A KNOWLEDGE BASE FOR CORALS OF THE MASCARENE ARCHIPELAGO: GENUS *POCILLOPORA*.

N. Conruyt[1], G. Faure[2], G. Ancel[3], J. Le Renard[4], M. Guillaume[4], O. Naïm[5], N. Gravier-Bonnet[6], D. Grosser[1]

[1]Institut de Recherche en Mathématiques et Informatique Appliquées, [3]Centre Multimédia, [5]Laboratoire d'Ecologie Marine, [6]Laboratoire d'Ecophysiologie, Université de la Réunion, BP 7151, 97715 Saint-Denis, La Réunion, France.
[2]Laboratoire d'Hydrobiologie Marine et Continentale, Université de Montpellier II, Pl. E. Bataillon, 34095 Montpellier, France.
[4]Laboratoire de Biologie des Invertébrés Marins et Malacologie, Muséum National d'Histoire Naturelle, 55 rue de Buffon, 75005 Paris, France.

corresponding author: Noël Conruyt
IREMIA
Université de la Réunion
15, av. René Cassin - 97715 Saint-Denis
Messag. Cedex 9, France

tel: (+0) 02-62-93-82-73
Fax: (+0) 02-62-93-82-60
email address: conruyt@iremia.univ-reunion.fr

Short running title: A knowledge base for corals of the Mascarene archipelago: genus *Pocillopora*

keywords: knowledge base, *Pocillopora*, Mascarene archipelago, expert system, descriptive model

A KNOWLEDGE BASE FOR CORALS OF THE MASCARENE ARCHIPELAGO: GENUS *POCILLOPORA*.

N. Conruyt[1], G. Faure[2], G. Ancel[3], J. Le Renard[4], M. Guillaume[4], O. Naïm[5], N. Gravier-Bonnet[6], D. Grosser[1]

[1]Institut de Recherche en Mathématiques et Informatique Appliquées, [3]Centre Multimédia, [5]Laboratoire d'Ecologie Marine, [6]Laboratoire d'Ecophysiologie, Université de la Réunion, BP 7151, 97715 Saint-Denis, La Réunion, France.
[2]Laboratoire d'Hydrobiologie Marine et Continentale, Université de Montpellier II, Pl. E. Bataillon, 34095 Montpellier, France.
[4]Laboratoire de Biologie des Invertébrés Marins et Malacologie, Muséum National d'Histoire Naturelle, 55 rue de Buffon, 75005 Paris, France.

ABSTRACT

A knowledge base is described which consists of computer aided systematics tools for describing, classifying and identifying coral specimens. An example is given involving 5 selected species of the genus *Pocillopora* from the Mascarene Archipelago (among more than 40 described species). The knowledge acquisition tools help the expert to build a descriptive model (what is observable), and to collect structured descriptions (observed cases) with a questionnaire. Then, the expert can apply the scientific method: experimenting (learning rules from cases with decision trees) and testing (identifying new observations with case-based reasoning). The method gives the expert the ability to update the knowledge base according to the results of determinations and comparisons, and thus improve iteratively his descriptive model and case base. The resulting expert system enables a user to identify the corals with a high probability of success.

INTRODUCTION

Scientific databases are becoming increasingly common in Biology. Some institutions distribute biological databases on CD-ROM (e.g. ETI in Netherlands, ICLARM in the Philippines, etc.) or make them available through the Internet (e.g. HBS in Hawaii, AIMS and CSIRO in Australia, etc.). Specialists of a domain, other biologists (students, amateurs) and professionals (environment, tourism) can easily consult these information systems of taxonomic and biogeographic significance.

In order to proceed, a lot of these databases require the knowledge of the specific name of the specimen under consultation. On the contrary, for a given taxon, knowledge bases (or knowledge-based systems) provide those ignorant of that name with some identification help, and they also give some classification help (class definition) to the specialists whose domain has been only partially studied so far (this is true for a great part of the marine fauna).

The term knowledge base comes from computer science and artificial intelligence. We prefer to use it here instead of the term database because it is more accurate for our purpose: knowledge is more general than data, i.e. we rely not only on observed facts (data, descriptions, cases, examples), but also on observable things (descriptive model, modelisation of data, metadata) and produced facts (classification or decision trees, rules, identification).

A lot of databases reproduce electronically textual descriptions and identification keys that already exist in books. This is of significant interest when it comes to a well-known and stable field. But it is not sufficient as regards evolving or not well-known groups. In such domains (sponges, hydroids, corals, etc.), specialists are not particularly keen on disseminating their taxonomic information. They need research-oriented systems dedicated to systematics.

