduced information.

2.1.4 Logic of exceptions

Whereas specialisation is the process of restricting the observable domain, exception is conversely a way to enlarge the current domain in order to handle particular cases. Suppose we learn that our farm does aquaculture; thus no more animals with four limbs, but fishes ("pisciculture") or even oysters ("ostreiculture") The descriptions will have to take into account characters about scales, fins, or shells. If those characters were not present in the general model of farm animals, it is needed for this particular case to be extended.

This process is complementary to the specialisation one even if it appears as a complication (like some "patches" that take place in computer programs). It seems better to follow this

process only in really exceptional situations, whenever it is justified to treat them apart rather than to integrate them in the general mold.

2.1.5 Iterative logic

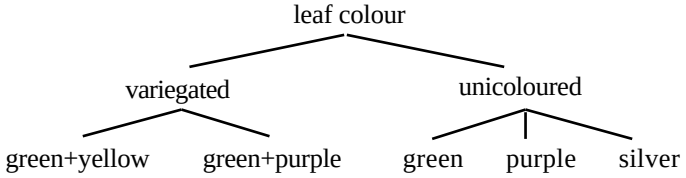
The study of above mechanisms was based implicitly on a matching between a description composed of sub-descriptions (or local descriptions) and a descriptive model composed of sub-descriptive models. The description is concerned with the observed facts whereas the model is concerned with observable facts.

It often occurs that, in a description, several characters, although they are not rigorously identical, are of the same "kind" and follow the same descriptive sub-model. Consider the example of mammal teeth. If we were to describe the human set of teeth (if we are afraid of the dog's one), we well see that there are several kinds of teeth, let say 3 or 4 kinds depending on our perspicacity. Those of us who are well informed will name them directly : incisors, canins, premolars and molars; but it is not necessary at to know all their names to describe them correctly. It is sufficient to follow a common sub-model of teeth description, and apply it iteratively as many times as necessary (here 3 or 4 times according to the ability of the descriptor to see the difference of nature between premolars and molars).

We pointed out that we had to respect the fundamental homology principle. If we had to compare in detail the set of teeth of the cat and of the dog, we must be sure to compare canins (or "fangs") of one with canins of other; otherwise we get lost. One must be aware of the interpretation risk (of being subjective) arising when venturing in "local identifications"; the descriptor who is not aware of the limits of his knowledge in the domain would make a mistake if he names canins the tusks of an elephant and the tusks of a morse; the consequence of this mistake is that objects to be compared are not homologous : the elephant tusks are modified incisors, whereas the morse ones do are canins, though of an exceptional size. It is right that it is difficult to only describe without searching to understand and to learn; but paradoxically, a good description should not call for intelligence because we are biased by our mental model and anybody may make a mistake.

Another situation may occur when describing. Suppose that we proceed to a local description of a plant inflorescence, and that the corresponding descriptive sub-model gives as a list of possible colours white, yellow and red, and that several answers are allowed (multiple choice). If we answer together white and yellow, that means that the colour is white or yellow, it is a lack of precision (why not an intermediary shade as white-yellowish ?). To express that we observe effectively the colours white and yellow simultaneously, it is necessary to make two successive local descriptions, one for describing flowers with white colour only, the other for flowers with yellow colour; in fact, there is a high probability to find other characters to differenciate the two types of flowers, as for instance their localisation in the inflorescence or also their sex, and that these flowers have not the same organical signification.

Remark : we need to distinguish this last case from the description of associations (of colours for instance) which are referenced under names like streaks, mosaic, etc. The fact that a leaf is variegated with green and yellow must not be expressed by the choice of green and yellow simultaneously, but by the single choice of the association green+yellow duly indexed. This can be represented in the descriptive model by a hierarchy of classified values like this :



Each time we have to express co-existing facts (noted simultaneously), the iteration process is

the one to be used.

2.1.6 Contextual conditions

The characters are generally dependent from one another. Rather than distorting the reality with an independence hypothesis (too rarely verified), it would be better to get the best of the information brought by these relations.

