ELSEVIER

Expert Systems with Applications 30 (2006) 337–351

Expert Systems with Applications

www.elsevier.com/locate/eswa

XKey: A tool for the generation of identification keys

M. Delgado Calvo-Flores[a], W. Fajardo Contreras[a], E.L. Gibaja Galindo[b,*], R. Pérez-Pérez[a]

[a]*E.T.S. de Ingeniería Informática, Departamento de Ciencias de la Computación e Inteligencia Artificial,*
Universidad de Granada, C/Periodista Daniel Saucedo Aranda, 18071 Granada, Spain
[b]*Escuela Politécnica Superior, Departamento de Informática y Análisis Numérico, Universidad de Córdoba, Campus de Rabanales,*
Edificio Albert Einstein, 3a planta, 14071 Córdoba, Spain

Abstract

This paper presents the development of XKey, a tool for generating taxonomical identification keys by means of decision tree construction. The tool is based on an XML standard for the representation of general taxonomical information, which makes it ideal for different fields of application. The article analyses the problem by examining the adaptation of machine learning techniques to the sphere of biology so as to incorporate the viewpoints of biologists and computer science experts. It also analyses the effect of using various division criteria on a set of real data: the Gymnosperm plant groups present in the Iberian peninsula.

Keywords: Dichotomous keys; Decision trees; XML; Division criteria; Interactivity

1. Introduction

Identification keys are fundamental in the study of biodiversity. This term, which arises from concern for the environment, and the need for conservation and the sustainable use of natural resources, is defined as the variability of living organisms on earth and includes all organizational levels: from the simplest of individuals to ecosystems (even the entire planet). The first step in its study is to identify the types of organisms present in the biotype being analyzed (in the field, the researcher finds an organism, but in a database register, its identification is stored). This identification (or determination) consists in recognizing characters in a sample, so as to give it a name which was previously given to a similar organism.

Given the age of taxonomy as a science, the most used identification tool in the last 200 years is the printed key; a tool which must also be included in field guides and floras. In general, experts design their keys manually, without the use of any computer support tool. The development of a key is therefore time-consuming, and also extremely costly when an error is detected: the later the error is detected, the more it will cost to rectify it.

Since the computer tools developed to support this process have been suggested by specialist biologists, we find that considerable improvements could be made. This article presents XKey, a tool for generating identification keys, and is organized in the following way: Section 2 details the background of the problem, focusing on the aspects related to the representation of taxonomical knowledge and the tools for the generation of previously designed keys; Section 3 analyses the problem and describes the division criteria used by the tool; Section 4 examines the characteristics of XKey and its contribution to its field of application; Section 5 outlines the test procedure and the results obtained; and finally, Section 6 presents a series of conclusions and the references consulted to produce this work.

2. Background

2.1. Identification keys

An identification key is a tree-shaped structure which presents the user with a series of choices at each step. We can distinguish between:

- *Printed keys*. Most printed keys have a treelike structure (see Fig. 1). They are very powerful tools when dealing

[*] Corresponding author.
E-mail addresses: mdelgado@decsai.ugr.es (M. Delgado Calvo-Flores), egibaja@uco.es (W. Fajardo Contreras), egibaja@uco.es (E.L. Gibaja Galindo).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.07.034

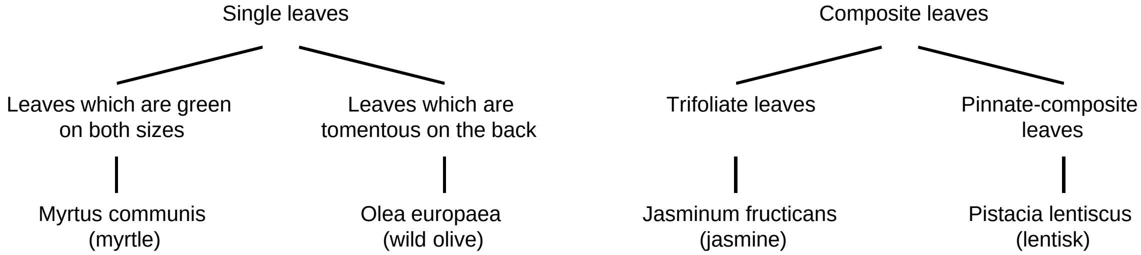

Single leaves
Composite leaves
Leaves which are green on both sizes
Leaves which are tomentous on the back
Trifoliate leaves
Pinnate-composite leaves
Myrtus communis (myrtle)
Olea europaea (wild olive)
Jasminum fructicans (jasmine)
Pistacia lentiscus (lentisk)

Fig. 1. Example of an identification key (Morales, Quesada, & Baena, 2001).

with primary identification characters (Fortuner, 1989), or rather characters which are capable of differentiating between species or groups of species with a very small risk of error. They proceed by elimination, and, for this reason, identification is prevented as they do not have the necessary information to go on to the next level. In order to avoid this, multi-entry or tabular keys are used whereby the elimination of species or groups depends on several characters. The problem of this type of key is that it is difficult to use with large genera.

- *Computerized keys*. This type of key is characterized by its interactivity: a computer enables a multi-entry key to be easily used. Within this group, we should also include the graphic key, in which the choices are presented in image format to represent the options.

It is not possible to say that one type of key is better than another; this aspect will depend on the sphere of its use. We shall focus our study on the development of tree-like printed keys, which are widely used and whose development is not particularly automated.

2.2. Digital representation of taxonomical knowledge

The automatic generation of identification keys requires highly structured descriptive information to be used. In spite of having described certain models such as Lif (UBio, 2003), Delta (Dallwitz, 1974; Dallwitz, Paine, & Zurcher, 2000) and Nexus (Maddison, Swofford, & Maddison, 1997), there is no widely accepted and used standard model, and this prevents a greater use of digital taxonomical information and leads to substantial inefficiency for taxonomy as a whole. At the present moment in time, information globalization is being pursued so that it may be used by different authors and with different ends (identification, key production, floras, field guides, etc.). Delta has been standard of the IUBIS-TDWG since 1991, and in September 1998, after analyzing the new challenges for the scientific community, the subgroup SDD was set up to develop an XML-based standard for representing and handling descriptive information of organisms. The aim is to design a sufficiently standardized knowledge representation model which is accepted by the entire community of biologists, whatever their discipline of origin (zoology, botany, etc.). SDD shall attempt to provide a flexible and independent platform to facilitate the exchange of data sets without loss of information between applications, in addition to using a same description for different purposes. Fig. 2 shows the general structure of an SDD document.

2.3. Tools for the generation of identification keys

There are certain tools for generating interactive keys manually. This is the case of LucID (CBIT, 1994), Linnaeus II (Schalk & Heijman, 1996; Schalk & Troost, 1999), Xid (Intelsys, Inc., 2000***–2001), Meka (Duncan & Meacham, 1986; Meacham, 1986–1996), NaviKey (Bartley & Cross, 2000; University of Toronto, Department of Botany, 1996). This type of tool does not give any information about the suitability of the selected characters for branching nor does it check the existence of inconsistencies in the key (unclassified objectives, dead ends, etc.). Pankey (Pankhurst, 1991; Pankhurst & Pullan, 1994) is a commercial system based on Delta and available for MS-Dos which incorporates two tools: Key3m3 (to generate keys totally automatically), and Kconi (to generate keys under the user's supervision). It is capable of showing the partial key resulting from the choice of characters made by the user and to suggest the best option at each step. We should also mention Delta which is a general system for the field of systematics based on the model with the same name which includes Key, a printed key generation program

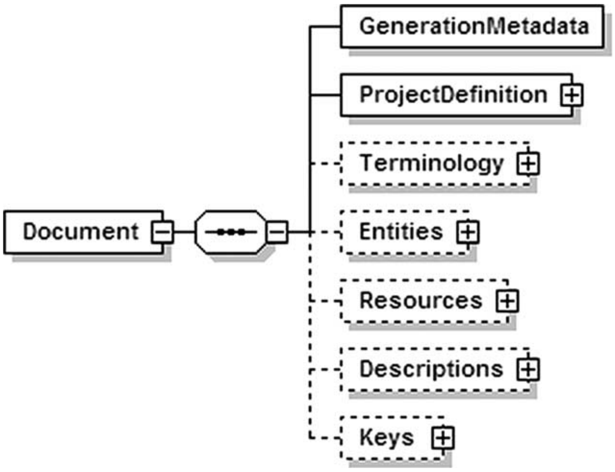


Fig. 2. General structure of an SDD document 1.0. In TDWG-SDD (2005).

which uses text files to register the user's preferences and thereby offer some interactivity.

This review reveals the shortage of tools, and also the fact that until recently, they were developed by biologists and have become obsolete. In addition, they incorporate few facilities for interactivity, and can be improved from the computing point of view. Consequently, we have developed XKey, an SDD-based tool which combines proven artificial intelligence techniques with the specific functional needs of taxonomy. It has been designed and implemented within the framework of a multidisciplinary team and from the study of previously developed tools, and therefore represents an integratory vision from the viewpoints of biologists and computer specialists.