In other domains such as botany, some researchers have come up with solutions for coding descriptions. Their programs provide comparability of descriptions and bring identification facilities to databases (e.g. Pankhurst 1970; Dallwitz 1974). The method of Dallwitz (1980) using DELTA (DEscription Language for TAxonomy) was applied to the coral genus *Acropora* by Wallace and Dallwitz (1981).

We use this kind of coding approach in our study, but we have been more influenced by the machine learning school in computer science (Quinlan 1986). The originality of our work is to implement the scientific method in Biology: experimenting (learning rules from examples) and testing (identifying new observations, improving the initial model and descriptions).

Unlike the work of Brown and Navin (1992) on scleractinian identification, we don't use traditional elicitation of rules by interviewing the domain expert or taking them from a monograph (like the one of Veron et al 1977). We start from pre-classified descriptions (i.e with the associated identification) that we call cases or examples. Then, we apply machine learning techniques on them for extracting these rules by induction. The input cases are compiled by the expert himself (here G. Faure). Their information content comes from different sources: literature (descriptions of the types of species, other monographs) and samples from a collection (descriptions of specimens). Thus, with many descriptions of the same class (or taxon) in extension, it is easier to take into account intra-specific variations.

Our knowledge base aims at providing computer aided systematics tools to experts for describing, classifying and identifying coral specimens of the south-west Indian Ocean (Mascarene Archipelago). As an outcome, the identification module may be used by non specialists.

A first study has been conducted on the genus *Pocillopora*, which includes 5 species (*damicornis*, *woodjonesi*, *eydouxi*, *verrucosa*, *meandrina*) among more than 40 species described at world level, and the 5 known ecomorphs of *P. damicornis* in the area (*bulbosa, acuta*, *brevicornis*, *caespitosa*, *favosa*).

The reference collection is the one of G. Faure (1982). It is located at the University of La Réunion (Lab. of Marine Ecology), at the Museum of Natural History of Saint-Denis de La Réunion and at the National Museum of Natural History in Paris (Lab. of Biology of Marine Invertebrates and Malacology).

METHODOLOGY

In order to apply the scientific method in biology (conjecture and test), our approach follows a natural process of knowledge acquisition which is divided in four steps:

- Acquisition of a descriptive model,
- Acquisition of examples (or cases),
- Processing of this knowledge,
- Validation.

Acquisition of a descriptive model

The descriptive model represents all what is observable as a structured scheme, so that the user can easily acquire the observed knowledge. Representation of individuals (specimens and/or species) is different than a flat one in a data matrix: in the former, there is a dependent composition of local descriptions, although in the latter the defined characters are independent one from each others.

In a descriptive model, there are logical rules for describing a species (e.g. *P. damicornis*): composition, point of view, specialisation, iteration, contextual conditions, etc. (Le Renard and Conruyt 1994). The structuration is a means to take into account good sense background knowledge that can be useful for description capture, management and processing (Allkin 1984).

All the observable components (called objects or parts) of the species of *Pocillopora* and ecomorphs of *P. damicornis* have been defined (Fig. 1). Their own characters (attributes) have been listed. There are 37 objects and 87 attributes for 9 descriptions (species and ecomorphs).

For example, the corallites on branch ends have three attributes: shape, distribution and disposition. For each attribute, the expert enumerates the observable values, in order that all possible descriptions of the genus can be covered.

In some specimens, some objects can be absent (e.g. on fig. 1, the object corallites on verrucae has a minus sign before). Other objects like wall, septa, etc. are dependent on the presence of the first: this background knowledge must be explicitly represented in the descriptive model to ensure the coherency of the description phase.

Some objects are physical (a box surrounds them) and others are fictitious (no box): these latters indicate viewpoints of the components at different levels of observation (e.g. macro structure, micro structure).

In fact, one of the roles of the descriptive model is to bring an observation guide for the end-user: the objects are linked together with relations that go from the most general to the most specific (from left to right), making the next description process easier for the non specialist.

In Conruyt (1994), we show that the acquisition of a descriptive model is the most important part of the method: the robustness of classification and identification results from the quality of processed descriptions, and thus from a well designed descriptive model.

