

# Managing complex knowledge in natural sciences

Noël Conruyt, David Grosser

IREMIA, Institut de REcherche en Mathématiques et Informatique Appliquées  
University of La Réunion  
15, av. René Cassin – 97715 Saint-Denis, Messag. Cedex 9, France  
{Conruyt, Grosser}@univ-reunion.fr

**Abstract.** In many fields dependant upon complex observation, the structuring, depiction and treatment of knowledge can be of great complexity. For example in Systematics, the scientific discipline that investigates bio-diversity, the descriptions of specimens are often highly structured (composite objects, taxonomic attributes), noisy (erroneous or unknown data), and polymorphous (variable or imprecise data). In this paper, we present IKBS, an Iterative Knowledge Base System for dealing with such complex phenomena. The originality of this system is to implement the scientific method in biology: experimenting (learning rules from examples) and testing (identifying new individuals, improving the initial model and descriptions). This methodology is applied in the following ways in IKBS: 1 - Knowledge is acquired through a descriptive model that suits the semantic demand of experts. 2 - Knowledge is processed with an algorithm derived from C4.5 in order to take into account structured knowledge introduced in the previous descriptive model of the domain. 3 - Knowledge is refined through the use of an iterative process to evaluate the robustness of the descriptive model and descriptions. The IKBS system is presented here as a life science application facilitating the identification of coral specimens of the family *Pocilloporidae*.

## 1. Introduction

In the natural sciences, data to be processed may be more complex than in other fields. In Systematics, attributes that describe organisms are numerous ( $> 100$ ) compared with the number of individuals by class which is mostly not representative ( $< 10$ ): the domain to describe is established deterministically (empirically) rather than probabilistically (statistically) [14]. In such domains, we must take into account diversity and incompleteness, and the exception is the only valid rule.

Learning systems intended to facilitate classification (class definition) and identification of natural organisms must adapt themselves to the representation and process of such reality.

For the necessities of representation, taking into account the structuring of biological knowledge [2], [5] is a progress that allows to consider useful common sense background knowledge in order to acquire, manage and process complex knowledge in a more elegant and efficient way.

The identification procedure that is described in this paper takes care of structured descriptions intelligently by reducing the number of eligible criteria for information

gain calculation and manages coherent consultations through a guide to observation (web questionnaire).

Nevertheless, the problem we are faced with in Systematics is more difficult: good identifications depend on previously good classifications from experts, and also good descriptions from other biologists. Nature is so conceived that giving a name to organisms can also be difficult for experts (synonymies' problem), especially when there is a great intra-specific variation. This is the case in coral taxonomy where the number of named species in the world is uncertain [18]. Thus, managing complex knowledge in natural sciences means to cope with such evolving knowledge.

We have designed an Iterative Knowledge Base System to build knowledge bases in natural sciences that responds to these requirements. The main goal of IKBS is to produce quality descriptions which is a key factor for getting better results in identification process [11] and avoid future revisions.

## 2. Methodology

In the computer sciences, knowledge is a controversial term [Kodratoff, 1997]. We thus offer a working definition for our purpose in biology that consists of three kinds of knowledge: domain, instantiated and derived.

*Domain knowledge* (or background knowledge) relates to the definition of what is *observable*, i.e., build a descriptive model that corresponds to the modeling of data, or metadata [6]. *Instantiated knowledge* refers to the description of *observed* instances (case descriptions). *Derived knowledge* can be compared with produced hypotheses (cluster definitions, decision trees, rules, identification) discovered from domain and instantiated knowledge. Obviously, knowledge is also grounded in expert's mind and what is "extracted" is but a minimal part of his or her experience.

Knowledge Discovery methodology views knowledge as an output of a linear process of input data handling [8]. In biological domains, our emphasis is placed on a different interpretation of knowledge which consists of both input (domain and instantiated) and output (derived). This viewpoint is more relevant to Case-Based Reasoning methodology: i.e. the CBR cycle described in [1] with an extensive use of domain knowledge in the processing phase.