3. Analysis of the problem

3.1. Construction of identification keys

An identification key is a tree-like structure in which each node represents a character, and the arcs its possible values. Because of its structure, it is constructed by performing recursive partitions of the set of training cases in order to finally obtain a decision tree. In short, when a node is constructed or expanded, the subset of training cases belonging to each class is considered. If all the examples belong to one class, or if some stopping rule is verified, the node is a leaf of the tree. Otherwise, the training set is divided into subsets (using a rule of heuristic division), and the same procedure is applied to each subset of the training set.

Due to the special characteristics of biology, an identification key does not totally fit the definition of a decision tree in the classic sense. Below, we shall describe the characteristics of these keys and how they are adapted by our algorithm to generate identification keys, as outlined in Fig. 3.

3.2. Objectives of the division criterion

The general aim is to obtain a compact decision tree. In accordance with the principle of Occam's razor, a small decision tree enables a better understanding of the classification model obtained. This principle, although allowing for the construction of easily understandable models, does not guarantee that these are better than other apparently more complex ones (Domingos, 1998, 1999). From the point of view of taxonomy, a good division rule (Dallwitz et al., 2000):

1. divides the *taxa* into uniform subgroups;
2. selects characters with high *reliability* and low *intra-taxon variability* (we understand *intra-taxon variability* to be the set of characters which can distinguish between individuals or populations belonging to a same taxonomical category);
3. discriminates the individuals which are going to be identified more frequently in the first steps (this frequency is also called *abundance*).

Although many division criteria have been proposed, we shall focus our study on four of these: the first three because they are classic and widely used, and the fourth because it has been proposed from the area of taxonomy.

3.3. Classic division criteria

Within this first group of criteria, we can distinguish:

- *Entropy* (Quinlan, 1986). This measures the impurity of a node of the tree and reaches its minimum when all the elements of the node are the same. It satisfies the requirements established by Dallwitz, i.e. it divides the taxa into more uniform groups and minimizes the intra-taxon variability. It also considers the abundance of an individual, so that the most frequent will appear in the first steps of the key. The entropy normally constructs decision trees with a high degree of branching, which favors questions with the most possible results. This aspect does not always reflect the work methodology of the expert biologist, who, depending on the objective with which the key is developed, may have other criteria which are different from the minimization of the length.
- *The gain ratio criterion* (Quinlan, 1993). This measurement is based on the entropy which normalizes the information gain obtained in order to avoid the construction of decision trees which classify the cases using their keys. It has been observed that, like the entropy, it favors partitions of the training set which are very uneven in size when any of them is very pure (all the cases which it includes correspond to the same class) even if it is not particularly significant (covering very few training cases).
- *The Gini diversity index* (Breiman, Friedman, Olshen, & Stone, 1984). This is a measurement for class diversity in a tree node which attempts to minimize the impurity existing in the subsets of training cases generated when the decision tree is branched.

3.4. Division criteria in taxonomy

In the area of taxonomy, very few criteria have been described for generating identification keys, and one such example is that used by Dallwitz et al. (2000). This criterion calculates the cost of using a certain character i according to the expression in Formula 1. Since it concerns a cost measurement, the aim is to minimize it, and consequently characters with a low cost will be preferred.

where:

- s is the number of values for the attribute;
- n_j is the number of *taxa* in the jth subgroup;

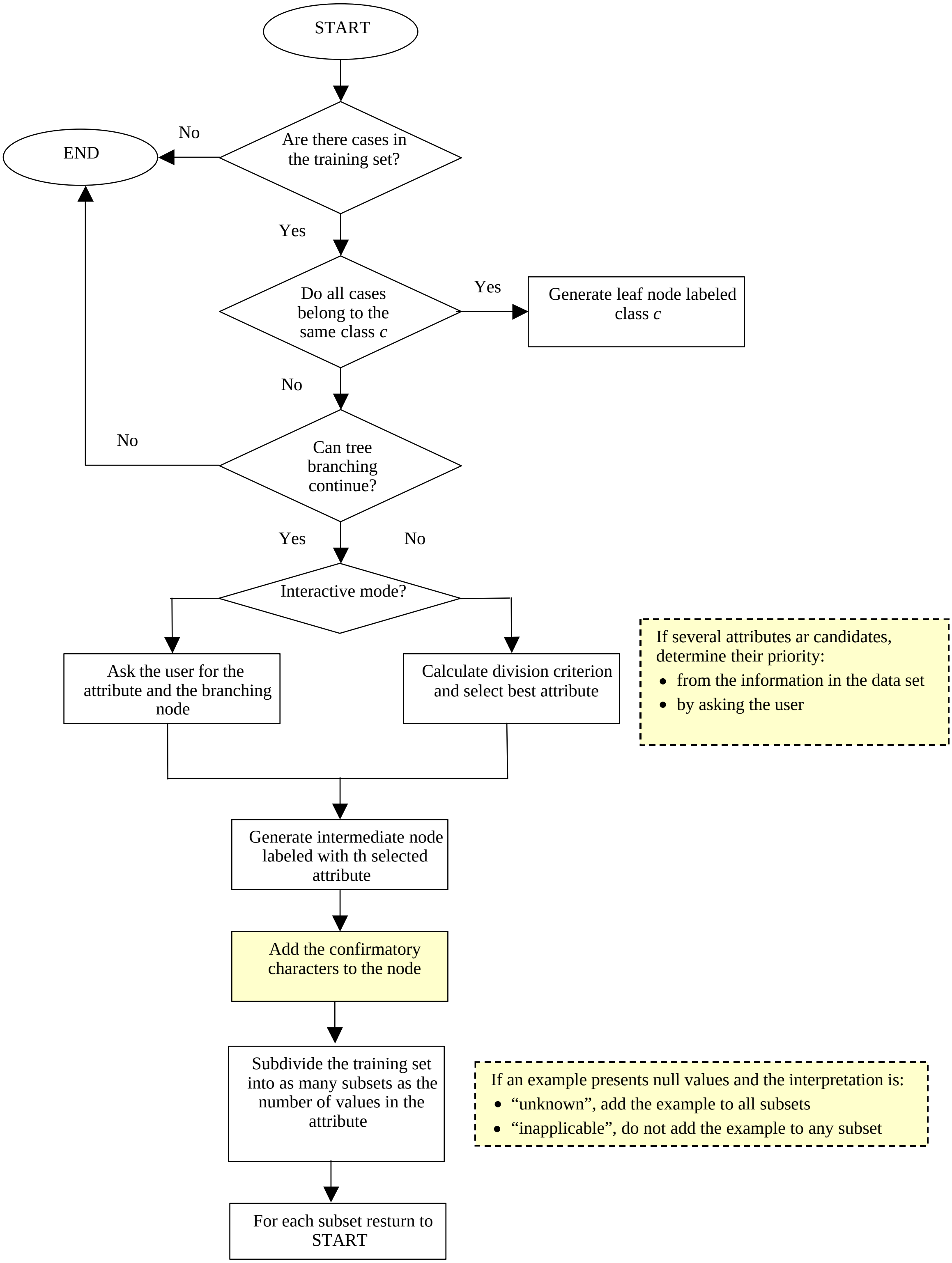

START
Are there cases in the training set?
No
END
Yes
Do all cases belong to the same class *c*
Yes
Generate leaf node labeled class *c*
No
Can tree branching continue?
No
Yes
No
Interactive mode?
Ask the user for the attribute and the branching node
Calculate division criterion and select best attribute
If several attributes ar candidates, determine their priority:
• from the information in the data set
• by asking the user
Generate intermediate node labeled with th selected attribute
Add the confirmatory characters to the node
Subdivide the training set into as many subsets as the number of values in the attribute
If an example presents null values and the interpretation is:
• "unknown", add the example to all subsets
• "inapplicable", do not add the example to any subset
For each subset resturn to START

Fig. 3. Algorithm for the generation of identification keys.

- c is the cost of a character. This is related to the *reliability* using the expression $c=\text{Rbase}^{5-r}$, where:
    - ○ r is the value ascribed to the *reliability* of the character (a value fixed by the expert) and takes values in [0–10].
    - ○ Rbase is a constant which takes values in the interval [1,5].
- $c_{\min}$ is the lowest cost for the characters being considered;
- f_j is the total frequency of the items in the jth group.

The frequency *f* of an item is given by the expression where:

  - ○ *a* is the *abundance* of the item (a value fixed by the expert) and takes values in [0–10]
  - ○ Abase is a constant which takes values in the interval [1,5].

- *V* controls the effect of the intra-taxon variability. In its expression:
  - ○ Varywt is a constant value. If its value is 0, those characters which have any intra-taxon variability are excluded from the key. If, on the other hand, the value of this parameter is 1, this aspect is not penalized. It takes values in the interval [0–1].
  - ○ *n* is the number of *taxa*.