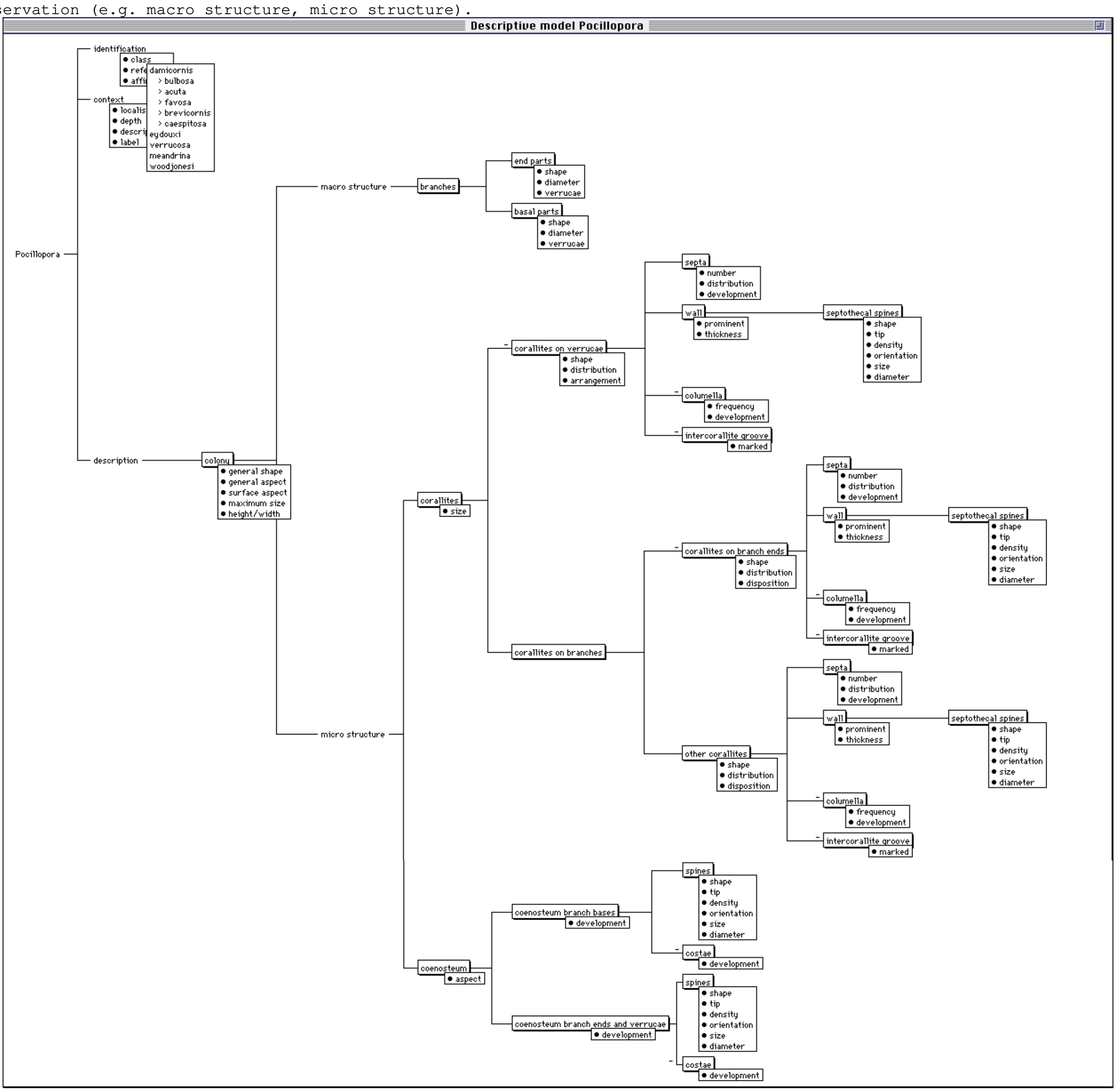

Fig. 1: the descriptive model of *Pocillopora* defines the structure of all observable descriptions of this genus in the Mascarene archipelago: 5 species and 5 ecomorphs of *damicornis*. The objects are the nodes of this description tree and the attributes are refered to each component. The figure shows here only the possible values of the attribute class of the object identification.

Acquisition of examples (or cases)

Starting from the definition of the descriptive model, a program builds a questionnaire automatically. It allows the expert and other biologists to acquire individual descriptions and make a case base. An identification is associated to each description in order to form a case.

The questionnaire follows the structure of the descriptive model: there are as many cards as there are objects and attributes. The observed descriptions are sub-trees of the descriptive model (see Fig. 1 and Fig. 2). Therefore, they can be directly compared with one another in making them pass cards by cards: this is easier than viewing lists of attribute-value pairs.

Each card is a local view of an object (Fig. 3). The user can navigate between cards in following the path of descendants and parents step by step. He can also jump from the description of an object to another through the use of the global view (Fig. 2).

