ELSEVIER

Artificial Intelligence 91 (1997) 257-279

Artificial Intelligence

ReTAX: a step in the automation of taxonomic revision *

Eugenio Alberdi *, Derek H. Sleeman [1]

Computing Science Department, King's College, University of Aberdeen, Old Aberdeen AB9 2UE, Scotland, United Kingdom

Received December 1995; revised August 1996

Abstract

Classification, and particularly taxonomic revision, have not been generally addressed by computational models of scientific discovery. In this paper we present a framework for the automation of taxonomic revision in biological domains. This framework views taxonomy formation as an interaction of: (a) observation; (b) creation and structuring of a taxonomic hierarchy; (c) identification of relevant taxonomic descriptors; and (d) use of background knowledge. We describe a prototype system for taxonomic revision, ReTAX, which implements relevant aspects of such a framework. ReTAX receives as input a pre-established taxonomy, and is presented with new items which contradict in some way the original classification. Using a set of consistency criteria, ReTAX identifies the inconsistencies between the new information and the taxonomy. The system then applies a set of refinement operators to modify the taxonomy and resolve the inconsistencies. ReTAX has been tested on a botanical domain, replicating taxonomic revisions which had been suggested by professional botanists for the family Ericaceae. Finally, we propose extensions to ReTAX, which we hope will enable us to further develop the framework, and subsequently create an aid which taxonomists can use to revise existing taxonomies. © 1997 Elsevier Science B.V.

Keywords: Classification; Taxonomy; Theory revision; Historical simulation; Empirical study; Scientific discovery

* A preliminary and abbreviated version of this article was presented in the *AAAI Spring Symposium on Systematic Methods of Scientific Discovery*, Stanford University, Stanford, CA (1995).

* Corresponding author. E-mail: e.alberdi@abdn.ac.uk.

[1] E-mail: sleeman@csd.abdn.ac.uk.

PII S0004-3702(97)00010-6

1. Introduction

Classification is an essential process in science. The creation (and refinement) of taxonomies is basic to many scientific tasks. Activities such as theory formation, law induction, and experimentation undoubtedly rely upon a sound classification of elements, e.g., physical, chemical, biological. More specifically, taxonomic *revision* can play an important role in scientific discovery. The detection of an anomalous experience which cannot be accommodated by an established scientific taxonomy may lead to the revision of the taxonomy; but it may also affect other scientific knowledge. In other words, the revision of taxonomic knowledge may challenge related scientific principles (or theories), which should then be reconsidered and investigated experimentally (Sokal [25]). As noted by Shrager and Langley [22], the formation (and revision) of taxonomies is one of the principal activities which should be considered in the computational simulation of scientific discovery.

In this paper, we focus on revision procedures related to biological classification. Biologists nowadays make an extensive use of computational methods, such as biological *databases* (Allkin and Bisby [4]), methods for biological *identification* (Pankhurst [21]), and numerical techniques for *taxonomy formation* (Sokal and Sneath [26]). However, to date, no computational system has been created to deal explicitly with taxonomic revision. Similarly, the task has not been addressed, in general, by the subfield of scientific discovery.

Current taxonomic practice is characterised by a great deal of *controversy*. Disputes have ranged from arguments about the nature of the categories being classified (Mayr [17]), to bitter battles among schools of thought regarding various methodological and theoretical issues (Hull [13]). Hence, one may view the need for a rational approach which provides independent criteria to assist in the various argumentations. The use of AI techniques seems a reasonable approach. However, a general reservation of taxonomists is that many aspects of their decision making cannot be appropriately captured by a computational model, as their reasoning is often based on highly subjective and intuitive considerations. We recognise that there may always be some subjectivity associated with classification, but believe that some principles can be abstracted to at least partially formalise the task. In this paper, we present a framework which outlines a series of such principles. This framework is aimed at generating AI tools which partially automate the taxonomic process. We have abstracted this framework from the following experience:

- A psychological study of *category induction*, in the domain of plant taxonomy, which we conducted at Aberdeen [2,3]. The purpose of this study was to identify strategies used by taxonomists when they encounter unexpected items which contradict their current beliefs. The results of the study were modelled computationally in a system called Proto-ReTAX.
- ReTAX, a *prototype system* for the revision of taxonomies. Some of the mechanisms implemented in the system are based on our psychological results, as simulated in Proto-ReTAX. ReTAX has been applied to replicate taxonomic revisions which have taken place historically in the botanical family Ericaceae (Middleton and Wilcock [19]). (See Sections 4 and 5.)

- Discussions which we have had, prior to and subsequently to the development of ReTAX, with *professional taxonomists* working in different biological areas (botany, entomology, bacteriology).

In the following section (Section 2), we give a brief introduction to taxonomic revision. The main focus of this paper is the overall framework (Section 3). We discuss how the framework has been implemented in ReTAX (Sections 4 and 5); how it relates to earlier research on revision and taxonomy formation (Section 6); and how we are planning to further develop it in the future (Section 7).

2. The revision of taxonomies

The purpose of *biological classification* is "to provide an information system, one that provides comparative information about organisms to biologists and to the general public" [1, p. 11].

The principal outcome of the taxonomic process is usually the grouping of organisms in an embedded hierarchical structure. A hierarchy reflects the relationships among different groups of elements. A group at any level of the hierarchy is known as a *taxon* (plural: *taxa*). The grouping of organisms into taxa is based on the similarities and differences observed among organisms with respect to a series of botanical aspects or features, known in the taxonomic literature as *characters*.[2]

Taxonomists describe their job as a never ending task [1, p. 13]. In fact, the history of plant taxonomy can be described as a cumulative process in which new classification systems supersede earlier ones as new biological information and taxonomic methods become available. Even now, plant taxonomy is not a completed activity, as there are still many types of plants in the world which have not been recognised and classified, and new biological findings continually shed new light on different aspects of plants.

Although taxonomic revision is unavoidable, the *stability* of classifications is often desirable. Classifications generally serve practical purposes, and unstable taxonomies which are modified with excessive frequency are of little use. As Simpson puts it, "there must be some compromise between the usefulness of up-to-date classifications and the usefulness of stable classifications" [23, p. 112].

Finally, the revision of taxonomies often involves a shift of focus from one type of taxonomic information (characters) to an alternative one. Sokal [25], for example, reports cases in which excessive emphasis on certain aspects of organisms had led to erroneous classifications, and so the modification of these flawed taxonomies was achieved by moving the focus of attention to alternative sets of characters which had been previously overlooked.

[2] Throughout this paper we will loosely use the term "character" as synonymous with "descriptor", "feature" and "attribute". We are aware that these terms are used with different meanings in the taxonomic literature. However we have favoured a usage consistent with AI terminology.

3. A framework for the automation of taxonomic revision

We view taxonomy formation as a task consisting of, at least, four interacting processes:

(a) Exhaustive *observation* of the specimens to be classified. Although essential, this process is difficult to emulate computationally. Information about specimens will need to be provided either by the user or by a biological database.
(b) Identification of the most important *descriptors* (characters) to represent taxa.
(c) *Grouping* specimens and assigning them to appropriate taxa at the different levels of a hierarchy. When dealing with revision, this may involve restructuring the taxonomy (i.e., regrouping the items).
(d) Use of *background knowledge* (or meta-knowledge). By background knowledge we mean: basic biological principles, criteria which establish when a taxonomy is consistent, information regarding the variability of specimens and taxa (e.g., genetic, geographical, etc.), knowledge about specimens and taxa which can be unambiguously classified, etc.