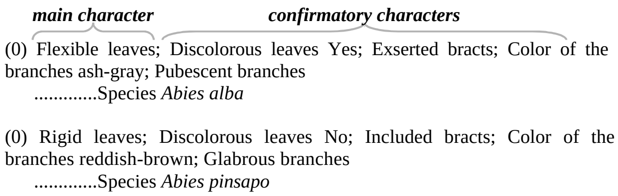
3.5. Treatment of null values

When a key is generated, it is common for null values to appear in the data set. The final result of the process depends on the interpretation made by the algorithm of these. In taxonomy, there could be three different interpretations for the appearance of null values:

1. The value does not appear because the attribute is inapplicable for a certain taxon. In this case, the taxon is not added to any branch of the tree. It should be observed that when this interpretation of the null values is used, the taxon to which the null value corresponds could remain unclassified.
2. The value does not appear because it is indifferent. This means that, for a certain taxon, the attribute can take any valid value. If this character is used for branching the decision tree, it is necessary to add the taxon with the null value to each branch of the tree.
3. The value does not appear because it is unknown. In this case, and in order to avoid loss of information, the same procedure is followed as in the previous case.

3.6. Reliability of a character

In taxonomy, not all the characters have the same relevance: some are more important than others, for example, owing to the ease with which they are observed, because they are distinguishing characters, etc. This aspect is called the *reliability* of a character (Dallwitz, 1974). If the expert provides information about the most reliable character for branching, it is possible to combine the measurement given by a division criterion and expert knowledge about the domain, in order to decide in those cases in which several characters present the same value of the division criterion. In this way, the key will adapt better to the characteristics of each taxonomical group and will be more satisfactory.


main character
confirmatory characters

(0) Flexible leaves; Discolorous leaves Yes; Exserted bracts; Color of the branches ash-gray; Pubescent branches
............Species *Abies alba*

(0) Rigid leaves; Discolorous leaves No; Included bracts; Color of the branches reddish-brown; Glabrous branches
............Species *Abies pinsapo*

Fig. 4. Key with confirmatory characters. Author: Eva Gibaja Galindo.

3.7. Confirmatory characters

It is normal in a key to include more than one character in the same node of the classification tree (see Fig. 4). This means that the node contains a main character and set of characters called *confirmatory characters* (Dallwitz et al., 2000). This is a fundamental aspect since it enables:

- identification to be continued when the main character is not available;
- confirmation of the decision to continue identification by a certain node.

The main character and the confirmatory characters are equivalent, and any of these can be used to branch the tree. This means that, within the *taxa* group being considered, the two characters:

1. have the same value for the division criterion;
2. have the same number of possible values;
3. generate identical *taxa* groups since they are used as the branching character.

3.8. Construction of knowledge bases

By taking advantage of the construction of the decision tree, it is possible to generate a knowledge base comprising rules from this. In botany, the classification problem and subsequent identification has singular characteristics: a fold does not, in every case, have the complete individual for its identification, and the samples gathered frequently depend on seasonality, age of the individual, etc. This can lead to a very well-informed consultation not offering conclusions if there is no value for any character. In order to tackle this aspect, various decision trees are constructed (which shall subsequently be converted into rules), taking a different generator node each time. For this reason, it is necessary to include a consistency reinforcer, and since the syntax of a rule is sufficiently restrictive, it systematically analyses each rule in the knowledge base to detect and correct unnecessary if conditions, redundant rules, conflictive rules, and subsumed rules (see Fig. 5).

3.9. Treatment of uncertainty

Given the characteristics of taxonomical identification, the application of some method for the treatment of

```
Let $n$ the number of rules in the knowledge base
FROM i=1 TO i=$n$-1
  FROM j=i TO j= $n$
    If rule$_i$ and rule$_j$ have the same consequent
      If rule$_i$ and rule$_j$ have the same number of antecedents
        If CF(rule$_i$)=CF(rule$_j$)
          • All the antecedents are the same except for one which is
            contradictory: UNNECESSARY IF CONDITION ⇨ Elimminate
            rule$_j$, and elimminate the unnecessary condition of rule$_i$.
          • All the antecedents are the same: REDUNDANT RULES ⇨
            Elimminate rule$_j$.
        If CF(rule$_i$)≠CF(rule$_j$)
          All the antecedents are the same
            CF(rule$_i$) = -CF(rule$_j$): CONFLICTIVE RULES ⇨ Elimminate the two rules.
      If rule$_i$ and rule$_j$ have different number of antecedents
        CF of the shortest rule is ±1
          The antecedents of the shortest rule are contained in those of the longest then
          SUBSUMED RULES ⇨ Elimminate the longest rule.
```

Fig. 5. Algorithm for the XKey consistency reinforcer.

uncertainty is extremely useful in order to enrich knowledge base consultation. There are various theories for treating uncertainty such as the theory of evidence (Dempster, 1967; Shafer, 1976), or the probability or fuzzy set theory (Zadeh, 1965). Due to the absence of historical data and the difficulties of defining fuzzy variables and probability mass functions, we have chosen the theory of certainty factors (Shortliffe & Buchanan, 1975). This model has been used successfully in many other recently published systems (Cabrero-Carnosa, Castro-Pereiro, Graña-Ramos, Hernández-Pereira, Moret-Bonillo and Martín-Egaña, 2003; Mahaman, Harizanis, Filis, Antonopoulou, Yialouris and Sideridis, 2002). Furthermore, the theory of certainty factors is a computationally simple and more natural model for a non-mathematical user than other approaches (Harrison & Kovalchic, 1998).

Taking all this into consideration, each rule in the knowledge base has an associated certainty factor (CF). This value, which is determined during generation and remains unchanged unless redefined by an expert, is ascribed according to the following criteria:

- if a pure leaf node is generated (all the elements belong to the same class), the associated rule has an associated certainty factor with a value of $+1$;
- if the leaf node is not pure, the result is that of a rule with disjunction in the consequent, which is fractionized into rules with simple consequents. The certainty factor of the original rule is also fractionized by the following Formula 2

where

- $p(x)$ is the probability of the objective in question occurring in the initial table;
- $p(x/a)$ is the probability of objective x occurring knowing that antecedent a occurs, regardless of whether this is simple or compound.

4. Characteristics of XKey

4.1. Use of different division criteria

XKey enables the division criterion to be selected, and includes the four criteria presented in Sections 3.3 and 3.4: entropy, the gain ratio criterion, the Gini diversity index, and the division rule proposed by Dallwitz. This allows generated keys to be compared from a same set of data but using different criteria and aids a better adaptation to each specific problem. The criterion is selected from the user options menu (Fig. 6).

4.2. Treatment of null values

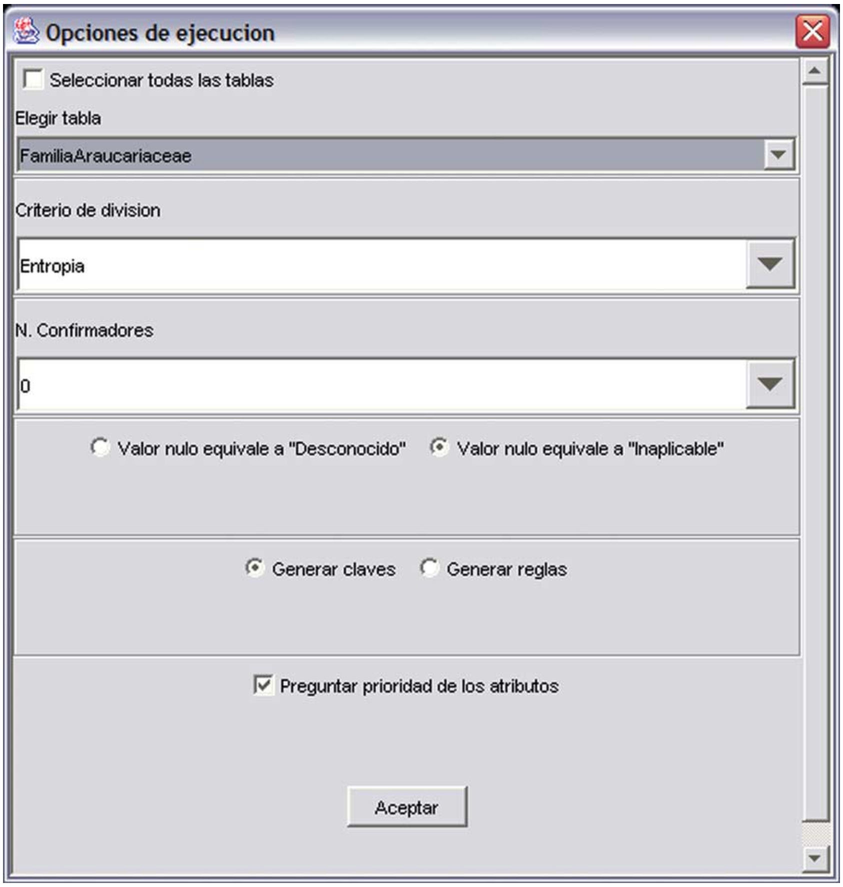
In its operation mode, XKey reflects the three different interpretations which we can give to the appearance of null values in taxonomy (see Fig. 5) and enables the distinction to be made between:

1. not applicable (the taxon with the null value is not added to any tree branch);
2. indifferent (XKey adds the taxon with the null value to each of the tree branches);
3. unknown (XKey acts as in Case 2).