Attribute cards are leaves of the description tree. As for objects, comments and illustrations for each of the values may be provided by the expert: it helps the user to interpret correctly the asked question (Fig. 4).

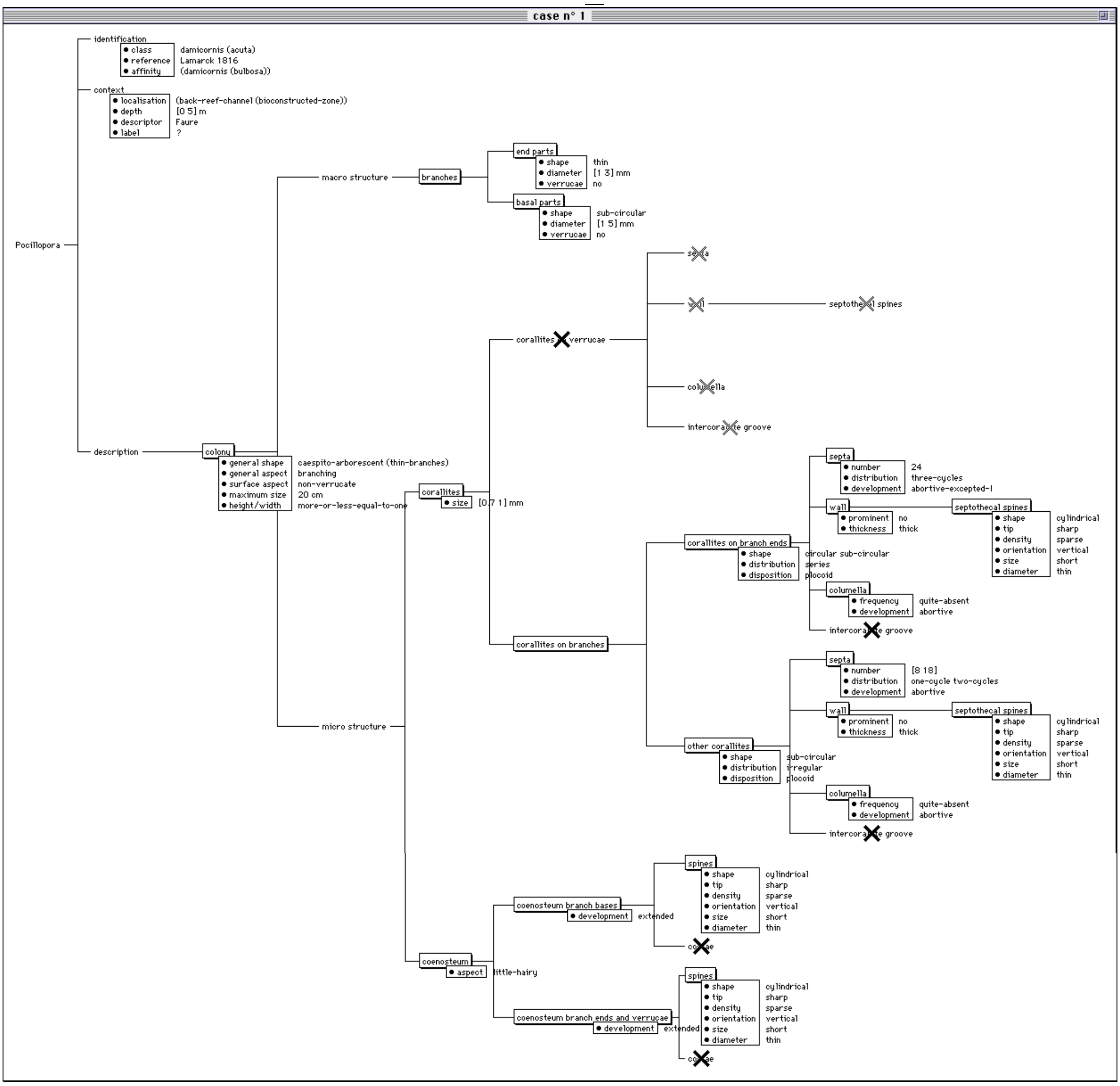

Fig. 2: A structured presentation of a description of *Pocillopora damicornis acuta*. The expert made a synthesis of descriptions for this ecomorph, based on specimens and literature (the attribute label of the object context is then unknown and the depth is an interval). Absent objects are crossed out (e.g. corallites on verrucae): other absent objects are deduced automatically (septa, wall, etc.). Most attributes are nominals and few are numericals. Some nominal attributes are classified (e.g. general shape of the colony) in order to specify some states, others are multi-valued (e.g. shape of corallites on branch ends): the object describes in fact a set of components that can share simultaneously different values. This remark is also valid for numerical attributes: the size of corallites can vary between 0.7 and 1 mm.