We can characterise taxonomy revision as a process which is needed when the following situations occur: an existing taxonomy may be incomplete/inconsistent or the description of the specimens to be accommodated in the taxonomy might also be incomplete/inconsistent. Given such a situation, there are three possible ways in which a computational system can tackle the task:

(1) *Assume* that the *taxonomy* is complete/consistent and see what changes in the characters (both adding and dropping characters) one would need to observe for a specimen in order to accommodate the specimen in the taxa at the different levels of the hierarchy.
(2) *Assume* the *specimen* is completely described and again note the changes in the taxonomy required to place the specimen in the taxa. The changes to the taxonomy can either be representational (i.e., related to process "(b)" above) or structural (i.e., related to process "(c)").
(3) *Assume* that *both* the specimen *and* the taxonomy may be incompletely/inconsistently described and note the changes needed in the item and the taxonomy to resolve the inconsistency. This, in fact, is the likeliest situation to arise.

During the above processes, one would plan to observe what changes the expert taxonomist makes, and would also encourage him/her to articulate explanations for the choices made. Our task, as cognitive scientists, would then be to record this argumentation and to represent it as background knowledge (or meta-knowledge) in a subsequent version of the workbench. Such background knowledge would then assist the system in deciding when a taxonomy is "consistent", whether the description of a new item conflicts with an existing taxonomy, what changes are most appropriate to resolve a given taxonomy, what changes (or groupings) are not desirable in a particular taxonomy, etc.

As will be evident in the next section, ReTAX, the prototype system we have developed, firstly attempts representational refinements of the taxonomy (i.e., process "(b)" above); and, secondly, addresses taxonomy restructuring (i.e., process "(c)"). Additionally, ReTAX has been designed to function under assumption number (2), mentioned above in this section. Further assumptions and simplifications are noted in Section 4.1.

4. ReTAX

4.1. Assumptions and simplifications

In a first approach to modelling taxonomic revision, we used a series of assumptions in the design of ReTAX, to make the task tractable. In particular, we can note the following simplifications:

- ReTAX deals with taxa represented as *monothetic* groups (as opposed to *polythetic* taxa). In a monothetic taxon, all its members share the values for *at least* one feature which is necessary to distinguish the taxon from other taxa at the same level of the hierarchy. In contrast, a polythetic taxon consists of organisms which are more similar to each other than they are to members of other taxa; but there need not be a single feature for which all members of a taxon *must* have the same value. The use of monothetic taxa makes membership recognition procedures simpler (see Section 4.6). However, this is not very realistic as it ignores the great variability existing in nature (e.g., many biological groups can only be described as polythetic taxa).
- ReTAX has been designed to identify the inconsistencies between an existing taxonomy and novel information, and to modify the taxonomy to resolve the inconsistencies. In order to simulate the *historical* evolution of a taxonomy, ReTAX is given an existing taxonomy, and a series of specimens which *traditionally* have been assigned to particular taxa in the taxonomy. The new description of the specimens includes aspects of the plants which had not been considered when the original hierarchy was constructed.
- Because a historical data set has been utilised, ReTAX's performance is partly *supervised* (as opposed to *unsupervised* [15]). In other words, the system requires the user to provide a classification for the items which motivate the revision. Real taxonomic practice is certainly not supervised. But, in the context of ReTAX's historical simulation, the classification provided by the user is meant to reproduce how the items *had* been traditionally (and "incorrectly") classified. Further, using the refinement operators for hierarchy restructuring which we mention in Section 4.8, ReTAX is able to propose changes to the hierarchy which challenge the user's classification.[3]
- The *meta-knowledge* currently incorporated in ReTAX is restricted to general principles for structural requirements which must be met among taxa (see "consistency rules" in Section 4.7). We suggest that these criteria are of general applicability, and, in fact, they have proved useful for a historical simulation of revision (Section 5). However, they do not incorporate the sort of domain knowledge which we highlighted in Section 3.

In summary, the main focus of ReTAX is *modifying* the taxonomic criteria (the descriptors) by which objects in a given taxonomy are either grouped together or differentiated from each other (see the end of Section 2). In discussions we have had with working

[3] We believe it would be theoretically possible to simulate an unsupervised revision of a hierarchy by using ReTAX systematically to explore all possible taxa in the hierarchy.

taxonomists, they have stressed that the search for alternative descriptors is a crucial (and demanding) taxonomic activity.

4.2. General description of the system

ReTAX receives as *input*:

- A taxonomic *knowledge base*, which consists of a botanical hierarchy and a set of "floating taxonomic descriptors" (see representational issues in Section 4.3).
- Some *membership recognition mechanisms*, which recognise a taxon as a child of another taxon (see Section 4.6).
- A set of *consistency rules*, which detect inconsistencies in a taxonomic hierarchy (Section 4.7).
- A set of *refinement operators* which generate refinements to a hierarchy in an attempt to resolve detected inconsistencies (Section 4.8).

The *output* of ReTAX is ideally an updated hierarchy which is consistent with the information associated with the new specimens.

4.3. Representation of botanical knowledge

A frame-based formalism has been used to represent most of the botanical knowledge implemented in ReTAX. The knowledge base essentially consists of a *hierarchy of frames*; additionally, as noted above, it contains a set of "floating taxonomic descriptors". Each frame in the hierarchy corresponds either to a botanical taxon or to a particular botanical specimen. Specimens appear as leaf nodes of the hierarchy. On the other hand, the "floating taxonomic descriptors" are features which have not been included in the hierarchy but are, nevertheless, potentially relevant for taxonomic purposes.

Fig. 1 shows the different pieces of information which constitute a frame. Fig. 1(a) shows the *slots* contained in a frame. Fig. 1(b) includes the *aspects* which describe each *feature* (or character) in the following slots: GENERAL FEATURES, FLOWER, FRUIT, LEAVES, and STEM in Fig. 1(a). Most of the elements presented in the figure are self explanatory and fairly standard. However, some deserve further explanation.

The *is-a* slot contains the name of the *parent* of the current frame, that is, the taxon at the immediately higher level of the hierarchy. Using a fairly standard terminology in AI, the current frame is said to be the *child* of the taxon contained in the *is-a* slot. All other higher taxa in which a taxon is embedded are known as the *ancestors* of the taxon. Similarly, taxa which, at the same level of the hierarchy, are members (children) of the same parent are considered to be *siblings* of one another.

A taxon is described by two different types of features: (a) features which apply to a plant as a whole (slot GENERAL FEATURES on Fig. 1(a)); and (b) features associated with specific structures of the plant (slots FLOWER, FRUIT, LEAVES, and STEM in Fig. 1(a)). These four plant structures are the ones usually considered to describe the members of the Ericaceae family (i.e., the botanical domain to which ReTAX has been applied).

a)
Frame *name*

:IS-A	*name of the* parent
:RANK	*("family"* or *"genus"* or *"species")*
:ABSTRACTION	*("taxon"* or *"instance")*
:CHILDREN	*list of* children
:INSTANCES	*list of specimens belonging to a taxon*
:GENERAL	
	feature-1; feature-2; feature-3;... feature-n
:FLOWER	
	flower-feature-1; flower-feature-2; flower-feature-3;...flower-feature-n
	:d.r. index for slot *flower*
:FRUIT	
	fruit-feature-1; fruit-feature-2; fruit-feature-3;... fruit-feature-n
	:d.r. index for slot *fruit*
:LEAVES	
	leaves-feature-1; leaves-feature-2; leaves-feature-3;... leaves-feature-n
	:d.r. index for slot *leaves*
:STEM	
	stem-feature-1; stem-feature-2; stem-feature-3;... stem-feature-n
	:d.r. index for slot *stem*

b)
Feature

:name	`e.g.: size of FLOWER`
:value or set of values	`e.g.: (big medium)`
:d.r. index	`0-1`

Fig. 1. Information associated with a frame.