SDD also enables explicit representation of null values. For this, it uses the global states ‘Unknown’ and ‘NotApplicable’. An attribute with the special value ‘Unknown’ will be treated by XKey as unknown (Case 3), whereas an attribute with the special value ‘NotApplicable’ will be treated as inapplicable (Case 1) independently of the general interpretation with which XKey has been configured.

4.3. Control of the reliability of a character

The SDD data sets do not include information about the *reliability* of a character, so when two attributes have



Opciones de ejecucion

☐ Seleccionar todas las tablas

Elegir tabla: FamiliaAraucariaceae

Criterio de division: Entropia

N. Confirmadores: 0

○ Valor nulo equivale a "Desconocido" ● Valor nulo equivale a "Inaplicable"

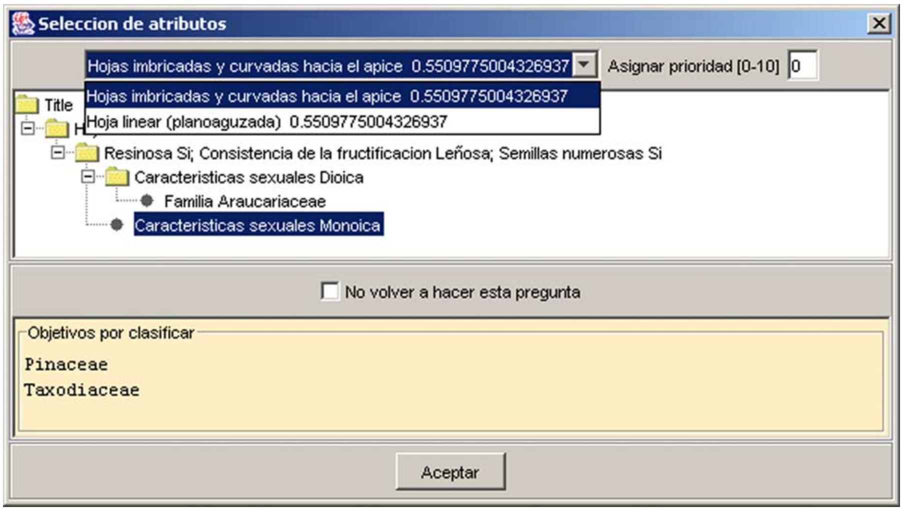
● Generar claves ○ Generar reglas

☑ Preguntar prioridad de los atributos

Fig. 6. Execution options of XKey.

the same value for the division criterion, the construction of the tree depends on the order in which the data are represented. In order to avoid this situation, XKey shows the user the different alternatives, the current state of the tree, and the objectives which must still be classified so that he/she may select the attribute which they consider to be the most suitable. It is also possible to ascribe a priority value in execution time to a specific attribute. This weight will be remembered by the system throughout execution and will be used to decide between several branching attributes with the same value for the division criterion (see Fig. 7).

At certain times, the data set can be too large to apply this strategy, or it could be that the expert is not interested in making this type of choice. For this reason, users can tell



Seleccion de atributos

Hojas imbricadas y curvadas hacia el apice 0.5509775004326937 Asignar prioridad [0-10] 0

Hojas imbricadas y curvadas hacia el apice 0.5509775004326937
Hoja linear (planoaguzada) 0.5509775004326937

Title
H
Resinosa Si; Consistencia de la fructificacion Leñosa; Semillas numerosas Si
Caracteristicas sexuales Dioica
Familia Araucariaceae
Caracteristicas sexuales Monoica

☐ No volver a hacer esta pregunta

Objetivos por clasificar
Pinaceae
Taxodiaceae

Fig. 7. Attribute selection in execution time.

XKey at any time that they no longer want to be asked about this aspect.

4.4. Inclusion of confirmatory characters

In order to adjust to this feature, XKey includes an option in the execution options menu so that users can indicate whether they want to include this type of character and the maximum number of confirmatory characters per tree node. The choice of the value will depend on the user's criterion and on the final use of the key; the final quantity included will depend on the characteristics of each data set. It is therefore possible to indicate that the inclusion of confirmatory characters is not required.

4.5. Operation modes

One difference between the learning of traditional classification models and identification keys is interactivity in character selection. The best character from the point of view of the division criterion may not be the best character from the expert's point of view: it might be difficult to visualize, at times it is preferable to eliminate the exceptions in the first steps of the key, etc. For these reasons, XKey offers various operation modes:

- *Automatic mode*. This applies the algorithm for the generation of decision trees without the user's intervention and uses the branching criterion selected in the options menu. The algorithm for the automatic generation of XKey matches the general guidelines of a TDIDT algorithm (*Top-Down Induction On Decision Trees*). What XKey offers is its capacity to add confirmatory characters, to treat null values according to the interpretation established by the user, and to select, in execution time, the best branching attribute (if several attributes have the same value for the division criterion).
- *Semi-automatic mode*. This applies the algorithm for the generation of decision trees automatically, but it consults the user in those cases where there are ties in the selection of the branching attribute. XKey shows the list of all the characters which can be used for branching, ordered according to the division criterion. In this way, it provides very useful statistical information which the expert does not have access to when generating keys manually with other tools.
- *Interactive mode (supervised by the user)*. This follows the schema for automatic generation, to which interaction facilities are added. The operations permitted during execution in interactive mode are:
    - *Add node*. The user selects the node to branch and the branching attribute, and generates the subtree corresponding to this node. XKey again shows the list of all the characters which can be used for branching, ordered according to the division criterion.
    - *Delete node*. In this case, the node and all its descendants are eliminated. For this, it is necessary to maintain an execution memory for each node so that it may return to previous states.
    - *End*. This option enables the user to establish where certain characters will appear so that the system can subsequently finish the execution automatically. When this operation is selected, XKey analyzes the classification model in order to detect the leaf nodes which do not contain classes of the example (intermediary nodes) so as to finish construction of these nodes automatically.

4.6. Diversity of output formats

Once the identification key has been generated, it is presented to the user in tree form (see Fig. 8). It is possible to save this key in the formats described below:

- *Text format*. The key is saved in a plain text file. This format facilitates the modification of the key with other more sophisticated text editors, and in turn, does not limit the use of any of these.
- *XML format*. The key is saved in an XML format file, with a similar structure to the expected format of the SDD keys. This format enables the keys to be edited once they have been generated and for information to be easily exchanged.

4.7. Construction of a knowledge base

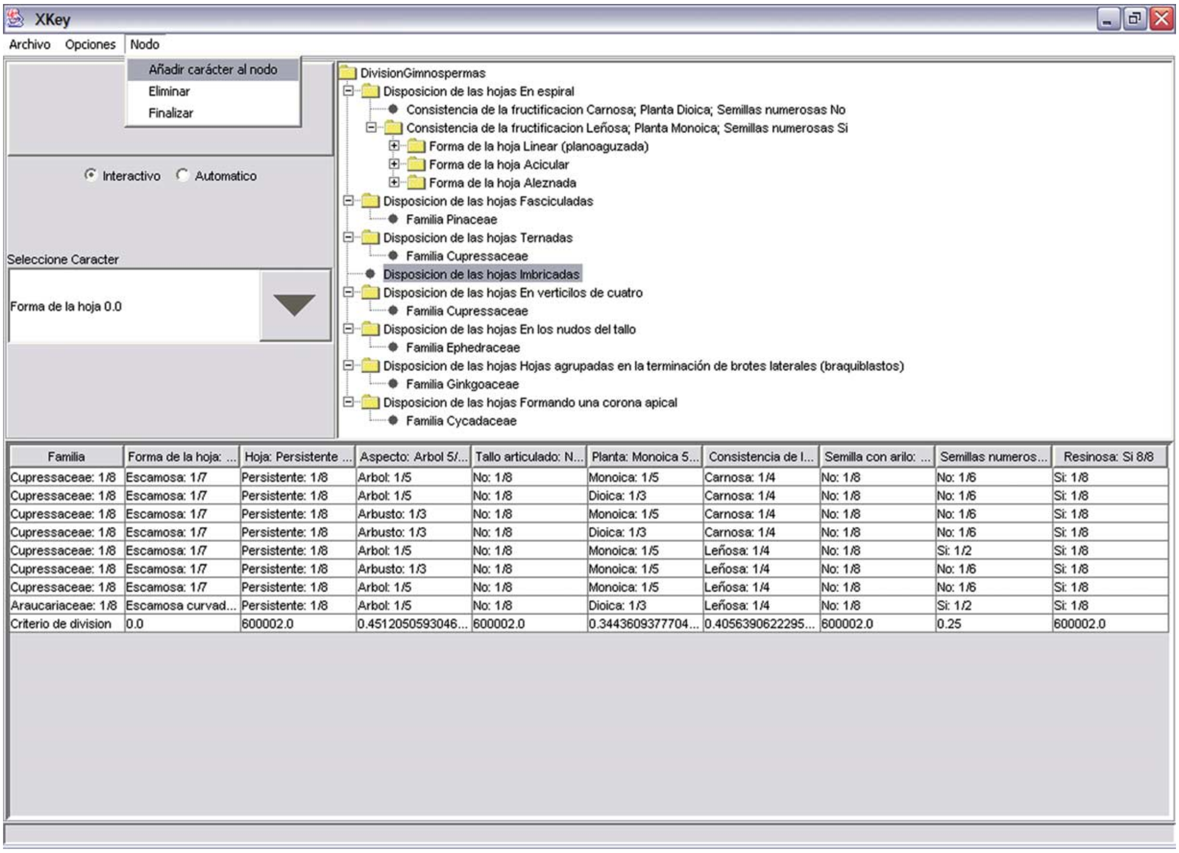
XKey incorporates the generation of a knowledge base into its functionality from the data set as described in Sections 3.8 and 3.9. If generation of a knowledge base is chosen, this can be saved in:

- CLIPS format (Giarratano, 1998). The key is saved in the form of rules for this well-known expert system *shell*, with which we can intercommunicate the standard XML model for taxonomical description with one of the most popular *shells* within the field of expert systems.
- GREEN format (Fajardo, Gibaja, Bailon, & Moral, 2003). We have mentioned that as well as generating keys, XKey can generate sets of rules which can be used directly by the GREEN system.