Processing of this knowledge

Depending on the goal to be achieved, two different types of method are used in order to process the case base: induction for classification, case based reasoning (CBR) for identification.

For classification purpose, a decision tree is built. Starting from descriptions (representation in extension) of classes to learn, an inductive method based on entropy and information gain measure (Shannon 1949; Quinlan 1986) gives a characterisation of these classes (representation in intension) with a set of rules. Each path from the root to a leave of the decision tree is a classification rule (also called a diagnose in biology). For *Pocillopora*, we obtain the following decision tree which classifies the 9 descriptions of species and ecomorphs (Fig. 5). This classification tree can be used in consultation mode to determine a new observation. It is one of the most discriminant decision trees: in only two questions, we are able to reach a conclusion. Nevertheless, when the user does not know how to answer to a question, the consultation of the decision tree is inadequate (Manago et al 1993).

For identification purpose, a CBR strategy is used (Bareiss 1989). Given a set of examples, it dynamically extracts the most efficient criteria from the ordered list of tests after each answer of the user (Fig. 6). The cases are selected from this reply. If the answer is unknown, the second most discriminant test is proposed to the user, and so on.

Even if this method of identification is more resilient to this noise (unknown responses), it is not as good when facing errors of description. This is due to the monothetic approach of this strategy (Pankhurst 1991): it is based on one and only one criterion at a given moment.

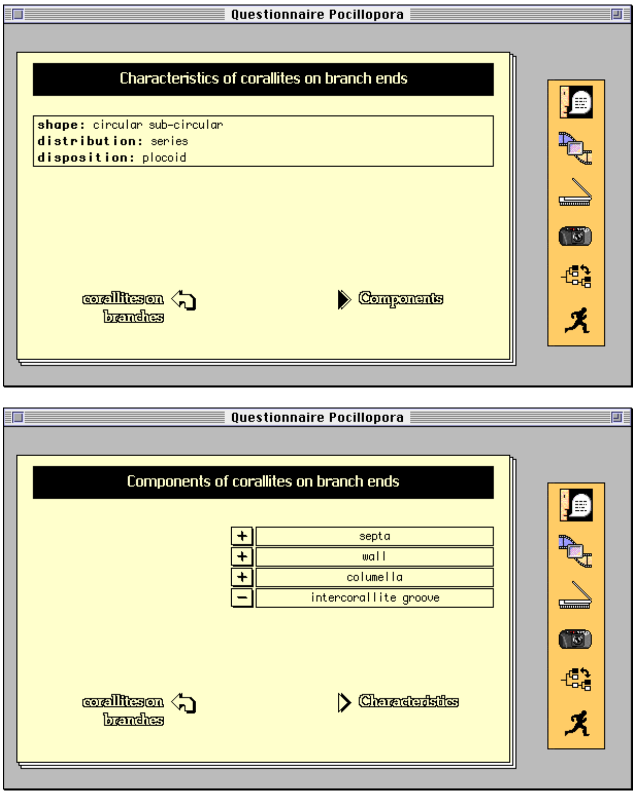

Fig. 3: A local view of the object corallites on branch ends from *P. damicornis acuta*. The characteristics of this object are above and the components are below. The user can switch from one local description to the other and navigate from this object to other related objects.