The aspect *value* of a feature is the only type of information that a frame can inherit from frames at higher levels of the hierarchy. Note when a feature has been assigned a value, we will say that the feature has been *instantiated.* Hence, in this paper, an *instantiated feature* is equivalent to what taxonomists refer to as a *character state.*

The aspect *discriminatory relevance index* (henceforth, *d.r. index*) consists of a numerical value which indicates the taxonomic significance or discriminatory power of a botanical feature (see Fig. 1(b)) or plant structure (e.g., FLOWER; see Fig. 1(a)). Those features which are useful to discriminate the frame from other taxa at the same level of the hierarchy have an index above a given threshold. For a given taxon, features whose index is above the threshold will be referred to as *dominant* features. A dominant feature which discriminates between two particular taxa at the same level of the hierarchy will be referred to as a *distinguishing* (or diagnostic) feature for that pair of taxa. Features with a *d.r. index* below the threshold will be referred to as *subsidiary* features. Additionally to differentiating between dominant and *subsidiary* features, the numerical values of the *d.r. indices* are used to determine the order in which features are considered ("activated") by some of ReTAX's procedures (see Section 4.8). In general, features with higher *d.r. indices* are activated first.

See [2] for a more detailed explanation of representational issues in ReTAX.

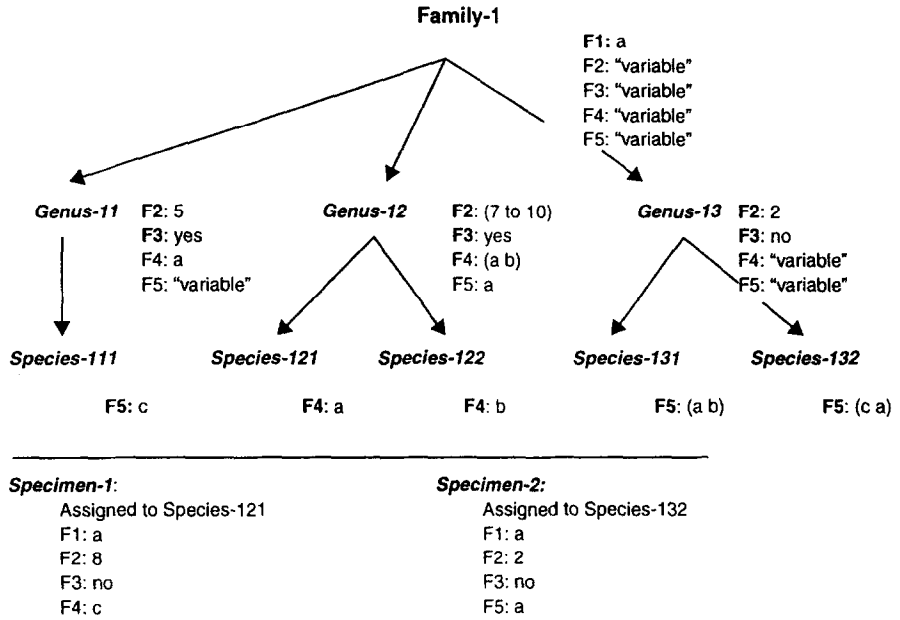

Fig. 2. Hypothetical hierarchy and specimens.

4.4. A hypothetical taxonomy

We will illustrate the different mechanisms implemented in ReTAX with examples drawn from the hypothetical hierarchy given in Fig. 2. This hierarchy consists of nine taxa, namely: one family (Family-1), three genera (Genus-11, Genus-12, and Genus-13), and five species (Species-111, Species-121 Species-122, Species-131, and Species-132). These taxa are characterised by five descriptors (F1, F2, F3, F4 and F5). The descriptors which are considered *dominant* features for each taxon appear in bold type in the figure. So, for example, F1 is the only *dominant* feature for taxon Family-1. Additionally, the figure includes two hypothetical specimens, Specimen-1 and Specimen-2; the values for their characters are also given; further each specimen is assigned to one of the hypothetical species.

4.5. Top-level algorithm

The performance of ReTAX consists essentially of three interrelated phases which are summarised in Fig. 3. Briefly, these phases are:

(a) An *interactive* phase, in which the system requests, from the user, information associated with the new specimen. Firstly, ReTAX asks the user for the name of the species the item belongs to (step 1.1 in Fig. 3(a)); this is equivalent to enquiring about how the item had been classified in the past. Subsequently, ReTAX requests information about different relevant taxonomic features (step 1.2 in Fig. 3(a)); this is equivalent to "observing" the plant.

a) **ReTAX (*new item*):**

1. User querying phase:
   1.1 Ask the user to assign the new item to a species
   1.2 Query the user about the values possessed by the item for a series of relevant features
2. Consistency Checking Phase: Check if any of the CRs is violated
   if some of the CRs is violated
   apply REFINEMENT CYCLE to solve detected inconsistencies
   else extend KB by adding information about new item
3. ***when*** the species reported by the user is not represented in KB
   search for distinguishing features between the new species and its *siblings*
   → do the same with the genus and the family if they are not represented in KB

b) **REFINEMENT CYCLE (*inconsistencies*):**

1. Apply refinement operator to deal with the least "critical" inconsistency
2. Apply Consistency Checking
   if there is any inconsistency
   apply REFINEMENT CYCLE to solve remaining inconsistencies
   else extend KB by adding information about new item.

Notation:
KB = Knowledge Base
CRs = Consistency Rules

Fig. 3. Top-level algorithm implemented in ReTAX.

(b) A *consistency checking* phase. In this cycle, the system essentially compares the information associated with the new item with the information stored in the knowledge base. When making this comparison, ReTAX is guided by a set of consistency criteria (see Section 4.7). The system detects an inconsistency if, as a result of incorporating new information, any of the consistency rules are violated.

(c) A *refinement cycle* which is activated if any inconsistency has been detected. After a refinement operator has been applied and an inconsistency is resolved, ReTAX executes a new consistency checking; this is needed because the resolution of one inconsistency may, on the one hand, have the side effect of solving other "waiting" inconsistencies, or, on the other hand, may lead to the appearance of further inconsistencies.

When ReTAX either does not detect any inconsistency in the knowledge base or has dealt with all the ones it has encountered, the new specimen is added at the appropriate place in the hierarchy.

4.6. Membership recognition procedures

ReTAX uses a series of mechanisms to determine whether an entity (be it a specimen or a taxon) is a member of a taxon. In general, such processes could be quite complex,

CR.1.- Each *instantiated* feature of a taxon must *either* be:

a. the same as the corresponding feature of its *parent*

b. *or* a specialisation of the corresponding feature of its *parent*

CR.2.- Each taxon will be distinguished from another sibling either by:

a. having a distinct set of values for at least one of the dominant features they share

b. *or* being the exclusive parent of its children

Fig. 4. General "consistency rules".

but we have reduced their complexity by only handling, in this version of ReTAX, *monothetic* taxa (see Section 4.1).

If the values for every *dominant* feature of a taxon match the values for the corresponding features in the entity, the entity is said to be a member of the taxon. A set of values (vs_1) for one feature is said to match another value or set of values (vs_2) for the same feature if vs_1 is identical with or contains vs_2. For example, the value for F1 in Specimen-1 (i.e., "a") is matched by the value for the same feature in Family-1 (see Fig. 2). Similarly, the value "8" for F2 of Specimen-1 matches the interval "(7 to 10)" for F2 in Genus-12. Similarly again, Specimen-2 would be recognised as a member of Genus-13 because the values it possesses for F2 and F3 (*dominant* features for the genus) are matched by the corresponding values possessed by the taxon. On the other hand, Specimen-1, would *not* be recognised as a member of Genus-12, as one *dominant* feature of Genus-12 (i.e., F3) is not matched by the corresponding value possessed by the specimen.