4.8. Execution results

In addition to the keys, the execution of XKey returns a set of measurements which facilitate analysis of the results obtained. This information includes:

- average length of the key;
- typical deviation of the length of the key (in order to compare some keys with others, the system returns Pearson's skewness coefficient);



● Interactivo ○ Automatico

Seleccione Caracter: Forma de la hoja 0.0

- DivisionGimnospermas
  - Disposicion de las hojas En espiral
    - Consistencia de la fructificacion Carnosa; Planta Dioica; Semillas numerosas No
    - Consistencia de la fructificacion Leñosa; Planta Monoica; Semillas numerosas Si
      - Forma de la hoja Linear (planoaguzada)
      - Forma de la hoja Acicular
      - Forma de la hoja Aleznada
  - Disposicion de las hojas Fasciculadas
    - Familia Pinaceae
  - Disposicion de las hojas Ternadas
    - Familia Cupressaceae
  - Disposicion de las hojas Imbricadas
  - Disposicion de las hojas En verticilos de cuatro
    - Familia Cupressaceae
  - Disposicion de las hojas En los nudos del tallo
    - Familia Ephedraceae
  - Disposicion de las hojas Hojas agrupadas en la terminación de brotes laterales (braquiblastos)
    - Familia Ginkgoaceae
  - Disposicion de las hojas Formando una corona apical
    - Familia Cycadaceae

Familia	Forma de la hoja: ...	Hoja: Persistente ...	Aspecto: Arbol 5/...	Tallo articulado: N...	Planta: Monoica 5...	Consistencia de l...	Semilla con arilo: ...	Semillas numeros...	Resinosa: Si 8/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbol: 1/5	No: 1/8	Monoica: 1/5	Carnosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbol: 1/5	No: 1/8	Dioica: 1/3	Carnosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbusto: 1/3	No: 1/8	Monoica: 1/5	Carnosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbusto: 1/3	No: 1/8	Dioica: 1/3	Carnosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbol: 1/5	No: 1/8	Monoica: 1/5	Leñosa: 1/4	No: 1/8	Si: 1/2	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbusto: 1/3	No: 1/8	Monoica: 1/5	Leñosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Cupressaceae: 1/8	Escamosa: 1/7	Persistente: 1/8	Arbol: 1/5	No: 1/8	Monoica: 1/5	Leñosa: 1/4	No: 1/8	No: 1/6	Si: 1/8
Araucariaceae: 1/8	Escamosa curvad...	Persistente: 1/8	Arbol: 1/5	No: 1/8	Dioica: 1/3	Leñosa: 1/4	No: 1/8	Si: 1/2	Si: 1/8
Criterio de division	0.0	600002.0	0.4512050593046...	600002.0	0.3443609377704...	0.4056390622295...	600002.0	0.25	600002.0

Fig. 8. Interactive selection of attributes with XKey.

- maximum and minimum length of the key;
- number of terminal and non-terminal nodes;
- number of OR terminal and exclusive nodes (the appearance of OR nodes indicates that it is not possible to clearly identify a particular node, and, therefore, that the data set is incomplete);
- number of total attributes in the data set and number of attributes used in the key;
- number of confirmatory attributes included in the key.

5. Test of the XKey tool

5.1. Execution data for the XKey tool and methodology followed

The information provided by XKey enables the keys obtained for the same data set to be compared by using, for example, different division criteria. In our case, experiments have been carried out with a real group: Gymnosperms present in the Iberian peninsula. The evaluation of the keys generated with XKey for this group offers us a sufficiently wide set of tests for determining what criteria, in what cases or conditions, and why they enable the most ideal key to be generated. It should be noted that in order to increase the diversity of casuistries which can be presented, not only have wild or naturalized taxa been included, but also ones which have been cultivated and which are widely used in gardening. The subgroups of Gymnosperms used are described in Table 1.

In order to carry out this experiment, a knowledge acquisition process has been performed with the participation of two expert botanists. Bibliographical sources have also been consulted, and in particular (López Gonźlez, 2001; López Gonźlez & Do Amaral, 1986), which are essential reference books for the taxonomical group being studied. In our study, we have analyzed a set of qualitative aspects which are directly related to the reliability of the generated key:

Table 1
Subgroups of *Gymnospermae* being studied

Division Gymnospermae	Family Araucariaceae
Family Cephalotaxaceae	Family Cupressaceae
Family Cycadaceae	Family Ephedraceae
Family Ginkgoaceae	Family Pinaceae
Family Taxaceae	Family Taxodiaceae
Genus Abies	Genus Calocedrus
Genus Cedrus	Genus Chamaecyparis
Genus Cryptomeria	Genus Cupressus
Genus Cycas	Genus Ephedra
Genus Juniperus	Genus Larix
Genus Picea	Genus Pinus
Genus Platycladus	Genus Pseudotsuga
Genus Sequoia	Genus Sequoiadendron
Genus Taxodium	Genus Tetraclinis

- *Average length of the key*. This is an indicator of the average number of steps for identification. This is an important aspect since the expert often searches from the key which leads to the identification in the least number of steps.
- *Typical deviation of the average length*. This is an indicator of the equilibrium of the keys: if the typical deviation is large, the different paths of the tree have very different lengths.
- *Terminal nodes/number of objectives ratio*. This is an indicator of the degree of branching of the keys: if there are many more leaf nodes than objectives to be classified, the tree is extremely branched.
- *Internal nodes/confirmatory characters ratio*. This is an indicator of the power of generating confirmatory characters of a certain division criterion. Those criteria which identify a greater quantity of confirmatory characters are best.
- *Characters used/total number of characters ratio*. With this, it is possible to determine the ratio of attributes in the data set which have been used to generate the key and what these attributes are.
- In addition to the quantitative interpretation of the results, another important aspect has been the analysis of the adaptation to the reality of the keys obtained.

5.2. Observations about the average length of the keys generated

From the expert botanist's point of view, the common factor is the inadequacy of the character selection of the keys when the Gini diversity index is used as the division criterion. This criterion always produces longer keys (see Fig. 9) and trees with a greater number of internal and external nodes (see Fig. 10).

If we consider the sum of differences in relation to the minimum average length (Fig. 11), the keys generated with the entropy criterion are practically in all the cases of minimum length. In second place is the profit gain ratio, followed by Dallwitz's criterion, which does not produce as good results due to the fact that the measurement it uses for counting the intra-taxon variability is cruder than that used by the two previous criteria: Dallwitz only considers the number of different *taxa* in each partition, whereas the other two also count the frequency of each taxon.

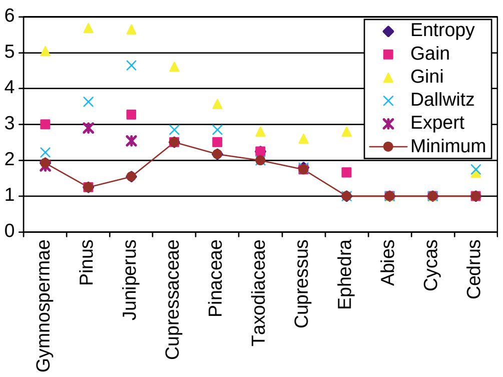


Fig. 9. Average length of the keys.

