4.4 IN NEW TOOLS FOR TRAINING AND CAPACITY BUILDING IN TAXONOMY PERSPECTIVES IN DEVELOPMENT AND CONSERVATION

Pierre Grard[1], Juliana Prosperi[1], and Khampeng Homsombath[2]

[1] CIRAD, UMR AMAP, Vientiane, Laos; CIRAD, UMR AMAP, Montpellier, F-34000 France
[2] NAFRI-LARReC, P.O.Box: 9108, Vientiane, Laos

Introduction

Signatory countries of the Convention on Biological Diversity (CBD) have recognized that taxonomists are highly sceptical about its implementation. They have identified the following bottlenecks: i) the number of described species is still low in comparison with the approximate number of existing species, ii) accessing information on species is difficult, due to the fact that the information is not easily available or, contrary to this, is dispersed throughout the internet without hierarchical ordering, iii) identification tools (flora) are old and out-dated or unavailable.

In the particular field of weed control, the identification of species is a major handicap for implementing the necessary measures recommended by the Convention on Biological Diversity, in many parts of the world, notably in South-East Asia. The drastic reduction in the number of taxonomists throughout the world and the irremediable loss of their knowledge have made the task more difficult for agronomists and ecologists in general.

Nevertheless, identification is a basic activity of life, essential for man to identify his surroundings and to classify his tasks. The mechanisms that enable us to establish identify are corollaries of those that help in classification. Indeed, it is not necessary to look at all the specimen of a species to identify one of them. Thus man, like animals to some extent, is capable of distinguishing groups of similar living and non-living things through some common features. We see that herbaria of herbalists existed even in the Middle Ages; these works helped the diffusion of information on medicinal plants in European monasteries. However, identification was possible only through comparison. It was only in the seventeenth century and the beginning of the Linnean era that identification keys were proposed. It is quite probable, at least in botany that a lucid classification was lacking for the preparation of identification keys. In his Flore Francaise (1778) Lamarck stated that keys should follow two basic rules:

1. Identification should be achieved through the most reliable path.
2. This path should also be the shortest one possible.

For a biologist, identification means giving an individual the name of the species to which it belongs. The normal scientific step in all domains of biology is to know the exact identity of the living material used in an experiment, otherwise no proper research can be carried out. Most botanical classifications follow dichotomous keys, which have not evolved much since the seventeenth century. Nevertheless, in the early 60s, i.e., about fifteen years after the first computers were invented, attempts were made to use the mathematical capacity of these machines to conceive of identification methods that would not be based on decision trees alone, but on other approaches which were until then inconceivable due to lack of computing tools.

1 Methods of plant identification

Classifications done by biologists are hierarchic. They rank the taxa (order, family, genus, species) in groups containing taxa with an adequate number of common characters. Thus in the following figure, we have level 1 (family) higher than level 2 with genera, which itself is linked to the next level, i.e., level 3, consisting of species. The classifications can therefore be represented in a form which mathematicians call a tree. Every tree has a root (here family level) and leaves (species level) connected by branches (genus level). Every branching point in the tree's architecture is a node.

Dichotomous keys are the outcome. Although the theory of decision trees is, in principle, very effective (identification of an individual from 65,536 taxa by asking 15 questions), direct applications in the form of identifications are much less effective.

The user may often have incomplete specimens (flowers, fruits, underground parts, etc., may be missing). However, some keys help identification from vegetative parts alone, but they do not tolerate errors in interpretation by the identifier.

On a mathematical scale, when we consider a key to identify T species, if two answers are possible for every question, we have:

$$Q = T - 1$$

Nevertheless, the Q questions, which will be the nodes of the tree, can be organized at different levels ranging

from

- a maximum number of levels that is equal to the number of questions: it is then an unbalanced key

to

- a minimum number of levels with a value of log2(T) – 1. This key is perfectly optimized because the number of questions asked to arrive at any taxon level will be minimal and always the same.

However, there are some risks of committing identification errors when using this organization of keys. If, due to intraspecific variation or observational error, the frequency of error for every question is f = 1/10, the success rate will be f' = 9/10. If this key is capable of identifying 256 species, we will have:

$$T = 256.$$

The number of levels in the key will be N:

$$N = \log2(T) - 1 = \log2(256) - 1 = \log2(28) - 1 = 7$$

If the dichotomous key is balanced, the success rate for the identification of each species will be:

$$fT = (9\text{-}10)7 = (0.9)7 = 0.47$$

In other words, in the case of a perfectly optimized key that enables the identification of only 256 species, the identifier will commit identification errors in more than half the cases.

Another way to identify species is to make identification of species by matching. A simple similarity coefficient can be defined as below. In a computerized form of comparison between a specimen and the taxa, the following formula can be used to calculate the agreement for each character:

$$\frac{\text{No. of common shared character states in both}}{\text{Total number of observed states in both}}$$

In this method, weights can be attached to the characters and even to the character states (Grard 1996), taking into account the confidence that can be given to them on the basis of intraspecific variation or if the character state is easy to observe. This method was used for the identification of weeds of Réunion Island (Le Bourgeois et al. 2001), for the identification of weeds of Indo Gangetic Plains and as been applied to the Oswald () implementation.