Finally, a feature is treated as a *distinguishing* feature between two taxa at the same level of the hierarchy if: (a) the feature is *dominant* in both taxa; and (b) its instantiation in one taxon does not overlap with its instantiation in the other taxon. A value or set of values (vs_1) for one feature overlaps with another value or set of values (vs_2) for the same feature if, at least, one of the values in vs_2 is contained in vs_1. For example, in Fig. 2 the values for feature F5 in Species-131 overlap with the values for the same feature in Species-132. Hence feature F5 is not a *distinguishing* feature between Species-131 and Species-132. On the other hand, F2 is a *distinguishing* feature for Genus-11 and Genus-13 because their corresponding instantiations do not overlap.

4.7. Consistency checking

A set of "consistency rules" has been defined for use in this domain; specifically, it is used when ReTAX is checking the inconsistencies which may appear in a hierarchy as a result of incorporating a new item. Fig. 4 shows the two types of rules considered. Rule CR.1 is concerned with the inconsistencies which may occur "vertically" in the hierarchy; that is, it specifies the requirements which must be met between one taxon and its *children*. On the other hand, rule CR.2 is concerned with "horizontal" inconsistencies;

that is, it specifies the relationships which must exist among taxa at the *same* level of the hierarchy.

CR.1 is basically concerned with the criteria used to determine whether an entity belongs to a given taxon. In ReTAX, every member of a taxon must possess for *each* of the *dominant* features of the taxon a value or set of values which is matched by the corresponding instantiation in the taxon. If, at least, one of the instantiated *dominant* features of a taxon does not match the corresponding instantiated feature of one of its members, CR.1 is said to have been violated. For example, Specimen-1 in Fig. 2 would *not* be recognised as a member of Species-121 because the item possesses the value "c" for feature F4 (the *dominant* feature for Species-121) and the corresponding value for the species is "a".

CR.2a establishes that siblings must be distinguished from each other in terms of their instantiated features. According to CR.2a, ReTAX detects an inconsistency in a hierarchy if there is not at least one feature which discriminates between the two taxa. For example, in Fig. 2 there is no *distinguishing* character between Species-131 and Species-132, because F5, the *dominant* feature which the two species share, does not discriminate between them. In fact, the respective instantiations of F5 in the two species (i.e., "(a b)" and "(c a)") overlap.

CR.2b specifies that taxa at the same level of the hierarchy cannot share a child (or an instance). ReTAX determines an inconsistency with respect to this constraint when a specimen can belong simultaneously to two taxa at the same level of the hierarchy. For example, Specimen-2 in Fig. 2 (which has been hypothetically assigned to Species-132) can be recognised as a member of either Species-131 or Species-132. This is because, as we saw, descriptor F5 is *not* a *distinguishing* feature between the taxa.[4]

These constraints have been procedurally implemented in ReTAX in a series of "consistency checkings" by which the system compares the new item with the information stored in the taxonomy. ReTAX creates three "inconsistency lists"; one for each of the three types of inconsistencies described above, namely, "CR.1-list", "CR.2a-list", and "CR.2.b-list".

4.8. *Refinement procedures*

The application of the refinement operators is guided by the type of inconsistencies the system is dealing with. For every type of inconsistency there are, at least, one and possibly more operators which are designed to cope with it.

[4] Naturally, the violations of CR.2a and CR.2b are closely interrelated. If the system detects that a specimen can be recognised as a member of two different sibling taxa (violation of CR.2b), it means that there is no *distinguishing* feature among those siblings (violation of CR.2a). For example, Specimen-2 in Fig. 2 (which has been hypothetically assigned to Species-132) can be indiscriminately recognised as a member of either Species-131 or Species-132. This is so because descriptor F5, although represented as *dominant* in both species, is not really a *distinguishing* feature for the taxa. However if the value possessed by Specimen-2 for F5 were "c" instead of "a", the item would only be considered a member of Species-132. The system would not detect an inconsistency in terms of CR.2b, but the violation of CR.2a would still exist. Consequently, a different consistency "checking" is applied for each of the rules.

FEATURE UPDATING OPERATORS

RO.1 - Generalise Descriptor

INPUT: A taxon's *dominant* descriptor whose instantiation does *not* match the instantiation of the corresponding descriptor in an item embedded in the taxon.

ACTION: Enlarge the set of values for the descriptor in the taxon so that it matches the instantiation of the corresponding feature in the item.

RO.2 - Decrease *d.r. Index* of Descriptor

INPUT: A descriptor which is inconsistently considered as a *dominant* feature for a taxon.

ACTION: Decrease the *d.r. index* of the descriptor so that the descriptor is no longer considered a *dominant* feature.

RO.3 - Search for Disjunctive Descriptor

INPUT: A set of items which cannot be recognised as members of a taxon in which they are embedded because of their "inconsistent" instantiation of one of the taxon's *dominant* descriptors.

ACTION: Search for a new feature which discriminates between that subset of the members of the taxon and the members of the taxon's *siblings*.

RO.4 - Search for Distinguishing Feature

INPUT: Two taxa which are not distinguished by any *dominant* feature.

ACTION: Search, among the subsidiary features of the taxa and the *floating taxonomic descriptors*, for a feature which discriminates between the members of one of the taxa from the members of the other.

RO.5 - Increase *d.r. Index* of Descriptor

INPUT: An originally *subsidiary* feature which has been identified as a *distinguishing* descriptor between two taxa.

ACTION: Increase the *d.r. index* of the feature so that the descriptor is considered a *dominant* feature

RO.6 - Specialise Descriptor

INPUT: An instantiated *dominant* feature of a taxon *matches* the corresponding feature in a specimen which is embedded in a *sibling* of the taxon.

ACTION: Remove from the description of the feature those values possessed by the specimen for the corresponding feature.

HIERARCHY RE-STRUCTURING OPERATORS

RO.7 - Merge Taxa

INPUT: Two taxa for which no *distinguishing* feature has been found.

ACTION: Create a new taxon which includes the members of the two conflicting taxa.

RO.8 - Elevate rank

INPUT: An item which is very distinctive; that is, it is very different from the members of its assigned *parent* and from the members of the siblings of its *parent*.

ACTION: Create a new *parent* taxon which embeds the item as a unique member.

Fig. 5. Set of "refinement operators" implemented in ReTAX.

The "refinement cycle" is executed under the *assumption* that simpler refinements should be dealt with before the more demanding ones. Hence, ReTAX starts dealing with the least "critical" of the inconsistencies it has detected (see Fig. 3(b)). The three different types of inconsistencies have been ordered, somehow arbitrarily, in terms of

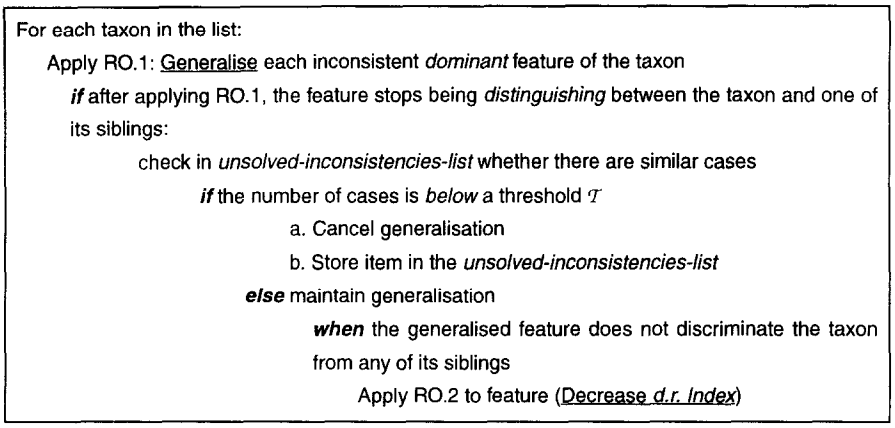

Fig. 6. Procedures to deal with inconsistencies in CR.1-list.

a "criticality" criterion (explained in detail in [2]). Briefly, the order in which the different inconsistencies are dealt with is the following: CR.1, CR.2b, CR.2a.