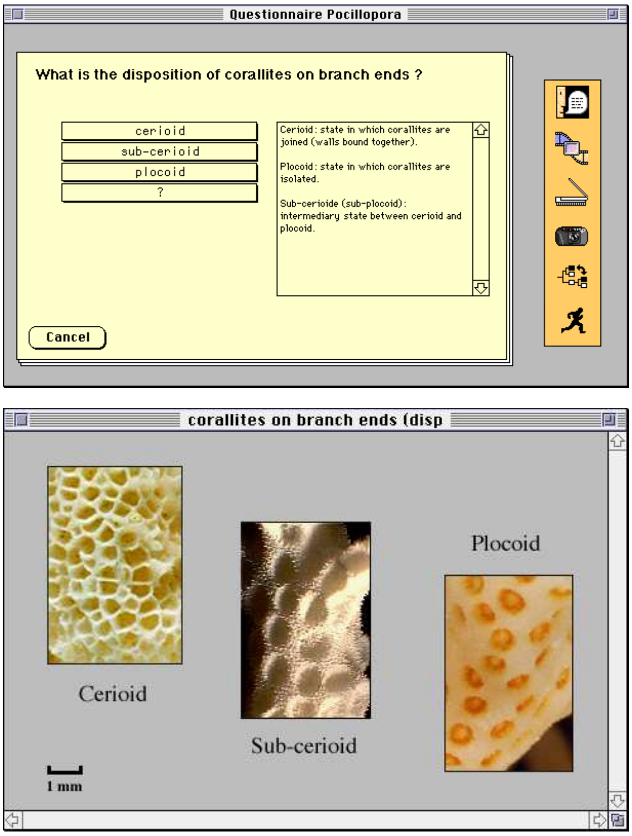

Fig. 4: An example of a commented and illustrated attribute: disposition of corallites on branch ends. The first icon on the top-right opens a comment for the values and the icon of the camera opens a window of illustrations (below).

Other methods of CBR are polythetic (i.e. rely on a combination of criteria) and are more robust to errors: the most similar cases method is derived from the k nearest neighbours one in data analysis: the examples are retrieved in calculating a similarity measure between descriptions. This is a comparison process which implies the whole set of characters (or attributes). A score between 0 and 1 gives the percentage of resemblance between two cases (Fig. 7).

For the consultation, there is an interest to combine these methods (induction and CBR) at different levels of integration in order to get more robust results (Auriol et al 1994). These knowledge processing tools are modules of the Kate™ software which is marketed by the company AcknoSoft in Paris.

Validation

With the help of these tools, the expert can evaluate the results of classification and identification, according to the quality of its own descriptions and descriptive model. Inductive learning as well as the repetitive use of the questionnaire remain useful for detecting possible inconsistencies within the case library, and thus for refining the knowledge base (observable and observed facts).

This experimentation and test method is also a means for the computer scientist to improve algorithms in order to fit with user needs. For example, in the decision tree of Fig. 5, two very closed species (*P. verrucosa* and *P. meandrina*) cannot be discriminated. This can be interpreted by the expert as an interesting result from a classification viewpoint (authors don't agree really if they are the same or not).

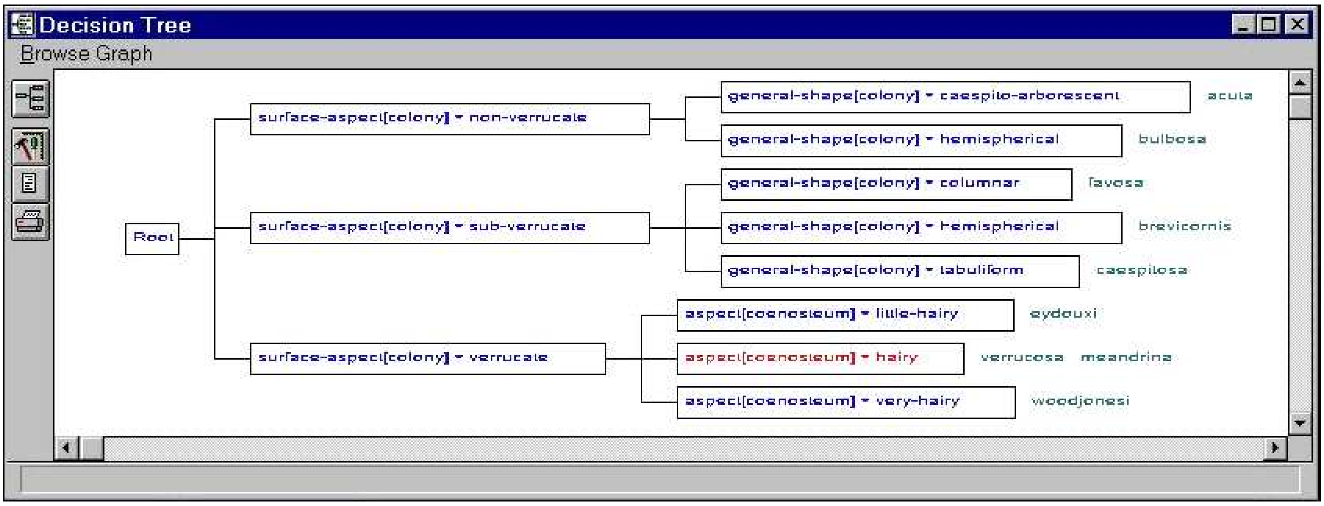

Fig. 5: A decision tree for classifying the genus *Pocillopora*.