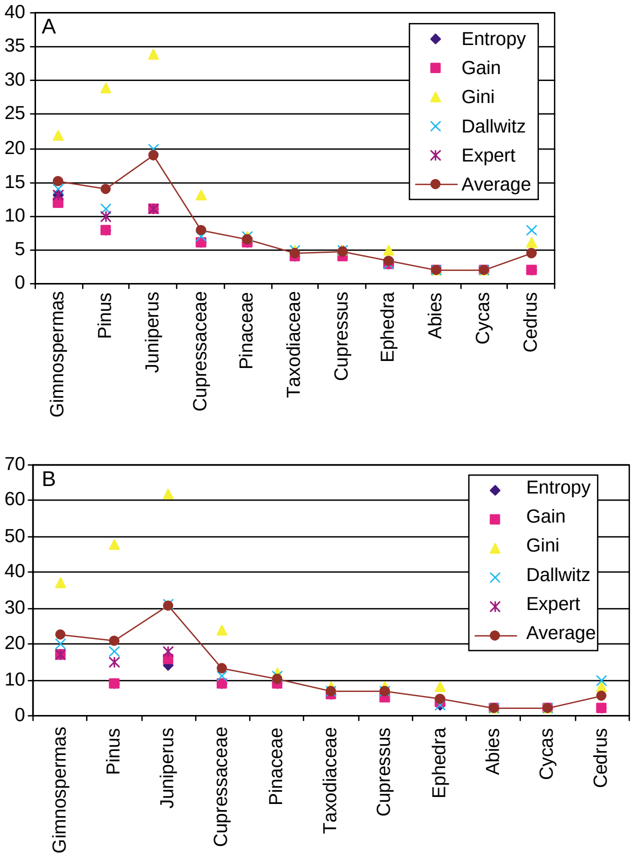


Fig. 10. Number of external (A) and internal nodes (B).

5.3. Observations about the equilibrium of the keys

The deviation of the length of the branches in relation to the average is always greater for the Gini measure, which indicates that the length of the tree branches generated with

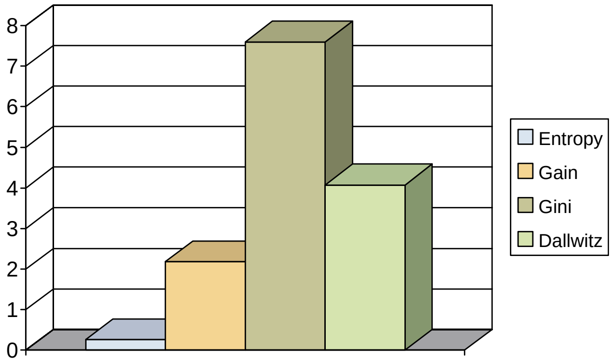


Fig. 11. Sum of the differences in relation to the minimum average length.

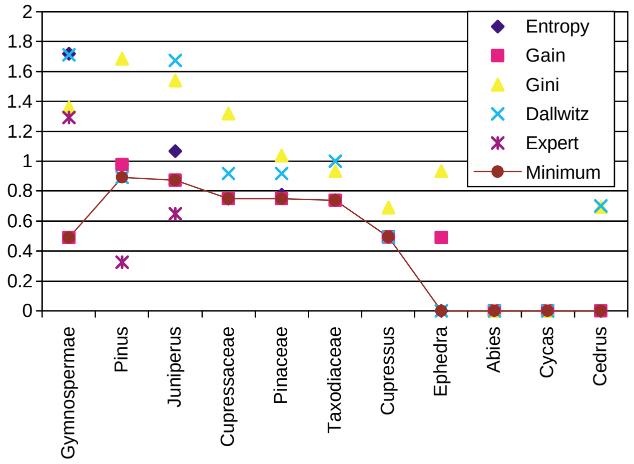


Fig. 12. Typical deviation of the average length.

this criterion is more variable. The most balanced keys are those generated with the gain ratio criterion, followed by those generated with entropy and Dallwitz's division criterion (Fig. 12). The sum of the differences in relation to the minimum typical deviation (Fig. 13) is less for the gain ratio, followed by entropy.

5.4. Observations regarding the degree of branching

The gain ratio and entropy criteria generate the least branched key, i.e. they tend to detect the smallest number of possible paths for a certain objective and, therefore, to generate more compact models than Gini and Dallwitz's criteria (Figs. 14 and 15).

5.5. Observations about the number of confirmatory characters included

The entropy criterion produces a better *number of confirmatory attributes/number of internal nodes* ratio, followed by the profit gain criterion (see Fig. 16). If we observe the differences regarding the maximum of the *confirmatory/internal nodes* ratio in each case, we can see how the entropy is the one which presents a smaller difference, followed by the profit gain criterion and, some way behind, by the division criteria of Dallwitz and Gini (see Fig. 17).

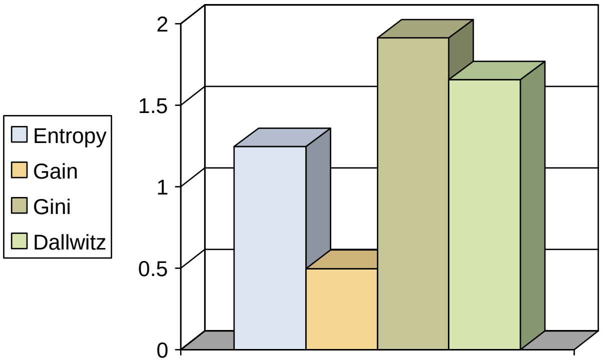


Fig. 13. Sum of the differences in relation to the minimum typical deviation.

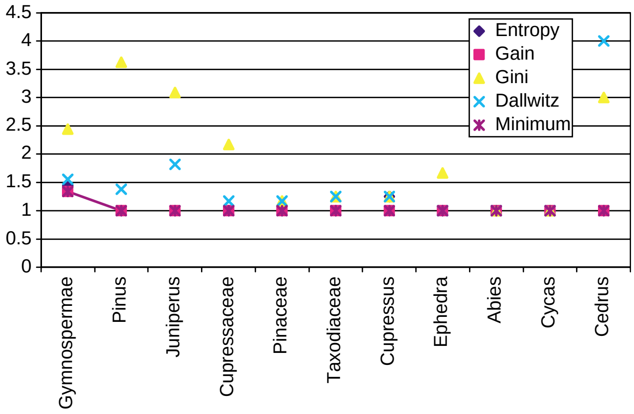


Fig. 14. *Terminal/objective ratio.*

5.6. Observations about the number of characters used

The entropy uses the least number of characters in the generation of keys, i.e. it locates the smallest amount of information in order to carry out identification. In second place is the gain ratio criterion, followed by Dallwitz's criterion. The Gini diversity index, in addition to producing longer keys, includes a greater number of attributes in the generation of keys (Figs. 18 and 19).

5.7. Qualitative comparison of the criteria

The tests carried out show that entropy and gain ratio offer the best results for all the measures proposed. We can see a summary in Table 2: 4 is the score given to the best criterion, 1 is the score for the worst criterion, and 2 and 3 are the scores for intermediary cases. Table 2 shows that it is precisely entropy and profit gain which are the best criteria.

The measure used by Dallwitz also produces good results, but in some unwanted cases since it measures the intra-taxon variability more crudely. The difference with Dallwitz's division criterion can be clearly seen in the case

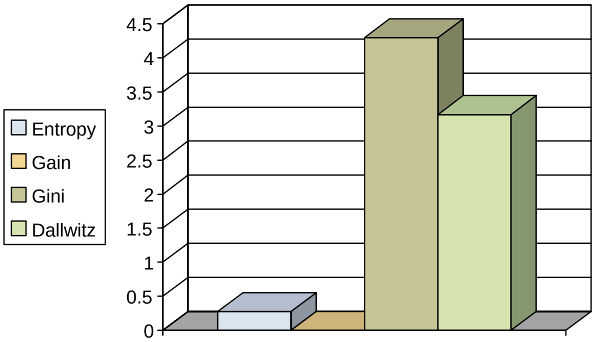


Fig. 15. Sum of the differences in relation to the minimum *terminal/objective* ratio.

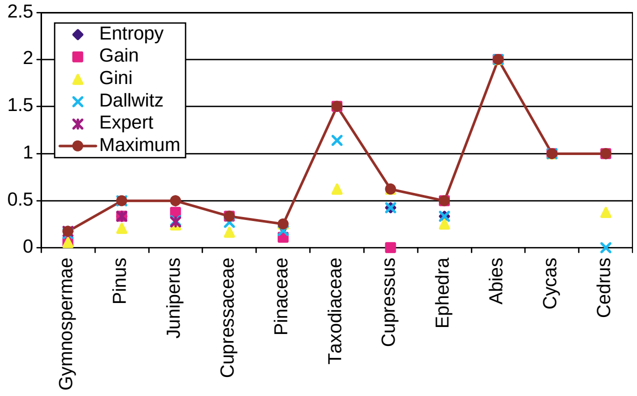


Fig. 16. Number of *confirmatory characters/internal node*.

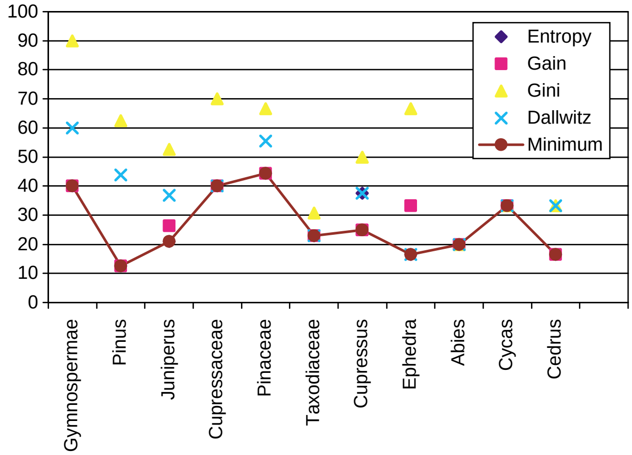


Fig. 18. Ratio of *total number of characters/characters used*.

of the keys of the genus *Cedrus*: in order to separate two *taxa*, Dallwitz gives five paths for *C. atlantica* and three for *C. deodara* (Table 3). With the entropy, we have two classes of *Cedrus* and one path to arrive at each species (Table 4).