ReTAX uses eight refinement operators whose names and specifications are listed in Fig. 5. Each operator is described in terms of the elements it acts on and the action performed. Two kinds of refinement operators are distinguished: operators which affect the representation of features in the elements of a hierarchy (i.e., *feature updating operators*); and operators which alter the structural relationships among elements of the hierarchy (i.e., *hierarchy restructuring operators*).

For the sake of brevity, the following description will omit three of the refinement operators outlined in Fig. 5, namely, RO.3 ("Search for Disjunctive Descriptor"), RO.6 ("Specialise Descriptor"), and RO.8 ("Elevate Rank"). Although these three operators have been implemented in the system, they were not needed during the historical simulation with which ReTAX was evaluated (Section 5). (A specification of these operators can be found in [2]).

Inconsistencies associated with CR.1

Fig. 6 summarises the main procedures followed by ReTAX to deal with "CR.1-list". The main operator used to deal with this type of inconsistency is RO.1, "Generalise Descriptor". When using RO.1, the system generalises those instantiated *dominant* features which, in a given taxon, do not match the corresponding instantiated features in the specimen. For example, we saw that in Fig. 2 there was an inconsistency between Specimen-1 and Species-121 as CR.1 was violated. In this case, ReTAX generalises feature F4 in Species-121; after the application of RO.1, the description of F4 in Species-121 is "(a c)".

However, before applying RO.1, the system checks whether the generalised feature would still discriminate between the taxon and all its siblings. For example, if F3 is generalised, the feature does no longer discriminate between Genus-12 and Genus-13. In cases like this, ReTAX may cancel the generalisation of the descriptor and store the case as an unsolved inconsistency in a list which will be referred to as the

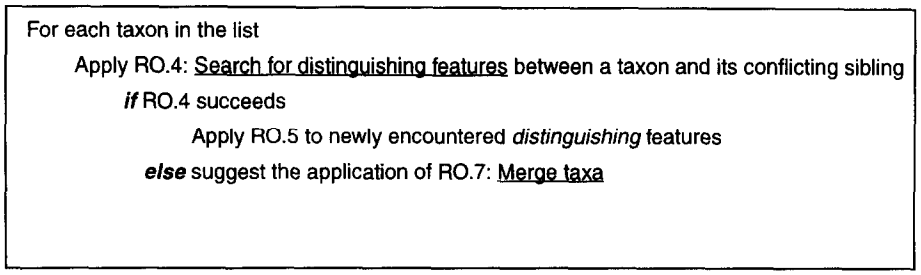

Fig. 7. Procedures to deal with inconsistencies in CR.2a-list or CR.2b-list.

unsolved-inconsistencies-list. A single specimen is not deemed to be sufficient evidence to generate an alteration in the representation of a feature which may ultimately lead to a more serious inconsistency, i.e., the violation of rule CR.2.

However, before ReTAX actually cancels the generalisation of a feature, ReTAX inspects the *unsolved-inconsistencies-list* to see if there have been previous cases similar to the one it is dealing with. The subsequent process can be summarised as follows:

- If the number of *similar* unsolved inconsistencies is above a given threshold, the system decides to give priority to the inconsistency associated with CR.1 over an eventual inconsistency with respect to CR.2a. In this case, ReTAX retains the generalisation of the feature. Further, the system checks whether this generalised feature is useful to discriminate the modified taxon from any of its other siblings. If it is not useful as a discriminant feature any more, ReTAX applies RO.2, that is, it decreases the *d.r. index* of the descriptor so that it is no longer considered a *dominant* feature. Subsequently, as the inconsistency has been solved, ReTAX removes from the *unsolved-inconsistencies-list* those items which have been used to generalise the feature.
- If ReTAX does not encounter in the *unsolved-inconsistencies-list* a sufficient number of similar unsolved inconsistencies, the system cancels the application of RO.1. Subsequently, ReTAX stores the unsolved inconsistency in the *unsolved-inconsistencies-list*.

Inconsistencies associated with CR.2a or CR.2b

Fig. 7 summarises the procedures used by ReTAX to deal with these inconsistencies. The main operator used to deal with them is RO.4 "Search for Distinguishing Features". As we saw above, an inconsistency associated with CR.2 implies that two taxa cannot be discriminated from each other on the basis of the *dominant* features that they share. The purpose of the operator RO.4 is to determine whether any of the features which are represented as *subsidiary* for one taxon may actually be used as *distinguishing* features between that taxon and a sibling. The application of the operator consists of comparing the members of a taxon with the members of the sibling. This comparison is done in terms of the values possessed by the members of each taxon for the following two types of features: (1) those features which, being *dominant* for one of the taxa, are *subsidiary* for the other one; and (2) those features which are represented as *subsidiary* in both taxa.

If, for a given feature, all the members of one taxon are represented with a set of values which does not overlap with the set of values which represents the members of the sibling taxon, then ReTAX decides that a new *distinguishing* feature has been encountered. As soon as ReTAX encounters such a feature, the system stops the search and updates the representation of the feature in the two taxa. The feature's *d.r. index* is increased so that it is considered a *dominant* feature (i.e., RO.5 is applied).[5]

As suggested in Section 4.3, ReTAX considers first those *subsidiary* features whose *d.r. indices* are the highest. In other words, those features which, being *subsidiary*, have the highest *d.r. indices* will be the likeliest to become *dominant* features. When the features with the highest *d.r. indices* prove to be inappropriate to discriminate between two taxa, the system considers the features with the next highest *d.r. indices*, and so forth until a feature is found to be discriminatory, or the set is exhausted.

If ReTAX does not encounter a descriptor among the *subsidiary* features which is able to discriminate between two taxa, it compares the members of the two groups in terms of the values they possess for the floating taxonomic descriptors. If, again, none of the floating taxonomic descriptors is found to be useful to discriminate between the two taxa, ReTAX decides that the groups are not sufficiently distinct, and proposes the application of RO.7 (i.e., a structural refinement), which merges the two taxa into a single taxon.

A more detailed explanation of ReTAX's refinement procedures can be found in [2].

5. Evaluation of ReTAX

This section reports the performance of ReTAX as it sought to revise the botanical genera *Pernettya* and *Gaultheria* in family Ericaceae. The purpose of the evaluation of the system was to determine whether ReTAX was able to mimic the historical evolution of the concepts of those genera as reported in [19].

5.1. A historical example: revision of the concepts of genera Pernettya *and* Gaultheria *in family Ericaceae*

One of the earliest and most influential studies of the family Ericaceae was made in the last century by Bentham and Hooker [5]. In Bentham and Hooker's classification, *Pernettya* and *Gaultheria* appear as two separate genera. The main differences detected between the two genera were the "type of fruit" and the "succulence of the calyx" that surrounds the fruit. The fruit in *Gaultheria* is normally a capsule surrounded by a succulent calyx while in *Pernettya* the fruit is a fleshy berry with a dry (non-succulent) calyx.