In practice, knowledge is extracted with IKBS by a cyclical process, divided into three parts:

1. Knowledge acquisition:
  - Acquire a descriptive model (domain knowledge or observable facts),
  - Acquire descriptions (observed facts or cases),
2. Knowledge processing:
  - Generate classification rules with decision tree induction,
  - Identify new observations (unknown specimens) with case-based reasoning,
3. Knowledge validation and refinement:
  - Verify the origin of misidentifications by analyzing differences of interpretation between the expert and the users of the knowledge base,
  - Iterate on the definition of the descriptive model (characters), update old cases.

For experts in biology, this approach is well suited to the natural process of their knowledge acquisition (conjecture and test) [16]:

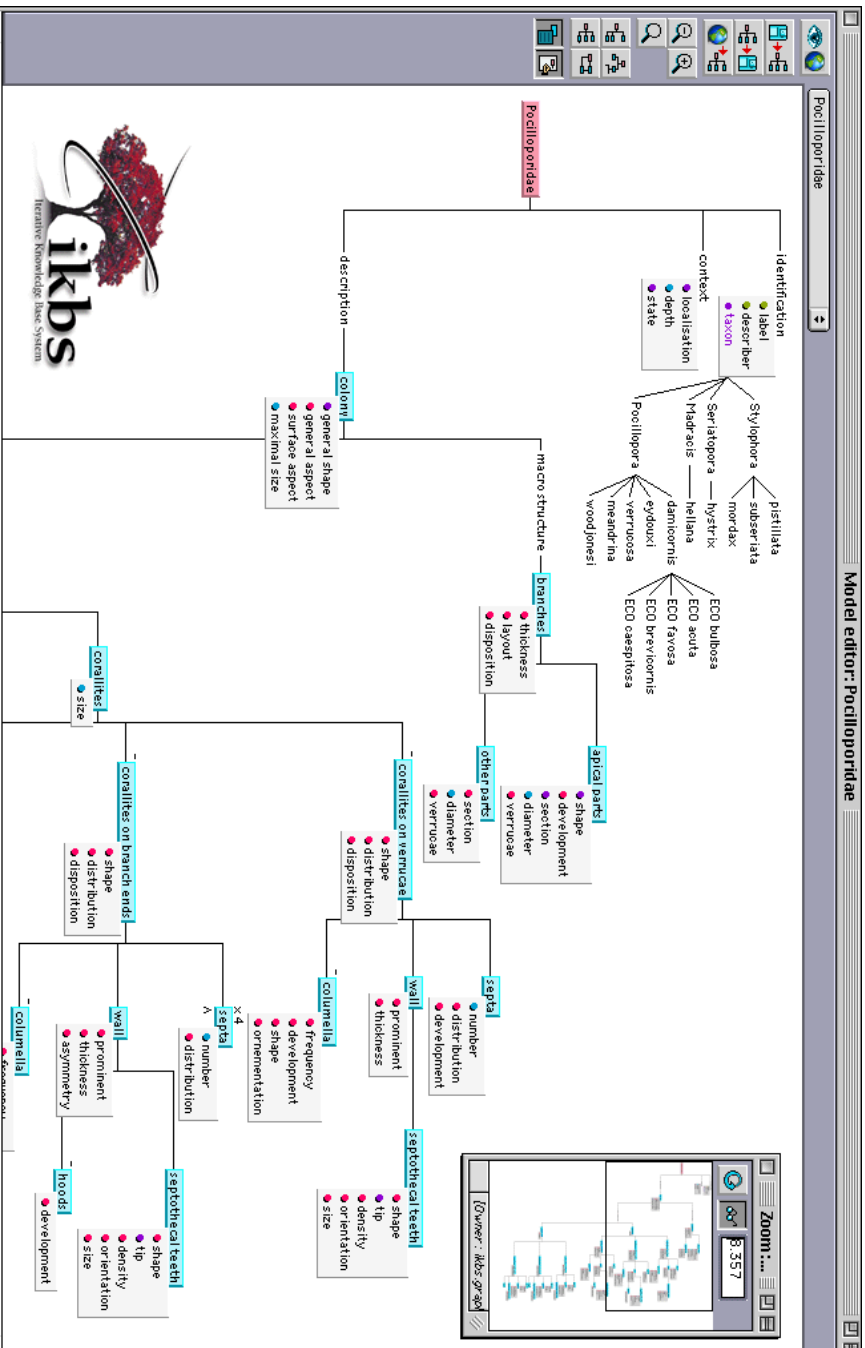


Fig. 1. Part of the descriptive model of the Family Pocilloporidae

1. Observe and familiarize oneself,
2. Represent observations, i.e. make descriptions,
3. Build hypotheses from descriptions (pre-classified), i.e. generate identification keys,
4. Test and experiment them with new observations, i.e. identify new specimens,
5. Refine their initial knowledge (new characters, cases and classifications).

The last point of the method is fundamental because the building of a knowledge base in natural sciences is very difficult. This was our experience in applications such as diagnosis of plant pathologies [12]. It is hard for experts to define the best representation of reality at once in a descriptive model. The challenge is to acquire the best character definitions and illustrations leading to interpretations of observations understood by anyone consulting the knowledge base.

### **3. Knowledge acquisition**

This part of the methodology is very important to acquire a case base with good quality descriptions, i.e. that are well structured in different dimensions with all the required information, with characters, illustrations and comments that are easily comprehensible for other biologists.

#### **3.1 The descriptive model**

The descriptive model represents all the *observable* characteristics (objects, attributes and values) pertaining to individuals belonging to a particular domain. It is organized in a structured scheme, the name of the domain being at the root of a description tree. Each node of the tree is an object (a component of the individual) defined by a list of attributes with their respective possible values. Designing a descriptive model is essentially an expert task.

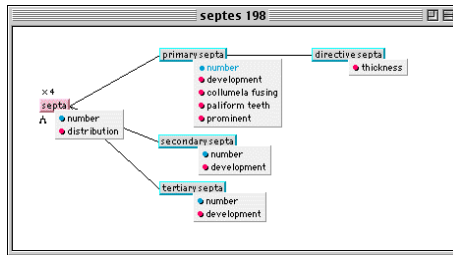