5.8. Discussion and biological interpretation of the results obtained/effect of interactivity

The entropy produces more optimized keys in terms of the number of questions, but in the field of biology, this aspect must be qualified. At times, there is no reason why the distinguishing character from the point of view of information theory need be the best character from the biological point of view, due, for example, to the difficulty in its observation (need for microscopic observation, temporality, size, layout, etc.). This is a fundamental aspect as the identification process is interrupted if no character is available. It is therefore necessary for those characters which as well as having a high distinguishing power can be easily observed to be included in the first steps of the key. All this entails considering not only the division criteria but also the characteristics of the taxonomical group being studied and the future receptors of the keys.

The possibility of choosing between the four criteria set out in our article enables different types of keys to be produced according to the type of work, research, and end user:

- If the key is going to be all the information about the taxa which shall be available in the end, we shall need long keys which gather the greatest number of possible characters.
- If the keys are the tool to access information and, once the taxon has been identified, this is accompanied by an exhaustive description, we can use more precise keys with a high level of discrimination.
- The generation of keys for experts, which assumes an advanced knowledge of the group and of the plant morphology and characteristics, prefers criteria such as that of the entropy.
- The case of considering keys as a tool for an intermediary user would probably entail using the ideal criterion according to the information theory, but combined with interactivity.

It might be that the attributes on which the entropy has generated its key are not the ones which an expert in the field would consider to be the most suitable. In these cases, the combination of interactivity with a division criterion

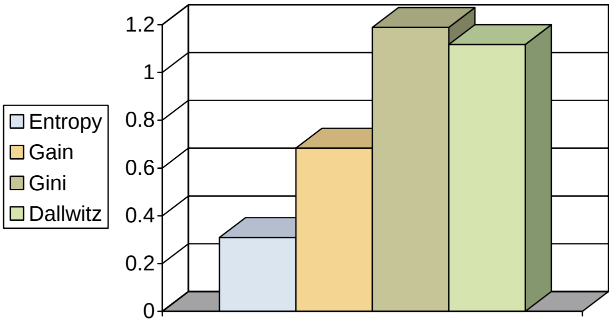


Fig. 17. Sum of the differences in relation to the maximum number of *confirmatory characters/internal node*.

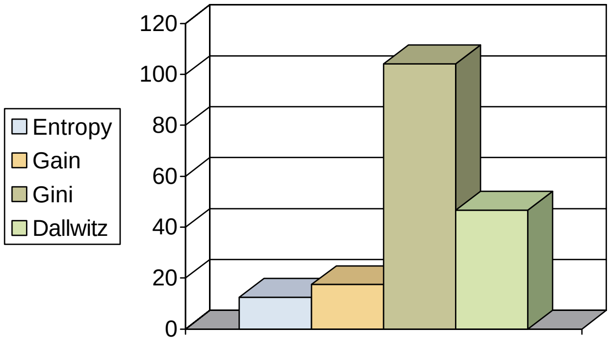


Fig. 19. Sum of the differences in relation to the minimum of the *total number of characters/characters used* ratio.

Table 2
Comparison of different division criteria on the *Gymnospermae* group

	Length	Balance	Branching	Confirmatory	No. of characters	Total
Entropy	4	3	3	4	4	18
Profit	3	4	4	3	3	17
Gini	1	1	1	1	1	5
Dallwitz	2	2	2	2	2	10

such as the entropy produces very satisfactory results: the entropy suggests and orders the characters, but it is the expert who ultimately decides which character to select by using his/her knowledge about the specific taxonomical group.

In the case of the keys for Gymnosperm families, we have managed to reduce the average length of the key in relation to the entropy by means of the use of expert knowledge (see Fig. 20).

A similar case occurs with the keys of the genus *Pinus*. In this case, the character in which the key is based is '*characteristics of the apophysis*', which rapidly separates the different species of pine trees. It is not easy, however, to see this character as it refers to the pine cone, a character which is not always present and which supposes knowledge of the plant morphology which is not within the reach of any user.

According to the experts' opinion, a good attribute is '*leaf shape*'. The interactivity can be used on other levels. Continuing with the example of Pinus, the second attribute was made to be the '*presence/absence of pine nuts*'. The final result is shown in Table 5, and this became the key which best responded to the expert's expectations (Fig. 21).

6. Conclusions and final notes

Throughout this article, we have observed that the results can be better adapted to the biological reality by adding meta-knowledge. For example, by ascribing a reliability value for the characters, it is possible to decide in those cases in which the system does not have the necessary information to make a decision. This alternative has its disadvantages: firstly, all this information must be gathered in the knowledge representation model and this makes design more complicated; secondly, the large number of characters normally being dealt with; and thirdly, all the additional meta-knowledge must be entered by the expert developing the data set. It is necessary to find a balance between the quantity of the additional information and the functionality of the tools. We should also highlight that the adaptation of the results also depends on the characteristics of the data set and on the specific problem being tackled. A good selection of identification characters will result in keys which are much more satisfactory and better adapted to reality (Fig. 22).

Table 3
Key for the genus *Cedrus* generated according to Dallwitz's criterion

Genus *Cedrus* /Dallwitz	
(0) Size of the leaves (long) between 2 and 2.5 cm	1
(0) Size of the pine cone (long) between 8 and 12 cm	
Species *Cedrus deodara*	
(0) Size of the pine cone (long) between 4 and 6 cm	
Species *Cedrus atlantica*	
(1) Size of the pine cone (long) between 6 and 8 cm	
Species *Cedrus atlantica*	
(0) Size of the leaves (long) between 2.5 and 3 cm	2
(2) Size of the pine cone (long) between 8 and 12 cm	
Species *Cedrus deodara*	
(2) Size of the pine cone (long) between 4 and 6 cm	
Species *Cedrus atlantica*	
(2) Size of the pine cone (long) between 6 and 8 cm	
Species *Cedrus atlantica*	
(0) Size of the leaves (long) between 3 and 5 cm	
Species *Cedrus deodara*	
(0) Size of the leaves (long) between 1 and 2 cm	
Species *Cedrus atlantica*	

We shall end this description of our work with a set of conclusions:

1. We have developed *XKey*, a tool for generating identification keys which operate directly from data sets developed with *SDD*. *XKey* can also operate with descriptions in *DeltaAccess* and *Delta* format (by means of the use of translation utilities).
2. We have shown the adaptation of artificial intelligence techniques for automatic learning and treatment of uncertainty to the area of taxonomical identification. The expert's satisfaction with the keys generated by XKey endorse the suitability of the techniques used.
3. The output of the tool is presented in various formats: text format, XML format, and CLIPS format, and is supplemented with statistical information which facilitates the study and comparison of the results obtained.
4. *XKey* generates identification keys rapidly and conveniently, which means a considerable saving in time for the expert. It is also versatile, since it enables different division criteria to be selected, the meaning of null

Table 4
Key for the genus Cedrus generated according to minimum entropy criterion

Genus *Cedrus*/entropy
(0) Size 6000 cm; Tree guide: recurved; Hanging branches: Yes
Species *Cedrus deodara*
(0) Size 5000 cm; Tree guide: not recurved; Hanging branches: Yes
Species *Cedrus atlantica*

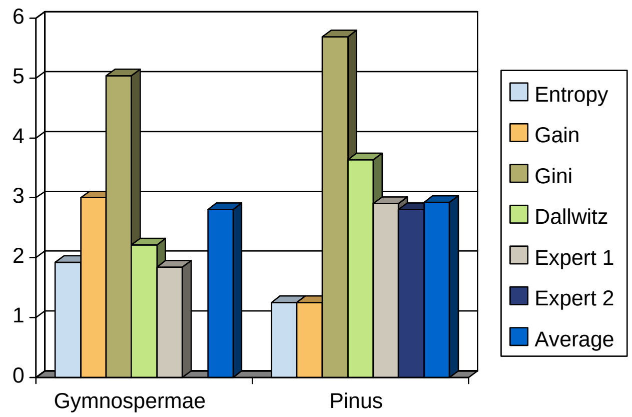


Fig. 20. Average length of the keys for the *Gymnosperma*e division and the genus *Pinus*.

values to be configured, distinguishing characters to be included, and weights to be ascribed to the characters in execution time. This added functionality produces keys with a more suitable biological meaning than those generated with classic decision trees.

5. In addition to generating identification keys, *Xkey* also generates complete, consistent knowledge bases which are compatible with the *CLIPS* system for subsequent use in expert systems.