Subsequent studies of the genera throughout this century showed that the two features highlighted by Bentham and Hooker were not appropriate to discriminate between *Pernettya* and *Gaultheria*. Taxonomists attempted to determine whether alternative descrip-

[5] This search for discriminating features among *subsidiary* descriptors replicates some of the procedures utilised by the botanists in our psychological study as simulated in Proto-ReTAX [3].

Botanical Findings inconsistent with current taxonomy	Violated CR	Historical Taxonomic Decisions	RO
1. Species of *Pernettya* have *fleshy calyces* AND Species of *Gaultheria* have *dry calyces*	CR.1	Character *calyx succulence* is deemed to be inappropriate to discriminate between *Pernettya* and *Gaultheria*.	RO.1 & RO.2
2. Species of *Pernettya* have *capsules* AND Species of *Gaultheria* have *berries*	CR.1	Character *fruit type* is deemed to be inappropriate to discriminate between *Pernettya* and *Gaultheria*	RO.1 & RO.2
3. *Fruit* characters do not discriminate between the genera.	CR.2a	Examination of further botanical characters: - leaf anatomy - presence of dioecism & vivipary	RO.4
4. Characters *presence of vivipary* and *presence of dioecism* are found to be useful to discriminate between the two genera.	--	--	RO.5
5. Further analyses of the genera show that no character discriminates between Pernettya and Gaultheria	CR.2a	Merge members of Pernettya and Gaultheria into a unique group (Middleton & Wilcock, 1990).	RO.7

Fig. 8. Summary of the revision of genera *Pernettya* and *Gaultheria*.

tors were useful to maintain the distinction between the genera; but further exhaustive examinations of the genera finally suggested that there are not sufficient differences between *Pernettya* and *Gaultheria* to maintain them as separate groups. In a recent revision of the genera, Middleton and Wilcock [19] reclassified all the species of *Pernettya* as members of *Gaultheria* (i.e., the two genera were merged).

This historical process is summarised in the first and third columns of Fig. 8. Additionally, Fig. 8 shows the consistency rules and refinement operators that ReTAX used to simulate each of the historical stages in the revision process (see second and fourth columns of the figure).

5.2. Simulation of historical data

The knowledge base used with ReTAX corresponds to Bentham and Hooker's [5] original classification of the family Ericaceae. Specifically, the knowledge base contains a subset of the taxa in Bentham and Hooker's taxonomy. In the knowledge base, taxa are represented at the levels of family, genus and species. Additionally to *Pernettya* and *Gaultheria*, the knowledge base represents five other genera. Fig. 9 gives a graphical representation of the groups included in the input hierarchy.

Bentham and Hooker's "Genera Plantarum" [5] and Middleton and Wilcock's revision [19] were the principal sources of the information used for the taxa in the hierarchy. This

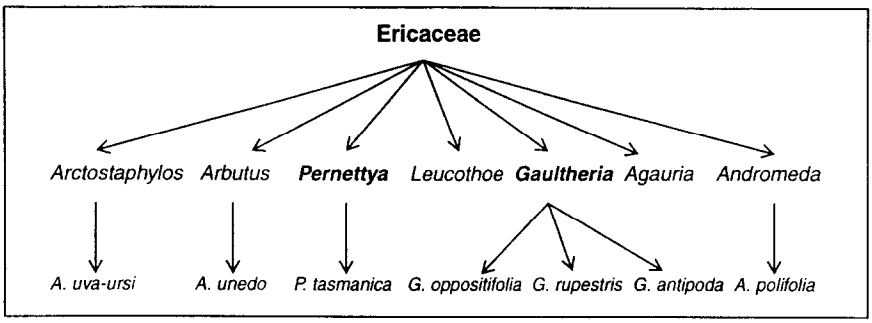

Fig. 9. Taxonomic hierarchy implemented in ReTAX to simulate the historical revision of the concepts of genera *Pernettya* and *Gaultheria* in the family Ericaceae.

information was complemented with the descriptions encountered in modern floras (e.g., [27]). When contradictions existed among the different taxonomic sources, priority was given to older texts, in particular to Bentham and Hooker's [5] descriptions, since the knowledge base was meant to reproduce Bentham and Hooker's classification. For example, the two features which were represented as *distinguishing* between *Pernettya* and *Gaultheria* were "calyx in fruit" and "fruit type".

As well as the hierarchical taxonomy, ReTAX was provided with a set of floating taxonomic descriptors. This set contains botanical characters which had not been considered, or given low priority, in early classifications of the Ericaceae. These floating taxonomic descriptors include:

- the general feature "presence of dioecism";
- the character "adherence", associated with the *fruit*;
- and characters related to leaf anatomy: "pith-type", "stomata-type", "hypodermis", "free-fibres", "lamina-thickness", "palisade-layers".

In total, 81 botanical characters have been used in the historical simulation, including both the features implemented in the hierarchy (i.e., 73 features) and the eight floating taxonomic descriptors.

Fifteen items were used in the evaluation of ReTAX. The items represent specimens which had been classified in early taxonomies (i.e., Bentham and Hooker's [5], and those based on it) as belonging to species of either *Pernettya* or *Gaultheria*. Fig. 10 shows the names of the species in which the specimens had been originally classified; additionally, it provides the order in which those items were presented to ReTAX during the simulation. The order of item presentation reflected, as accurately as possible, the order in which the items had been investigated historically by taxonomists.

Given the knowledge base and the fifteen items just outlined, ReTAX was able to replicate some of the taxonomic decisions which led to the change in the status of *Pernettya* and *Gaultheria*. ReTAX used the consistency rules (especially CR1 and CR2a) to recognise inconsistencies in the original classification given novel observations. The assessments provided by the consistency rules, combined with the refinements executed to accommodate new evidence (see used refinement operators in Fig. 8), led ReTAX to modifications similar to those proposed by modern taxonomists.

Items "inconsistent" with respect to feature Calyx-succulence of Fruit:
Pernettya nana
Gaultheria nubigena
Gaultheria itatiaiae
Pernettya macrostigma
Pernettya insana
Items "inconsistent" with respect to feature Type of Fruit:
Gaultheria procumbens
Pernettya lanceolata
Gaultheria rigida
Gaultheria glomerata
Gaultheria sinensis
Items "inconsistent" with respect to General feature Presence of Dioecism:
Pernettya prostrata
Gaultheria rupestris
Gaultheria oppositifolia
Gaultheria tenuifolia
Gaultheria wardii

Fig. 10. Set of items presented to ReTAX during the historical simulation.

Name the SPECIES of the plant you want to classify: ***pernettya nana***

What is the value of ANTHER-AWNS in the FLOWER of this plant? *short*
What is the value of COROLLA-LOBES-NUMBER in the FLOWER of this plant? *5*
What is the value of TYPE in the FRUIT of this plant? *berry*
Has this plant STAMEN-APPENDAGES in the FLOWER? *no*
What is the value of -MEROUS in the FLOWER of this plant? *5*
(...)
Consistency Checking
CR.1-LIST
(PERNETTYA (FRUIT (CALYX)))
CR.2a-LIST
NIL
CR.2b-LIST
NIL
*** Refinement Cycle ***
(< RO.1-GEN >)
Trying to generalize...
Failed generalization of CALYX of FRUIT in PERNETTYA
Consistency Checking
[Not new inconsistency found]
NO-REFINEMENT

Updating the Hierarchy with New Item...
What is the value of ANTHER-ARISTS in the FLOWER of this plant? *2*
What is the value of FILAMENTS-SHAPE in the FLOWER of this plant? *linear*
What is the value of APEX in the LEAVES of this plant? *obtuse*
Has this plant PETIOLE in the LEAVES? *yes*

The distinguishing features between PERNETTYA-NANA and PERNETTYA-TASMANICA are:
(INFLORESCENCE ANTHER-ARISTS) of FLOWER
(CALYX-SUCCULENCE) of FRUIT

Fig. 11. Excerpt of the output produced by ReTAX for the item *Pernettya nana*.