For helping them, we have set up logical rules for case description covering: decomposition, viewpoint, iteration, specialization, contextual conditions, etc. [11]. These rules were constructed from the analysis of expert's process of creating monographs of organisms or diseases.

To serve as an example, we present the descriptive model of one of the world's most widespread family of corals: *Pocilloporidae* [7] (see Fig. 1). 51 objects and 120 attributes have been defined by the expert. With them, biologists are able to describe 4 genus and 14 species (see attribute called "taxon" in Fig. 1).

There are multiple benefits in such a representation. Viewpoints divide the descriptive model into homogeneous parts, thus giving a frame of reference for describing organisms at a particular level of observation (see objects identification, context, description, macro and micro structure in Fig. 1).

Sub-components introduce modularity into the descriptions making it possible to structure the domain from most general to most particular parts. This object representation of specimens is semantically better than the flat feature-value one: in the former, local descriptions of attributes depend on the existence of parent objects, although in the latter the defined characters are independent of one another. Some of the possibly missing objects are marked with a minus sign (e.g. columella).

Fig. 1 shows the partitioning dimension of objects (subpart links for disjoint classes). For some of them (i.e. septa), other dimensions such as multi-instantiation (x symbol) and specialization (^ symbol) of objects can be seen. The former enables users to describe several sorts of the same object by descriptive iteration (there are 4 possible instances for septa in Fig. 1) and the latter lets users name each sort with the help of the following classification tree of objects (specialization links in Fig. 2).



**Fig. 2.** Classification tree of object “septas”

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 by 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 (see below).

### 3.2 The descriptions

Starting from the selection of a descriptive model, the program automatically generates a questionnaire [4]. It permits the less informed biologists, as the expert, to acquire personal descriptions and create a case base. An identification name is associated to each observation in order to form a description or a case (Fig. 3).

The description process generates sub-trees of the descriptive model (Fig. 1 and Fig. 3). Therefore, observed descriptions can be directly compared to one another by leafing through page by page: this navigation process is easier than viewing different lists of attribute-value pairs.