Co-existence and exclusive relations appear frequently in descriptions. They give respectively a condition of presence or absence of a character depending on the "context" made by other characters. For example, in the Mammals' classification, there is the fact that some have a placenta and others don't (distinction between Placentalia and Aplacentalia); it is obvious that this should not be observed on male individuals; if a bull is described, it is "not relevant" to know if it is gravid, or to know the number of dugs carried on its udders. One can notice that, as for the "value «unknown»", it is non-sense to speak about the "value «not relevant»" unless one needs to fill empty boxes in data matrices : if the sex of the bull is male, this carries all the information related to the "non relevance" of the gravid character, and shows the general fact of exclusion between masculinity and pregnancy. Nature is so made.

One can easily imagine co-existence relations, when the presence of a character is deduced "automatically" from the context. Such relations are sometimes perceptible only by specialists, and that constitute their expertise. We will take a real example from the diagnostic of plant diseases : the expert notes a withering of leave extremities and will focus on the most inexpected part of the plant (the collar at the base of the stem) to see if there is not a "canker" or a tumor that stops sap circulation. He thus uses this way a co-existence relation, and more precisely here a cause to effect relation.

Because of the variety of nature, dependencies between characters are not absolutely marked. For instance, some witherings are not due to a collar canker, and Nature does not like "rules" or "laws" without exceptions. Thus it is important, not only to consider dependency relations, but also to specify their applicability conditions, that is to say exceptions and related contexts.

In a lot of situations, a part that should be theoretically observable is not; or on the contrary, a local description is only possible under some conditions. This can be turned into contextual rules, for instance : if the dog is nasty, then don't observe its teeth; or : if the bird is flying, then describe the marks that are under its wings. These conditions are common sense knowledge and can be well used to guide cleverly the observations.

2.2 Structured representation with a descriptive model

For a given domain, the descriptive model is created by the expert. He must represent all what is observable as a structured scheme.

Furthermore, the major goal of the descriptive model is to be transposed in an observation guide to help the user to describe. It must be a way to translate without constraints the set of mechanisms or observational logics shown up precedently. So it is a representation of the set of all the observable knowledge, well suited for acquiring the observed knowledge.

The descriptive model can take several aspects equivalently, depending on the target user. In depth, it is represented under a data processing aspect adapted to observable knowledge bases; one can find objects like "frames", lists, matrices, rules, pictures etc, written with a syntax that translates as exactly as possible the different observation mechanisms and the "background knowledge" of the domain. This form is not to be read by the naturalist; it is only a technical representation, used as input and/or output to the different modules of description treatments.

The processing model must of course follow a formalism that can be transcribed immediately to a mathematical plan, in order to be able to process knowledge with symbolic data analysis programs, inductive ones or others. Our individuals (or subjects) are represented as boolean

symbolic objects (Diday, 1991) of type "synthesis objects" because of the use of the iteration logic that introduces "hords" in our descriptions. Furthermore, we introduce the notion of "composite objects" (Conruyt *et al.*, 1992) to deal with the compositional logic as explained above.

The descriptive model must also be presented in a more practical and synthetical way for the naturalist, specialist of his domain who elaborates and updates it. Its structuration is logically depicted by a tree or a graph showing parts and subparts with their own relations and characteristics. The "object" manipulation (in a computer science meaning) to create, modify, move, associate pictures to them etc, is better made in a graphical way with interactive tools that are easily used by biologists, who are not computer programmers.

One last issue, perhaps the most important practically, allows to present the descriptive model like a real observation guide : we called it "questionnaire" (Manago *et al.*, 1992) in our developped applications because it is put in the hands of the user, under a flexible but logical navigation form between different input screens. Each screen (called a "card" because of the HyperCard tool which is used) corresponds to the acquisition of a local description, matching exactly the equivalent part in the descriptive model. Notice that the descriptive model can provide some gradation for answers' accuracy, for instance giving intervals of numerical values, and at last permit to use the answer "?" to express the absolute uncertainty. This is essential for real descriptions, where context or particular events do not allow complete descriptions.