Briefly, when ReTAX detected that neither of the fruit-related characters was appropriate to discriminate between *Pernettya* and *Gaultheria*, the system searched for alternative *distinguishing* features (operator RO.4). The feature "presence of dioecism" was encountered as a new *distinguishing* descriptor. However, new evidence ruled out this feature as a discriminant one. A new search for further *distinguishing* features failed, thus ReTAX proposed a structural refinement, that is, merging the two genera (operator RO.7; see Fig. 8).

A sample of the output produced by the system is shown in Fig. 11. This figure summarises ReTAX's output for the first item used in the simulation, i.e., *Pernettya nana*. In this figure, small parts of the trace have been omitted for brevity. All the text which appears in the figure, except those words in italics, corresponds to the actual output of the system. The phrases in italics which appear in square brackets are explanatory notes of certain aspects of the output. The rest of the text in italics corresponds to the replies of the user in the interactive phases.

Although the performance of ReTAX succeeded in simulating the taxonomic evolution of the concept of genus *Pernettya*, there are certain aspects of the historical process which have not been completely reproduced. For example, taxonomists considered, in their investigations of the genera, a *wider range* of specimens and botanical characters than the ones used in the simulation. Additionally, taxonomists studied the classification of *Pernettya* and *Gaultheria* in the context of a larger number of taxa than the ones implemented in the knowledge base used by ReTAX. Nevertheless, the taxa and features included in the knowledge base are deemed to be sufficiently representative of the diversity which exists in the Ericaceae family and, hence, reflect the complexity involved in the revision of a relatively large botanical taxonomy.

A more detailed account of ReTAX's simulation of historical data is given in [2].

6. Related work

ReTAX combines general features of *theory revision systems* (Section 6.1) with some of the characteristics of *conceptual clustering methods* (Section 6.2).

6.1. Theory revision and knowledge refinement

Revision has been addressed by various modern discovery models. Systems such as PHINEAS [10], HYPGENE [14], and several others (see [6]), deal with the revision of scientific theories or models. However, they do not address the revision of taxonomies.

ReTAX, as other theory revision systems, revises a knowledge base which cannot account for inconsistent data. Similarly to these systems, the task of ReTAX is to search for new explanations of the data, and to modify the original knowledge base accordingly. However, in contrast with earlier models, ReTAX is concerned with revision in the context of *classification*. Consequently, the input knowledge base is a taxonomic hierarchy, and the search for new explanations of the data is performed by either identifying novel descriptors to characterise taxa or by structurally reorganising the hierarchy.

On the other hand, knowledge base refinement is the subfield which deals with modifying the knowledge bases used with expert system shells [7]. Many of ReTAX's refinement operators are similar to the operators used by knowledge refinement systems. ReTAX adapts the standard operators used by knowledge refinement systems to deal with a frame-based hierarchical representation. For example, the generalisation and specialisation of rule conditions in standard refinement systems are represented in ReTAX by operators RO.1, "Generalise Descriptor", and RO.6, "Specialise Descriptor" (see Section 3.4.1).

6.2. *Systems for taxonomy formation*

The task executed by ReTAX is, to a great extent, related to *conceptual clustering*. Conceptual clustering was developed as a symbolic AI extension to numerical taxonomy [12]. Whereas *numerical clustering* methods group items on the basis of purely numerical similarity measures [9], conceptual clustering systems generate not only a hierarchical organisation of objects but also a conceptual description which characterises each group in the hierarchy. In this context, a *conceptual* description (a concept) can be interpreted as a *rule* which makes explicit the conditions under which a set of objects are considered members of a given cluster [12]. However, although these systems were originally developed to overcome limitations of numerical clustering, to our knowledge, conceptual clustering methods have not been applied to biological taxonomic domains. Numerical classification techniques, in contrast, are still used by modern biologists as a significant tool for taxonomy formation [21]. As suggested in Section 1, the purpose of ReTAX (or a subsequent extension) is to apply AI techniques to real biological tasks.

The conceptual clustering methods which are most closely related to ReTAX are those which are *incremental*, that is, methods which construct classifications from partial data, and extend them as new information becomes available. Relevant examples of incremental conceptual clustering are COBWEB [11] and ARACHNE [18]. Similarly to these methods, ReTAX conceptually characterises groups of entities in terms of shared descriptors. Further, ReTAX uses a restructuring operator (RO.7, *Merge taxa*) which is very similar to operators used by COBWEB and ARACHNE. In contrast with ReTAX, conceptual clustering methods (as well as numerical techniques) are unsupervised (i.e., they do not require that the user provides a classification for items), and generate polythetic taxa, as opposed to the monothetic groups considered by ReTAX. In the concluding section we suggest how our system can be extended to do unsupervised learning and deal with polythetic taxa.

An important difference between ReTAX and previous clustering methods is that our system focuses on revision. We argue that incremental conceptual clustering is unsatisfactory to deal with taxonomic revision. Systems, like COBWEB and ARACHNE, are biased towards updating a hierarchy every time a new item is presented. Their application of domain-independent restructuring operators can lead to significant structural modifications on the basis of *a single* instance. However, in real taxonomic practice, the modification of hierarchies does not normally take place until a considerable amount of evidence has been gathered and analysed (see Section 2). In fact, ReTAX takes

into account both the weight of prior evidence which supports established taxonomies, and the information associated with newly encountered evidence. Some of the refinement operators used by ReTAX are not applied until sufficient empirical evidence has been gathered to support the need for a refinement. By using this approach, the system manages to replicate, to some extent, the desirable compromise between preserving the stability of a classification and updating that classification to accommodate novel information.

Further, the performance of conceptual clustering systems is generally confined to the detection of surface similarities among objects, and rarely benefits from the use of background knowledge. In contrast, taxonomic revision is clearly guided by taxonomists' scientific expertise and deep knowledge of the objects being classified. Currently, ReTAX does *not* use background knowledge, but we have previously indicated (Section 3) how such knowledge could be acquired and incorporated in future extensions of the system (see also Section 7).

7. Concluding remarks

We have presented a computational model for taxonomic revision which takes into account evidence from different sources: computational, psychological, historical, as well as contributions from practising taxonomists. We have discussed how a subset of this model has been implemented in ReTAX, a prototype system for taxonomic revision.

The initial success of ReTAX's historical simulation is encouraging. It suggests the potential of the system as an aid to taxonomists as they perform the demanding and time consuming tasks of taxonomic revision. In fact, discussions with working taxonomists have indicated a number of potential uses of ReTAX, including:

- The system could be used to determine which of two *alternative pre-existing classifications* is more consistent. The systematic application of ReTAX's current consistency rules (or an extension of these) should assist taxonomists in deciding between two conflicting arrangements of taxa.
- The system could help taxonomists *keep track* of all the characters which are being considered while constructing a taxon.
- The system could have interesting uses for curators, as an aid for the revision (and construction) of *identification keys*. This feature is particularly useful if we consider the large amount of descriptors and specimens which are being examined when constructing and revising identification keys. However, many computational systems already exist which deal with taxonomic identification (e.g., [8, 16, 20, 28]). Although ReTAX has not been designed to address identification, it may be interesting to explore how the system could assist in the revision of keys.
- An additional application of ReTAX could be its use as a *student modelling* system [24].