In Fig. 3, we illustrate possibilities of IKBS for rendering complete and comprehensive descriptions of a given sample.

Different types of attribute are used: taxonomic ones (e.g. general shape of object colony), numerical intervals (e.g. diameter of apical parts) and multi-nominal values (e.g. section of apical parts). The latter shows variation in objects displaying a set of multiple elements.

The visualization of objects differs graphically according to their status: black if present, black with a cross if absent, dimmed if unknown (see object “hood” at the bottom-right side of Fig. 1 and Fig. 3).

At last, an object can be specialized (e.g. the septa of calices from apical parts, see Fig. 1): the result is a substitution of its name by a more precise one (e.g. primary septa, see Fig. 3) with its associated attributes (inherited or not, see Fig. 2).

It is important for the user to visualize structured descriptions: so doing brings better clarity and comprehensibility to the acquisition phase. This is the most important part of our methodology for acquiring good results of classification and identification.



## 4. Knowledge processing

This section highlights how usual inductive learning algorithms can be stretched to complex data processing by using domain knowledge to generate accurate and meaningful decision trees (from pre-classified examples).

### 4.1 Tree-based classification using domain knowledge

Starting from the well known decision tree builder algorithm C4.5 [17] which works on discrete and continuous attributes, IKBS extends some functionality of this algorithm for dealing with:

1. Structured objects
2. Taxonomic attribute-values
3. Multi-valued attributes:

Let  $E = \{\omega_1, \dots, \omega_n\}$  be a set of observed examples,  $M = \{N, Y\}$ , a set of observable components and attributes defined in the descriptive model with  $N = \{n_1, \dots, n_m\}$  a set of structured components and  $A = \{A_1, \dots, A_p\}$  a set of attributes depending on  $N$ .

Let  $\text{dom}(A)$  be the definition domain (range) of  $A$ .

#### Structured objects

The algorithm for building decision trees from structured objects is the following:

```
BuildDecisionTree (E, M)
  Y = SelectClassifier (root(M))
  BuildTree(E, Y)
end BuildDecisionTree

BuildTree(E, Y)
  if stop Criterion(E, Y) then BuildLeaf(E)
  else
    A = BestTest(E, Y) // A = y(n)
    di = BuildNode(A)
    Y = FilterClassifier(A) // depending on type of A
    partitioning(di) = R(E)
    // R(E) :  $\forall \omega \in E, Q(v_i, A(\omega)) = 1 \Leftrightarrow \omega \in E_i$ 
    for each Ei  $\in$  partitioning(di)
      BuildBranch(vi)
      if (A = exist(n))  $\wedge$  (vi = "Present")
        Y = SelectClassifier (n)
      end if
      BuildTree (Ei, Y)
    end for each
  end if
end BuildTree
```

```

SelectClassifier (r)
  Y' = ∅
  if (possiblyAbsent(r) = "yes") then
    Y' = Y' ∪ {exist(r)}
  else
    Y' = Y' ∪ Att(r)
    for each nf ∈ depend(r)
      SelectClassifier (nf)
    end for each
  end if
  return Y'
end SelectClassifier

```

The original aspect of the algorithm is the classifier's selection function. The tree of the descriptive model is followed from root to leaves, component by component in depth search first. If one of it *can* be absent (e.g. calices on verrucae of Fig. 1), an "exist component" test is dynamically generated and placed in the eligible classifiers' list with values "Present" or "Absent". The sub-tree of this object is not yet visited in order to avoid inapplicable sub-objects and attributes as other classifiers. On the other hand, if components are always present (e.g. septa), dependent attributes are placed in this list.

In the identification process (see further), if the test *exist* of a component is chosen as the "best" one and the user answers that it is really present, then the classifier's selection function is recursively called on the sub-tree of the descriptive model.

#### Taxonomic attribute-values

For attributes which values are structured by relations of hierarchical type (classified values), an extension of the discrete classifier partitioning process is proposed.