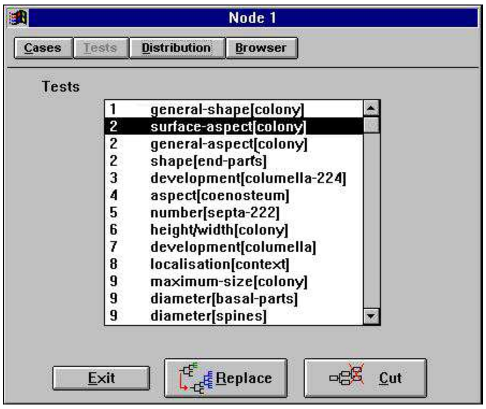

Fig. 6: The list of ordered tests at a node (here the root node) is computed with an inter-class discrimination measure: the information gain. The higher is the gain, the more homogenous is the repartition of cases between classes. At each node, the best test (i.e. the most discriminatory attribute) is chosen to separate the cases and generate a tree. Nevertheless, the user can modify the default decision tree by replacing a test: the example of Fig. 5 was made with the choice of the second most discriminant attribute at root node.

But in fact, there exists some little differences on multi-valued attributes between the two species (e.g. distribution, number of septa). The presence of multiple states was interpreted by the computer scientist as a disjunction of values due to imprecision and not as a conjunction of values due to intra-specific variation. So, an object representing a set can share different states simultaneously (e.g. Fig. 2: shape of corallites on branch ends: circular and subcircular). This background knowledge must be differently processed: in case of doubt, we have to refrain from discriminating whereas in case of variation, the discrimination must carry on (i.e. finding other criteria after the aspect hairy of coenosteum in Fig. 5).

Another requirement is to easily update the case base after making modifications in the descriptive model. The descriptive model supports all the validation process.

DISCUSSION

For the moment, only 9 descriptions of species and ecomorphs have been recorded in the knowledge base, corresponding to the 9 classes of *Pocillopora*. Those cases were described by the expert. The descriptive model was refined twice and now looks like the one of Fig. 1. Another work was to illustrate each attribute from the questionnaire, like the one of Fig. 4.

Now, we are able to take some specimens in collection and tell other "naive" persons (like the co-authors) to describe them through the questionnaire. We want to integrate two other dimensions in the case base: the observed intraspecific variability and the different interpretations of vocabulary and illustrations on the same sample from many users.

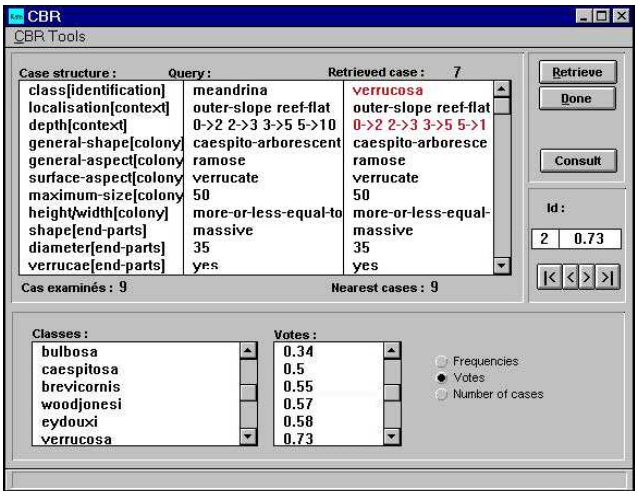

Fig. 7: The CBR strategy of the nearest cases makes identification of classes (i.e. species and ecomorphs) by comparing the whole set of characters of descriptions. A similarity measure counts the associations of states for each character and delivers a resemblance score between descriptions. We can see here that *P. verrucosa* is near from the query *P. meandrina*: other classes have a lower score.

Effectively, because of its very large geographical and ecological distribution, this genus shows a great intraspecific, even intracolonial variability (Veron and Pichon 1976).

This is why in the descriptive model we have integrated the different sort of corallites within a species and/or a colony: corallites on verrucae, corallites on branch ends, other corallites (Fig. 1). Each of these objects shares the same sub-tree or descriptive structure but with different possible values.

Because of this variability, the species present a very complex taxonomy which is often contradictory in the literature. Some very significant characters of species for an author are not recognized, or even in contradiction for other authors.