2 Application to weed species

The objective of this work has been to provide to scientists working on biodiversity-related subjects with a software to help identify these species in spite of missing information and, within certain limits, errors, and which can provide information of the users' choice. This is done entirely graphically by constituting an identikit picture. After identification, the user can access a descriptive file of the species where botanical terms are defined in a hypertext manner. The information available on the CD-ROM will be printed in:

- a «paper» flora in the form of a field book making all the information available on the CD-ROM accessible.
- an updateable web site giving information on these species (www.oswaldasia.org).

The software, in this version, is able to identify about 113 species of weeds of paddy fields of Cambodia and Laos. Besides botanical, ecological it includes information on their control. Devoted to non-specialists, it is useful in capacity building for training and self-training of young agronomists.
This work is a collaborative project of the European Asia IT&C program: "Open Source for Weeds Assessment in Lowland Paddy fields".

The Identikit:
The identikit has been developed with vectorized drawings. On this version on weeds species, the software needs to use 19 different layers, which compound the picture of the species that the user is trying to identify.

All these graphics were then converted to the Windows Metafile format which are directly readable by Windows. Access to the different files, which constitute the identikit is managed by MS-Access and data are formatted following the rules defined by IDAO System (Grard 1996).

In this application, 138 character states belonging to 18 characters are used to identify a flora of 113 species (see list of species). A weight ranging from 1 to 9 is given to each of the characters, depending on the easily observable nature of the character, its variability and the « confidence » that can be given to it. Thus, the phyllotaxy will have a greater weight than the hairiness of the leaves or the colour of the flowers. What is important while attributing weight to a character is the note given to each character state in a relative comparison to the

others.
The second database shows the hypertext mark-up language (HTML) pages after a species is identified. This database contains a table listing the species that can be identified, as well as all the information on these plants to assemble their description. For each species, the number of photos and their titles appear on the respective HTML pages, and then the descriptive text of each species and the legends of the drawing, if included. All the digitized and processed photographs are then listed in a folder with this database.

The identification process:
This software acts through a graphic interface with an identikit of the species to be identified as the main support. This identikit contains 19 elements (one for each character), which are modified during the search:

The choice of characters can be identified when the cursor is moved, activating the sensitive zones: on certain colours, or on a frame on the identikit.

Once the selection is made, one simply clicks on the character state that closely resembles the species to be identified. The identikit page responds to the selected character in a newly created cell.

Thus the identikit refines as and when the descriptions of character states are seen in the species while being identified.

The number of species closely matching with the identikit appears on the lower right-hand corner of the screen, along with a percentage of confidence index.
When a certain number of characters is collected, it is possible to obtain the results of a search, i.e., the list of species that closely resemble the description obtained, by clicking on the "Results" button. In fact, it is not necessary to know all the characters to get the result. This list is classified in descending order of the confidence percentage of the species.

When a species is selected, the page with its description and illustration appears; the botanical drawings can be enlarged many times by using the zoom and clicking on the illustration.

It is possible to obtain definitions of botanical terms in the descriptions through hypertext links.

Identification can be done in a second way: any botanist knows that some character state is "strong". For instance, in the case of weeds, it is easy to imagine that there are not so many different species having a spiny attachment on the leaves. The selection of such a character state is strongly discriminant.

Conclusion:

This computer aided identification software is based on recognizing mainly vegetative characters. This capacity helps identify trees in any season, without bothering about the phenology of the species. Moreover, identification is made easy because the characters to be identified are more easily visible than floral structures. This graphic interface also enables people who are not specialists in botany, and who do not know its descriptive terms and the logic of identifying a plant, to use this software. For identification, it is enough to simply compare the character states of the plant to be identified with the different character states proposed. Moreover, the possibility of getting a botanical description, good photographic illustrations and accurate botanical drawings enables the user to visually confirm his identification and be assured of his result.

The user need not define all the descriptive characters to obtain results. Hence he does not stop if no response to his question is received, as is the case when using traditional floras.

Finally, after confirming the identification of a species, it is possible to find errors in the choice of characters, if there were any during identification, through the photographs. For this, it is enough to use the « contradiction » button after identifying and selecting a species. This will appear on the results page only if the species identified does not have a 100% identification index. When it is activated, it shows the characters selected by the user on the identikit that are not the characters of the species identified. These non-conformable characters are marked with a red cross in the cell showing the character.

All these capacities are end-user oriented and have been endorsed by the global taxonomy initiative (GTI) of the Convention on Biological Diversity (CBD) as following the rules defined for training and self-training of non-specialists, for information dissemination and capacity building. The next step will be to propose the extension of this software to other paddy fields of the world and to apply this work to other species identification sets.

Bibliography

Hansen B. and Rahn K., 1969. «Determination of angiosperm families by means of a punched card system», Dansk Botanisk Arkiv, 26.

Le Bourgeois, T., Jeuffrault, E. and Fabrigoule, S., 2000. «Advenrun, Principales mauvaises herbes de la Réunion», Les Editions de Canne Progrès, 125 pp.

Saldanha, J. and Rao, K., 1975. «A punched card key to the dicot families of South India», Centre for Taxonomic Studies, St. Joseph College, Bangalore.

Whalley, P.E.S., 1976. Tropical leaf moths. British Museum (Natural History), London.