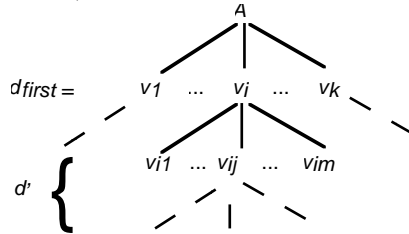


Fig. 4. Classified values of attribute A

The method consists, when such a classifier is selected, in creating a set of partitions corresponding to the first level of the hierarchy (noted  $d_{first} = \{v_1, \dots, v_i, \dots, v_k\}$  with  $k$  elements). Each case is assigned to the partition that generalizes its value. Let  $A$  be a taxonomic attribute with the domain  $d = \{v_1, \dots, v_i, \dots, v_n\}$  of  $n$  modalities and  $d' = \{v_{i1}, \dots, v_{ij}, \dots, v_{im}\} \subset d$  is a subtree of  $m$  submodalities of  $v_i$  [Fig. 4]:

Let  $Q$  be a Boolean application (called question) which determines if the modality  $v_i$  generalizes a value  $v_{ij}$ .  $Q$  is defined by:

$$Q(v_i, v_{ij}) = 1, \text{ if } v_{ij} \in d \cup \{v_i\} \text{ else } Q(v_i, v_{ij}) = 0$$

Then, we can generate  $k$  partitions from  $d_{first}$ :

$$E_{A1} = \{\omega \in E_A / Q(v_1, A(\omega)) = 1\}, \dots, E_{Ak} = \{\omega \in E_A / Q(v_k, A(\omega)) = 1\}$$

In the next step, we create temporarily  $k$  attributes  $\{A_1, \dots, A_j, \dots, A_k\}$  in each partition  $E_{A_1}, \dots, E_{A_k}$  with a set of modalities defined by the subvalues of  $\{v_1, \dots, v_i, \dots, v_k\}$ . These ones can be picked by the test function (information gain, gain ratio) and the method is recursively reapplied.

### Multi-valued attributes

When modeling the descriptive model, a discrete attribute (nominal or taxonomic) can be defined as multi-valued. It can express doubt (disjunction of imprecision) or the simultaneous presence of states (conjunction of variation) like in the following expression:

$$v = (v_{11} \& \dots v_{1i} \dots \& v_{1m}) \mid \dots \mid (v_{j1} \& \dots v_{ji} \dots \& v_{jn}) \dots \mid (v_{k1} \& \dots v_{ki} \dots \& v_{kp})$$

where  $cf_j = (v_{j1} \& \dots v_{ji} \dots \& v_{jn})$

Depending on the semantic associated with a conjunctive form of a case ( $cf$ ), IKBS can apply three processing methods:

1. If  $cf$  is true information (association of co-existing facts), create  $k$  partitions corresponding to each conjunction of  $v$ , and dispatch cases with such value in each partition:  $cf$  is seen as a new possible value of  $\text{dom}(A)$ .
2. If  $cf$  expresses fuzzy information (the intrinsic variability of multiple objects is an adding source of noise), treat conjunctions as disjunction.
3. Allow the user to customize the degree of similarity  $\partial$  between two conjunctive forms.

The default method is the third one with  $\partial = 1$  because it gives a good compromise between the tree size (number of nodes) and the discrimination accuracy. Indeed, the first method don't generate a deep tree, but carries a major risk of misidentification: each  $cf$  of the selected attribute at a node of the decision tree must match exactly the  $cf$  of the tested case. The third method is more flexible because it makes a fuzzy matching for dispatching cases in each partition, depending on the number of differences between the two conjunctive forms and  $\partial$ .

## 4.2 Identification process

Given a set of examples, IKBS dynamically extracts the most efficient criteria from the ordered list of tests after each answer of the user. The cases are selected from this reply. If the answer is unknown, the second most discriminate test is proposed to the user, and so on. This procedure is the same as in KATE [13]. Nevertheless, IKBS processes cases that are in an object oriented formalism, while KATE starts from cases represented in a data table. In the former, the "exist (objects)" tests are directly exploitable at a node of the decision tree, while in the latter, these tests are deduced from the appearance of at least one not-applicable value in the object's attribute column of the table. Our approach is semantically better because it guarantees that the inapplicability of an attribute's value depends on the absence of an object and not the contrary.