For example, the development of columella is today a character that discriminates the group *P. verrucosa* - *P. meandrina* (absent or abortive) from the group *P. eydouxi* - *P. woodjonesi* (well developped). But in the original description of *P. eydouxi* (Edwards and Haime 1860), the columella is absent. Furthermore, the presence of columella is not a specific character of two species placed in synonymy with *P. eydouxi* (*P. grandis* and *P. elongata*) as it is described by Dana (1848): "columella obsolescent" (*P. grandis*), "a minute columella sometimes seen" (*P. elongata*). At last, the two first authors (Edwards and Haime 1860) point out the presence of a "saillie columellaire bien développée" in *P. verrucosa*.

Given such relative observations, the expert must define a description structure of all observable objects (past and present) in order to be exhaustive. This is the most difficult task, but the descriptive model is the key entry for communicating descriptions (preferably of specimens) between specialists and other biologists. The descriptions will be comparable on the same rigorous basis.

Thus, starting from the same descriptive model and the descriptions of multiple specimens from a species by different observers (specialists and apprentices), this will allow us to give more robustness to learnt classification rules and to resulting identifications.

CONCLUSION

After applying this methodology, every one concerned with corals of the genus *Pocillopora* can have the use of a robust expert system, and can thus identify them with a high probability of success.

In future, we will apply the proposed method to the identification of all coral genera of the Mascarene archipelago (about 55 genera). We also plan to make a cooperation with AIMS (Australian Institute of Marine Science) to build an identification system of indo-pacific corals (about 70 genera) that will be added to their CoralBase project.

ACKNOWLEDGMENTS

We want to thank P. Gigord, Director of IREMIA, for finding funds that allowed us to build our prototype on *Pocillopora*. We also thank D. Choussy for taking the photos that illustrate the application, and W. Birnie for english improvement. We are also grateful to J. McManus who critically revised the linguistic content of the manuscript.

REFERENCES

Allkin R (1984) Handling taxonomic descriptions by computer. In; Allkin R and Bisby FA (eds.), Databases in systematics. Systematics Association London, Academic Press, (26) pp 263-278

Auriol E, Manago M, Althoff KD, Wess S, Dittrich S (1994) Integrating induction and case-based reasoning: methodological approach and first evaluations. EWCBR-94 - Second European workshop on case-based reasoning. M Keane, JP Haton & M Manago (Eds.), AcknoSoft Press, pp 145-155

Bareiss R (1989) Exemplar-based knowledge acquisition: a unified approach to concept representation, classification and learning, London, Academic Press inc

Brown D, Navin K (1992) An interactive knowledge base for identification of scleractinian corals. Proc 7th Int Coral Reef Symp Guam 2 : 661-664

Conruyt N (1994) Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques. Thèse de doctorat, Univ Paris-IX Dauphine, pp 1-281

Dallwitz MJ (1974) A flexible computer program for generating identification keys. Syst Zool 23 : 50-7

Dallwitz MJ (1980) A general system for coding taxonomic descriptions. Taxon 29 (1) : 41-46

Dana JD (1846-1849) Zoophytes. U.S. Exploring Exped, 1838-1842, (7) pp 1-740

Edwards H Milne, Haime J (1857-1860) Histoire naturelle des Coralliaires, Paris (3) (1860) pp 1-219

Faure G (1982) Recherche sur les peuplements de scléractiniaires des récifs coralliens des Mascareignes. Thèse es sciences, Univ Aix-Marseille II, (2) pp 1-206

Le Renard J, Conruyt N (1994) On the representation of observational data used for classification and identification of natural objects, IFCS'93, Lecture Notes in Artificial Intelligence, Springer Verlag, pp 308-315

Manago M, Althoff KD, Auriol E, Traphöner R, Wess S, Conruyt N, Maurer F (1993) Induction and reasoning from cases. First European workshop on case-based reasoning (EWCBR-93), MM Richter, S Wess, KD Althoff and F Maurer (Eds.), Springer Verlag, (2)

Pankhurst RJ (1970) A computer program for generating diagnostic keys. Computer J. 13 : 145-151

Pankhurst RJ (1991) Practical taxonomic computing. Cambridge University Press, Cambridge, pp 1-202

Quinlan JR (1986) Induction of decision trees. Machine Learning 1 : 81-106

Shannon CE (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Veron JEN, Pichon M, Wijsman-Best M (1976-1984) Scleractinia of eastern australia, vol. I to V, Australian Institute of Marine Science Monograph Series

Veron JEN, Pichon M (1976) Scleractinia of eastern australia, vol. I, Part I, Australian Institute of Marine Science Monograph Series

Wallace CC, Dallwitz MJ (1981) Writing coral identification keys that work. Proc 4th Int Coral Reef Symp Manila 2 :187-190