We reconsider the assumptions and simplifications which guided ReTAX's design (see Section 4.1) and propose a series of extensions to the system which would allow it to execute revision in a more realistic way, thus further implementing the general framework discussed in Section 3.

- We are planning to test ReTAX (or an extension of the system) on *new unsolved problems*, as opposed to historical cases. The system could be used to deal with recently encountered data which challenge modern established taxonomies, or to revise modern taxonomies in the light of new biochemical or cytogenetic findings. We are currently exploring applications in the botanical and medical (bacteriological) domains.
- If we want ReTAX to deal with novel *non-classified* data, the system should be able to operate without external supervision, as opposed to the current partly supervised approach. In fact, ReTAX could be easily adapted to perform unsupervised classification. The system could use the membership recognition procedures (or an extension of these) to autonomously identify a new specimen as a member of one or more of hierarchy's taxa. Subsequently, the system would be able to perform the currently available consistency checking and refinement cycles (see also footnote 3).
- We saw that ReTAX's restriction to handle only *monothetic* taxa was not realistic. However, the system could be easily adapted to incorporate contributions from numerical and conceptual clustering which, as we saw, construct polythetic taxa. Additionally, ReTAX's current consistency criteria should be relaxed to accommodate polythetic taxa. [6]
- Finally, the consistency criteria could be enhanced with *background knowledge* obtained by using the knowledge acquisition procedures suggested in Section 3. This would certainly improve the system's general consistency checking and refinement abilities, as these would then be supported by theoretical (evolutionary or otherwise) explanations.

In this article, we have stressed the relevance of classification in science. In our view, if the field of scientific discovery wishes to create a general architecture, it is essential that the automatisation of taxonomic tasks be included as a component. We believe that the framework proposed here for taxonomic refinement is a useful contribution to this overall architecture.

Acknowledgements

We would like to thank the following experts for their advice on taxonomy: Gordon Smith (University of Aberdeen), Andy Whittington (National Museums of Scotland), T.H. Pennington (Department of Medical Microbiology, University of Aberdeen), David Middleton (Royal Botanic Gardens of Edinburgh), and Chris Wilcock (Department of Plant and Soil Science, University of Aberdeen). Further, the work described in this paper has benefitted from comments provided by a number of people among whom we would like to mention: Lindley Darden, Pat Langley, Jeff Shrager, Herbert Simon, and Raúl Valdés-Pérez. The first author was supported by a research studentship granted by the Spanish Government (Ministerio de Educación y Ciencia).

[6] As a result, rule CR.2a, for example, would need to be modified to express the following criterion: "two polythetic taxa will be distinguished from each other if the members of one taxon have more instantiated features in common with each other than with the members of the other taxon" (see Section 4.7).

References

[1] L.A. Abbot, F.A. Bisby and D.J. Rogers, *Taxonomic Analysis in Biology* (Columbia University Press, New York, 1985).

[2] E. Alberdi, Accommodating surprise in taxonomic tasks: a psychological and computational investigation, PhD Thesis, University of Aberdeen, Scotland (1996).

[3] E. Alberdi and D. Sleeman, Taxonomy revision as shift of representational focus, in: *Proceedings European Conference on Cognitive Science* (1995) 47–54.

[4] R. Allkin and F.A. Bisby, *Databases in Systematics* (Academic Press, London, 1984).

[5] G. Bentham and J.D. Hooker, *Genera Plantarum*, Vol. II, Part 2 (Reeve & Co., London, 1876).

[6] P.C.-H. Cheng, Approaches, models and issues in computational scientific discovery, in: M.T. Keane and K.J. Gilhooly, eds., *Advances in the Psychology of Thinking* (Harvester-Wheatsheaf, Hemel Hempstead, 1992) 203–236.

[7] S. Craw, D. Sleeman, R. Boswell and L. Carbonara, Is knowledge refinement different from theory revision?, in: S. Wrobel, ed., *Proceedings MLnet Workshop on Theory Revision and Restructuring* (1994) 33–35.

[8] M. Domingo, Evaluating the expert system approach to biological identification through application to Porifera, in: R.W.M. van Soest, T.M.G. van Kempen and J.C. Braekman, eds., *Sponges in Time and Space* (Balkema, Rotterdam, 1993).

[9] B. Everitt, *Cluster Analysis* (Gower, Aldershot, 2nd ed., 1986).

[10] B. Falkenheimer, A unified approach to explanation and theory formation, in: J. Shrager and P. Langley, eds., *Computational Models of Scientific Discovery and Theory Formation* (Morgan Kaufman, San Mateo, CA, 1990) 157–196.

[11] D. Fisher, Knowledge acquisition via incremental conceptual clustering, *Mach. Learn.* **2** (1987) 139–172.

[12] D. Fisher and P. Langley, Conceptual clustering and its relation to numerical taxonomy, in: W.A. Gale, ed., *Artificial Intelligence and Statistics* (Addison-Wesley, Reading, MA, 1986) 77–116.

[13] D.L. Hull, *Science as a Process* (University of Chicago Press, Chicago, IL, 1988).

[14] P.D. Karp, Design methods for scientific hypothesis formation and their application to molecular biology, *Mach. Learn.* **12** (1993) 89–116.

[15] P. Langley, Machine learning and concept formation, *Mach. Learn.* **2** (1987) 99–102.

[16] J. Lebbe, R. Vignes and S. Darmoni, Symbolic-numeric approach for biological knowledge representation: a medical example with creation of identification graphs, in: Diday, ed., *Data Analysis, Learning Symbolic and Numeric Knowledge* (Nova Science Publishers, 1989).

[17] E. Mayr, *The Growth of Biological Thought* (Belknap, Cambridge, MA, 1982).

[18] K.B. McKusick and P. Langley, Constraints in tree structure in concept formation, in: *Proceedings IJCAI-91*, Sydney (1991) 810–816.

[19] D.J. Middleton and C.C. Wilcock, A critical examination of the status of *Pernettya* as a genus distinct from *Gaultheria*, *Edinburgh J. Botany* **47** (1990) 291–301.

[20] R.J. Pankhurst, Botanical keys generated by computer, *Watsonia* **8** (1971) 357–368.

[21] R.J. Pankhurst, *Practical Taxonomic Computing* (Cambridge University Press, Cambridge, 1991).

[22] J. Shrager and P. Langley, Computational approaches to scientific discovery, in: J. Shrager and P. Langley, eds., *Computational Models of Scientific Discovery and Theory Formation* (Morgan Kaufman, San Mateo, CA, 1990) 1–25.

[23] G.G. Simpson, *Principles of Animal Taxonomy* (Oxford University Press, London, 1961).

[24] D.H. Sleeman, Inferring student models for intelligent computer-aided instruction, in: R.S. Michalski, J.G. Carbonell and T.M. Mitchell, eds., *Machine Learning: An Artificial Intelligence Approach* (Morgan Kaufmann, Palo Alto, CA, 1983) 483–510.

[25] R.R. Sokal, Classification: purposes, principles, progress, prospects, *Science* **185** (1974) 1115–1123.

[26] R.R. Sokal and P.H.A. Sneath, *Principles of Numerical Taxonomy* (Freeman, San Francisco, CA, 1963).

[27] C. Stace, *New Flora of the British Isles* (Cambridge University Press, Cambridge, 1991).

[28] J.B. Woolley and N.D. Stone, Application of artificial intelligence to systematics: Systex—a prototype expert system for species identification, *Syst. Zoology* **36** (1987) 248–267.