An illustration of a decision tree built with 30 training cases is shown in Fig. 5. The numbers refer to previously discussed classifiers' types: taxonomic attributes (1 and 2 highlight the use of the same classifier at two different levels of the values' hierarchy), multi-valued attributes (case *bulbosa* with size of spines "short&long" goes in three branches (3)) and structured objects (4).

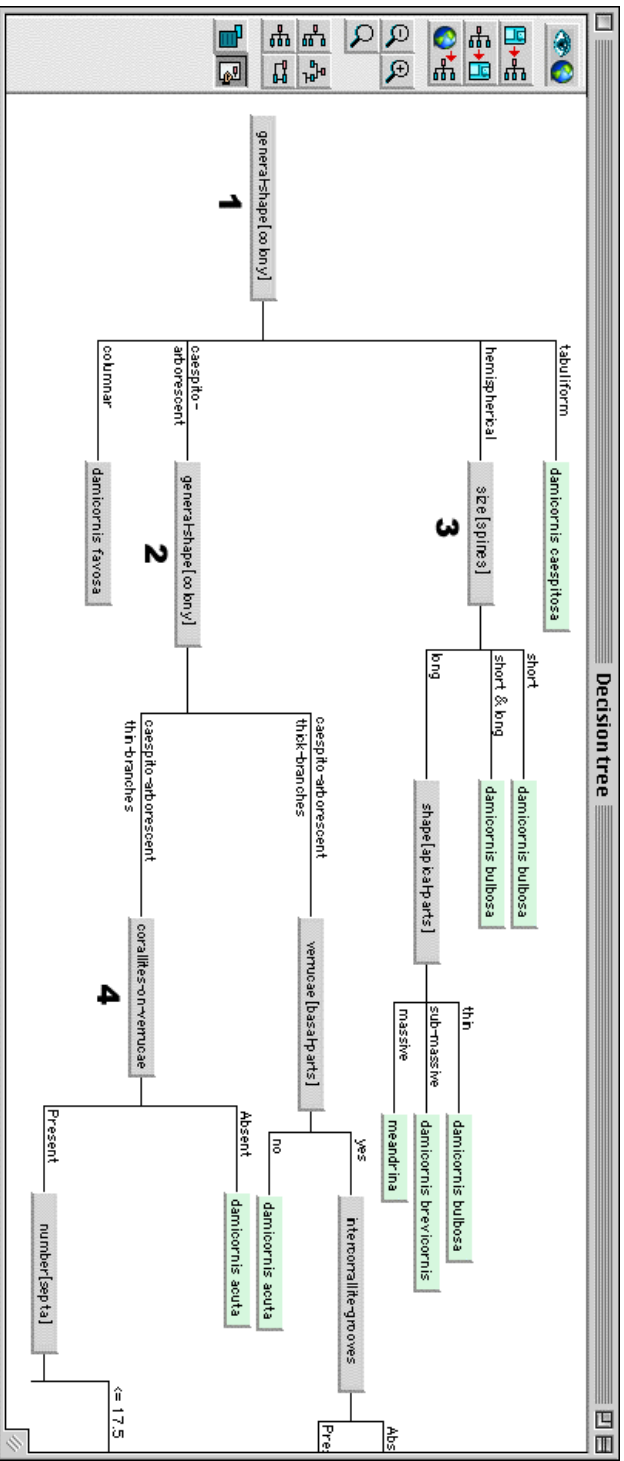


Fig. 5. Part of a decision tree that makes use of domain knowledge

## 5. Validation and refinement

We experimented on corals to test the reliability of IKBS identification with different users. We tested two consecutive descriptive models in a sub-domain of *Pocilloporidae*: the genus *Pocillopora* (9 species and ecomorphs). The validation of both descriptive models was *qualitative*. It led to modification of the initial descriptive model ( $dm_1$ ) and case base to the one shown in this paper ( $dm_2$ ). The first test with  $dm_1$  is called A. Later, another test B on  $dm_2$  was carried out. The experiments were made with a sample of 15 specimens of the Genus, each one of them being described completely in both descriptive models by 3 different biologists ( $x_1, x_2, x_3$ ), and the expert (E). With this training set of 60 cases, the expert added 22 other descriptions of *Pocillopora* (37 expert cases). The four experiments that were led with IKBS are the following:

- $A_1$ : 15 cases of  $x_n$  tested against 37 cases of E.
- $A_2$ : 15 cases of  $x_n$  tested against 67 cases (E+other  $x_n$ ).
- $B_1$ : 15 cases of  $x_n$  tested against 37 cases of E.
- $B_2$ : 15 cases of  $x_n$  tested against 67 cases (E+other  $x_n$ ).

The results on 15 consultations are shown in table 1.

**Table 1.** Number of good identifications with IKBS

	$x_1$	$x_2$	$x_3$
$A_1$	6	7	8
$A_2$	9	8	10
$B_1$	9	10	11
$B_2$	11	10	12

The results show that updating the first descriptive model and case base on *Pocillopora* brought better results. When testing the expert training set, it gave 20% (3/15) of improvement in identification process (from 46% to 66%). If we integrate other biologists' descriptions of the same specimens in the reference case base, the score goes up from 60% to 73%.

The reasons of these improvements are principally:

1. the expert was able to detect inconsistencies in the first case base (omissions or errors in descriptions) and descriptive model (misunderstood characters, faulty illustrations). He could verify the answers of other biologists in regards to decision tree questioning that lead to misidentifications. He noticed the difficulties of interpretations of observation of specimens on some noisy comparative attributes and refined them into a new descriptive model.

2. Consequently, the expert, aware of the importance of transmitting his knowledge to other biologists, postulates more precise and relevant characters that may be easier to observe and/or offer less ambiguous values (easier to interpret) in his descriptive model. For example, he will refine on the basis of mutually exclusive values, monosemic attributes, frames of reference, warning signals, enhanced illustrations.

## 6. Discussion

IKBS has been implemented in Java language and is fully operational on (<http://www.univ-reunion.fr/~ikbs>). Experts unaccustomed to computers are able to model and describe, and any non-specialists interested in the field can describe and identify new observations. IKBS is used directly by experts for creating descriptive models and filling cases without any help from a computer scientist. They find the interface very pleasant and enjoy the effectiveness of the tool.

In our methodology, it is important that the case base contains descriptions of specimens made by biologists other than the expert. This, in order to counterbalance his interpretation of observations (inter-observer variation) when consulting the knowledge base. The results of the identification process are more dependable when we mix descriptions of different users for the same specimens (shown in Table 1). As they are labeled with the correct identification name from the expert, we can integrate the noise due to misinterpretations from end-users directly into the case base.

Similarly, because of the intra-specific variability, the number of described specimens by species must be increased. Insofar as *Pocilloporidæ* is concerned, this family is one of the sixteen families of corals containing the greatest intra-variability, and its complex diversity was covered with detailed precision.

The difficulty arises due to the number of attributes applicable to each case. Thus, the building of an exhaustive knowledge base is time-consuming for describers: updating a case with the latest descriptive model on *Pocilloporidæ* requires nearly a whole day's work!

## 7. Related work

In other domains such as botany and zoology, some researchers have come up with solutions for coding descriptions [5]. Their programs enable to compare descriptions and facilitate identification process from databases [10], [15].

In Case-Based Reasoning methodology, IKBS can be compared with AcknoSoft's KATE, Isoft's RECALL and TecInno's CBR-Works. These decision support systems have been designed to cope with industrial fields and very large databases [3]. In the life sciences, our objective is to deal with more complex descriptions and less data (cases) by class.

## 8. Conclusions and future work

In collaboration with three experts, we are presently experimenting with IKBS on three other families of corals of the Mascarene archipelago (*Fungiidae*, *Poritidae*, *Thamnasteriidae*). The meticulous choice of terms, drawings and images seems decisive for generating a dependable knowledge base and managing the complexity of natural objects.