The final descriptions, the consistency of which is ensured by complying to the descriptive model and the completion verified at the end of the data entry, can be presented in different ways too. The initial form is the one of the filled questionnaire. It can be imported again to bring corrections or further descriptive informations. But it is sometimes useful to be able to visualize a description as an instanciated subgraph of the descriptive model. This form allows to highlight the underlying structure of the description that is somehow lost sight during the questionnaire navigation. In fact, these two forms complement each other and the user must be given the possibility to switch easily from one to the other. Moreover, it is nearly necessary to be able to present the user with descriptions under a natural langage text form, as it always existed. It is not difficult here to offer a choice of several target langages. At last, for best efficiency and homogeneity, input descriptions are recorded with the same syntax representation as the descriptive model. Therefore, observable and observed facts benefit from the same well adapted formalism, that allows to use them jointly and give more consistency and power to the programs that treat them.

We will not detail here the different technical solutions which allowed us to represent the different observational mechanisms. "Frames" are applied as a structure basis. They are "objects" with their own slots (characters or attributes). Each slot can take one or several possible values (in a list, possibly in a hierarchy for nominal classified values; in an interval for numerical values); once valued, each slot expresses a descripted character or a feature. When objects correspond to subparts (but not to points of view), their stated absence is recorded as significant. The specialisation and particularisation mechanisms are expressed by "class" instanciation (in a computer science meaning) with inheritance. The iteration mechanism is delt using "variables" and a first order logic. At last, the context conditions are represented as rules or demons.

It is now possible, by using AI knowledge representation methods, to formalise such complex descriptions that are required from the "truth" of nature, without transposition bias, without resort of subjectivity, and with as little loss of information as desired.

There is a good way to make sure that obtained descriptions satisfy our quality criteria. One has only to compare such descriptions produced under their natural langage form, with those directly written by specialists. It is then very easy to estimate drawbacks of ones and others; this is independant of the fact that "conform" descriptions (to the descriptive model) have the great advantage to be comparable to each others and easily mobilizable.

2.3 Processing descriptions for classification and identification

In our applications, all descriptions that are processed are pre-classified : we don't use descriptions for aggregation classification (or categorization). We call a *case* or an *example* the association of a description with an identification from the specialist. For classification purpose, a decision tree is grown using inductive learning from examples. This general knowledge extracted from the examples allows the classes to be characterized (or intentionally defined). For identification purpose, a case based reasoning strategy is used to directly compare the examples in extension (Conruyt *et al.*, 1992).

3. A brief illustration

A real world application of the ideas presented above has been developped by Pr. C. Lévi, a renowned specialist of Marine Sponges. Using the interactive tools we have designed, he defined a descriptive model concerning the genus *Hyalonema* (including morphological and histological characters, as well as contextual informations about the specimens). Using a *questionnaire* automatically built from the descriptive model, he acquired most of the descriptions already published in the literature in a "standard" manner, allowing him to test the different systematical concepts he was using for many years, and to iteratively derive an improved method of description. He presently applies this method in order to propose a new set of the descriptions that he had already published himself, as a way to transmit to other or future specialists of Sponges an optimized package of information about his domain of expertise (Conruyt *et al.*, 1993).

Acknowledgements

Thanks to Prof. E. Diday, Drs J. Lebbe and R. Vignes at INRIA for their valuable advices. And to M. Manago for extensive discussion and emendation of the text.

References

Conruyt N., Manago M., & Le Renard J. (1992) "Modélisation, Formalisation et Analyse d'Objets biologiques en vue de leur identification: application au domaine des éponges marines.", Actes des 3èmes journées "Symboliques- numériques", Université Paris-IX-Dauphine, 1992.

Conruyt N., Manago M., Le Renard J. & Levi C. (1993) "Une méthode d'acquisition de connaissances pour la classification et l' identification d'objets biologiques.", actes des treizièmes journées sur les systèmes experts et leurs applications, EC2, Avignon, 1993.

Diday E. (1991) "Des objets de l'analyse des données à ceux de l'analyse des connaissances", in "Induction symbolique-numérique à partir de données", vol 1, p. 9-75, Kodratoff Y., Diday E., editions Cepaduès, Toulouse, 1991.

Manago M., Conruyt N. & Le Renard J. (1992) "Acquiring Descriptive Knowledge for Classification and Identification". in: Th. Wetter, K.-D. Althoff, J. Boose, B. Gaines, M. Linster & F. Schmalhofer (eds.), Current Developments in Knowledge Acquisition - EKAW ´92, Springer Verlag.