Table 5
Key for the genus *Pinus* generated according to the expert's criterion

Genus *Pinus*/expert criterion
(0) Number of leaves per fascicle 2 1
(1) With pine nut: No; Crown shape: pyramidal; Persistent, winged seed: Yes 2
(2) Characteristics of the apophysis: prominent and sharp
Species *Pinus pinaster*
(2) Characteristics of the apophysis: not very prominent 3
(3) Colour of the bark (rhytidome): Ash-grey; Shiny pine cones: Yes; Leaf colour: intense green; Leaves: flexible
Species *Pinus nigra subsp. Salzmannii*
(3) Colour of the bark (rhytidome): Reddish-brown; Shiny pine cones: No; Leaf colour: light green; Leaves: rigid
Species *Pinus sylvestris*
(2) Characteristics of the apophysis: very prominent, hooked
Species *Pinus uncinata*
(2) Characteristics of the apophysis: not very convex
Species *Pinus halepensis*
(2) Characteristics of the apophysis: very prominent and sharp
Species *Pinus radiata*
(1) With pine nut: Yes; Crown shape: umbrella-shaped; Persistent, winged seed: No
Species *Pinus pinea*
(0) Number of leaves per fascicle 3 4
(4) Size 2500 cm; Characteristics of the apophysis: very prominent, hooked
Species *Pinus uncinata*
(4) Size 4000 cm; Characteristics of the apophysis: very prominent and sharp
Species *Pinus radiata*
(4) Size 6000 cm; Characteristics of the apophysis: prominent
Species *Pinus canariensis*

$$C_i = \left[c_{\min} \left(\sum_{j=1}^{s} f_j \log_2 n_j \right) \Big/ \left(\sum_{j=1}^{s} f_j \right) \right] + V$$

$$V = \left(\frac{1 - \text{Varywt}}{\text{Varywt}} \right) \left(\frac{n+8}{n \log_2 n} \right) \left(\sum_{j=1}^{s} n_j - n \right)$$

$$c = R\text{base}^{5-r}$$

$$f = A\mathit{base}^{a-5}$$

Fig. 21. Division criterion proposed by Dallwitz.

6. The adaptation of the key to reality has been shown to depend to a large extent on the attributes selected for its branching. In view of various equivalent branching alternatives, *XKey* stops its execution and resorts to the user's criterion. This semi-automatic mode of execution can be deactivated in order to operate totally automatically.
7. In addition to the automatic and semi-automatic execution modes, *XKey* has been provided with the capability of generating keys interactively. In this way, the user can select what node to branch and what attribute to use at any given moment; it is also possible to eliminate nodes from the tree, and to change from interactive to automatic mode at any time so as to finish the key generation.
8. In order to help the user, *XKey* presents the available characters in each step, ordered according to the division criterion. This aspect is particularly important as it enables the discrimination capacity of division rules (such as the entropy) to be combined with the human expert's criterion and for the keys obtained to have a biological content which is more satisfactory to the expert.
9. Finally, a comparative study has been carried out of the effect of several division criteria on the generation of dichotomic keys with a sufficiently complex, real taxonomical group: Iberian Gymnosperms. This study reveals that the division criterion of the entropy is the one which produces the shortest keys and which favors the inclusion of a greater number of distinguishing characters. The gain ratio criterion generates somewhat longer keys, but in return, they are more balanced (the length of the paths is less variable). These two criteria detect the most important information for the classification of a set of *taxa*. Dallwitz's criterion does not offer as good results in terms of the average length of the key, power of generating confirmatory characters, etc.

$$CF = \begin{cases} \dfrac{\min\{p(x), p(x \mid a) - p(x)\}}{1 - p(x)} & si \quad p(x) \le p(x \mid a) \\[2ex] \dfrac{\max\{p(x), p(x \mid a) - p(x)\}}{1 - p(x)} & si \quad p(x) > p(x \mid a) \end{cases}$$

Fig. 22. Calculation of the certainty factor of a rule with OR in the consequent.

In spite of this, it can be useful in cases where the objective is not to minimize the length of the key. The same happens with the Gini diversity index, with the observation that this last criterion is not advisable with large data sets since it produces excessively complicated keys.

References

Bartley, M., & Cross, N. (2000). *Navikey 2.0* (En linea, Consulta: 03/06/2003, http://www.huh.harvard.edu/databases/legacy/navikey/index.html).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. California: Wadsworth.

Cabrero-Carnosa, M., Castro-Pereiro, M., Graña-Ramos, M., Hernández-Pereira, E., Moret-Bonillo, V., Martín-Egaña, M., et al. (2003). An intelligent system for the detection and interpretation of sleep apneas. *Expert Systems with Applications*, *24*, 335–349.

CBIT (Centre for Biological Information Technology, University of Queensland) (1994). *LucID version 2.1* (En línea. Consulta: 25/05/2003. http://www.lucidcentral.com/default.htm).

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, *38*(2), 325–399.

Dallwitz, M. J. (1974). A flexible computer program for generating identification keys. *Systematic Zoology*, *1*, 50–57.

Dallwitz, M. J., Paine, T. A., & Zurcher, E. J. (2000). *User's guide to the DELTA system: A general system for processing taxonomic descriptions, 4.12 edition*. Canberra: CSIRO Division of Entomology (En línea, Consulta: 25/05/2003, http://biodiversity.uno.edu/delta/).

Domingos, P. (1998). *Occams's two razors: The sharp and the blunt Proceedings of the fourth international conference of knowledge discovery and data mining (KDD-98), August 27–31, New York City, USA* pp. 34–37.

Domingos, P. (1999). The role of occams's raxor in knowledge disvorvery. *Data Mining and Knowledge Discovery Volume*, *3*, 409–425.

Duncan, T., & Meacham, C. A. (1986). Multiple-entry-keys for the identification of angiosperm families using a microcomputer. *Taxon*, *35*, 492–494.

Fajardo, W., Gibaja, E., Bailon, A., & Moral, P. (2003). An application of expert systems to botanical taxonomy. *Expert Systems with Applications*, *25/3*, 425–430 (Elsevier/2003).

Fortuner, R. (1989). *Nematode identification and expert system technology*. New York: Plenum Publishing Corp..

Giarratano, J. E. (1998). *CLIPS user's guide*. NASA, Artificial Intelligence Section, L.B. Johnson Space Center.

Harrison, P. R., & Kovalchik, J. G. (1998). Expert systems and uncertainty. In J. Liebowitz (Ed.), *Handbook of applied expert systems* (pp. 1–11). West Palm Beach, FL: CRC Press.

Intelsys Inc. (2000–2001). *XID authoring system 3.0 (Demo Version)* (En línea, Consulta: 3/06/2003, http://www.xidservices.com).

López González, G. (2001). *Los árboles y arbustos de la Península Ibérica e islas Baleares Tomo I*. Madrid: Ediciones Multiprensa. 861 pp.

López González, G., & Do Amaral Franco, J. (1986). In S. Castroviejo, et al., *Gymnospermae*. *Flora Ibérica* (Vol. I) (pp. 161–195). Madrid: Real Jardín Botánico, CSIC.

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). NEXUS: An extensible file format for systematic information. *Systematic Biology*, 46.

Mahaman, B. D., Harizanis, P., Filis, I., Antonopoulou, E., Yialouris, C. P., & Sideridis, A. B. (2002). A diagnostic expert system for honeybee pests. *Computers and electronics in agriculture*, *36*, 17–31.

Meacham, C. A. (1986–1996). *Meka version 3.0* (En línea, Consulta: 3/06/2003, http://ucjeps.berkeley.edu/meacham/meka/).

Morales, C., Quesada, C., & Baena, L. (2001) *Guías de la Naturaleza. Árboles y Arbustos*. Diputación de Granada, Granada.

Pankhurst, R. J. (1991). *Practical taxonomic computing*. Cambridge: Cambridge University Press.

Pankhurst, R. J., & Pullan, M. (1994). *Pandora user guide version 3.1* (En línea, Consulta:8/06/2003, http://www.ibiblio.org/pub/academic/biology/ecology + evolution/software/pandora).

Quinlan, J. R. (1986). Induction on decision trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Los Altos, CA: Morgan Kaufmann1-55860-238-0.

Shafer, G. (1976). *Mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Schalk, P. H., & Heijman, R. P. (1996). *ETI's linnaeus II taxonomic software: A new tool for interactive education*, Vol. 3. UniServe.Science News, University of Sydney (En línea, Consulta 7/5/2003, http://www.eti.uva.nl/Home/Articles.html).

Schalk, P. H., & Troost, D. G. (1999). Computer tools for accessing biodiversity information. *Nature and Resources*, *35*(3), 31–38.

Shortlife, E., & Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*, 351–379.

TDWG-SDD (2005). *SDD part 0: Introduction and primer to the SDD standard* (En línea, Consulta 7/02/2005, http://160.45.63.11/Projects/TDWG-SDD/Primer/index.htm).

UBio (2003). *X:ID, version* (En línea, Consulta 30/7/2003, http://ubio.org/offerings/applications/key/index.html).

University of Toronto. Department of Botany; the University of Toronto Libraries; the Royal Ontario Museum (1996). *Pollyclave a multi-entry identification key version 1.6* (En línea, Consulta 6/06/2003, http://prod.library.utoronto.ca/pollyclave/index.html).

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, *8*(3), 338–353.