This is why we are designing IKBS to build cooperative knowledge bases. The aim is to encourage experts to draw up a common thesaurus of vocabulary and illustrations (i.e. the questionnaire) on the same Family.

Collections of specimens, like experts, are distributed around the world. Thanks to satellite high-speed broadband networks, we have been able to demonstrate Telesystematics using video-conferencing and IKBS. At ATM Developments'98, experts were able to share their interpretations of observations of specimens under a microscopic examination synchronously between La Reunion (South-West of Indian Ocean) and Rennes (France).

Nowadays, expertise in natural sciences is precious (it becomes very rare). It is therefore urgent to develop tools that will ensure that expertise be collected and safeguarded for transmission to future generations. If this is not done, we will be left only with monographic descriptions and museum collections.

## Acknowledgement

We would like to thank the French experts G. Faure, M. Pichon and M. Guillaume for their valuable contributions on the applications on corals. We are also grateful to Philippe Sills for English improvement. This work is supported by the Conseil Régional of La Reunion Island.

## References

1. Aamodt A., Plaza E., Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, *AI Communications* 7(1): 39-59, 1994.
2. Allkin R., Handling taxonomic descriptions by computer, In; Allkin R. and Bisby F.A. (eds.), *Databases in Systematics*. Systematics Association London, Academic Press, 26: 263-278, 1984.
3. Althoff K. D., Auriol E., Barletta R., Manago M., A review of Industrial Case-Based Reasoning Tools, *AI Intelligence*, Oxford, 1995.
4. Conruyt N., Grosser D., Faure G. Ingénierie des connaissances en Sciences de la vie: application à la systématique des coraux des Mascareignes. *Journées Ingénierie des Connaissances et Apprentissage Automatique (JICAA'97)*, Roscoff, pages 539-566, 1997.
5. Dallwitz M.J., Paine T.A., Zurcher E.J., User's guide to the DELTA System. A general system for processing taxonomic descriptions, Canberra: CSIRO, Div. Entomol., 4th ed., 1993.
6. Diederich J.R., Milton J., Creating domain specific metadata for scientific data and knowledge bases, *IEEE Trans., Knowledge Data Engineering* 3(4): 421-434, 1991.

7. Faure G., Recherche sur les peuplements de scléactiniaires des récifs coralliens des Mascareignes. Thèse es sciences, Univ Aix-Marseille II, 1982.
8. Fayyad U., Piatetsky-Shapiro G., Padhraic S., From Data Mining to Knowledge Discovery in Databases, *AI magazine*, 17(3): 37-54, Fall 1996.
9. Kodratoff Y. L'extraction de connaissances à partir des données. Journées Ingénierie des Connaissances et Apprentissage Automatique (JICAA'97), Roscoff, pages 539-566, 1997.
10. Lebbe J., Systématique et informatique. Systématique et biodiversité, Bourgoin T. (Ed), *Biosystema*, 13:71-79, Paris, 1995.
11. Le Renard J., Conruyt N. On the representation of observational data used for classification and identification of natural objects. IFCS'93, Lecture notes in Artificial Intelligence, Springer-Verlag, pages 308-315, 1994.
12. Manago M., Conruyt N. Using Information Technology to Solve Real World Problems, *Lecture Notes in Computer Science subseries*, 622: 22-37, Springer Verlag, 1992.
13. Manago M., Althoff K.D., Auriol E., Traphoner R., Wess S., Conruyt N., Maurer F., 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), 1993.
14. Mingers J. Expert Systems – Rule induction with statistical data. *Journal of the operational research society*. 38(1): 39-47, 1987.
15. Pankhurst R.J., Practical taxonomic computing. Cambridge Univ. Press, Cambridge, 1991.
16. Popper K.R., La logique de la découverte scientifique. Payot (Eds.) Press, Paris, 1973.
17. Quinlan J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, CA, 1993.
18. Veron J.E.N., Pichon M., Scleractinia of eastern australia, vol. I, Part I, Australian Institute of Marine Science Monograph Series, 1976.