Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 107-112.
ISBN 978-88-8303-295-0. EUT, 2010.

MyKey: a server-side software to create customized decision trees

David Gérard, Régine Vignes Lebbe

Abstract — To facilitate the identification of specimens, biodiversity informatics has developed numerous new computer-aided tools which compete with the old, printed single access keys. Free access keys are accessible for many different taxonomic groups, but key-generating software are also helpful to construct single access keys. This paper presents a midfield solution to create customized decision trees (single access keys) through an interactive web-based interface. This solution offers an online service to create keys according to the parameters and context chosen by the final users themselves. It is also useful for the administrator of this online system, because of low-maintenance needs, limited to configuration files when new knowledge bases are added on the server side. Presently, the software is available with a French interface at the following URL: http://baron.snv.jussieu.fr/cgi-bin/david/MyKey.cgi.

Index Terms — polytomous key, customized key, decision tree, web-based interface.

1 Introduction

Accessing relevant and critical taxonomic information is often a privilege for the specialists [1], who can take profit of natural history collections, taxonomic monographs or low-circulation journals. Lawyers, border guards, epidemiologists, as well as ecologists or any other biologists, may also have identification requirements. For their needs, printed dichotomous or polytomous keys are often included in monographs, floras and faunas, and in practical field guides. A key has a graph structure, comparable to a decision tree of Artificial Intelligence [2] (in this paper we will use as synonyms the terms single access key, and decision tree).

A negative property of classical keys is their static nature: if you cannot answer to a question (for example if you have no flower and characters of the flowers are frequently used in botanical keys), the key is useless. Moreover, to create a

D. Gérard was student in the Laboratoire Informatique et Systématique, University of Paris 6, UMR7207 (MNHN, CNRS, UPMC), CP 48, 57 rue Cuvier 75005 Paris, France. E-mail: dagerard@gmail.com.
R. Vignes Lebbe is with the University of Paris 6, UMR7207 (MNHN, CNRS, UPMC), CP 48, 57 rue Cuvier 75005 Paris, France. E-mail: regine.vignes_lebbe@upmc.fr.

key is a time-consuming task and each taxonomist adapts his/her key to a given context. But it could be necessary to offer different keys to different user groups (e.g. an autumn key to trees based only on trunk and leaves characters, a key based on fruits, a key limited to a geographic area, a key taking into account immature stages, etc.).

In the late 1960s, biologists [3], [4] began to use computers to produce more flexible free access keys or computer-aided-identification (CAI) systems [5]. Since the 1980s, knowledge bases formats for structuring descriptive data appeared, like the DELTA format [6], and user-friendly software (e.g. IntKey [7] and XPER [8]) were implemented for creating knowledge bases and enabling CAI. Storage of data became easier, as well as the retrieving of specific information in a pool of data [9]. The reader can find a good report comparing these tools in [10], [11], on the DELTA website, and on the BD tracker of the European project EDIT [http://www.e-taxonomy.eu/].-

Causse & Lebbe [12] have demonstrated the strong similarity between CAI and single access keys, and their common elimination procedure. These authors introduce the idea of a unique system able to propose identification from a free access key to a single access key by continuously improving the strategy advertisement expressed by the taxonomist.

To adapt single access keys to the users, one finds different proposals (see for example [18] and [19] in this book). This paper offers another solution: it combines a program to compute automatically single access keys and a web interface for the final user to define himself the input parameters of the key constructor. This original prototype, MyKey, is a server-based program. It uses knowledge bases stored with the XPER system and the key generator MAKEY [13]. Running on a server of the Laboratoire d'Informatique et Systématique of the University Paris 6, it is available at the following URL (http://baron.snv.jussieu.fr/cgi-bin/david/MyKey.cgi).

2 The tools XPER and MAKEY

XPER and its current version Xper² 2.1 is a complete software package for managing knowledge bases [14]. It provides tools for structuring and using taxonomic descriptions and for identifying specimens. The basic program allows to save structured descriptions, comparable with the DELTA format and consisting of three main elements: taxa, descriptors (characters for DELTA) and states of descriptors. An import/export in SDD XML schema the new standard proposition of TDWG, is also available (SDD= Structured Descriptive Data. See : http://wiki.tdwg.org/SDD).

MAKEY [13], [15] is a key generator software. It selects step by step the best question (descriptor) to create a node. By default, the choice is based on the discriminating power (ability to split the pool of taxa in equal disjointed classes, two or more classes) and so MAKEY tries to generate short and well-balanced polytomous keys [16].

MAKEY creates keys discriminating the different taxa of the input knowledge base. Other versions create a key to discriminate groups or classes of the taxa; these classes are sets of taxa defined by different character states of a selected

descriptor. For example, if a descriptor is the toxicity of mushrooms, MAKEY can create a key to identify the toxicity of a mushroom even if the specimen is not identified at the species level; in the same manner it is possible to create keys to identify genera within a knowledge base describing species if a descriptor associates each species to its genus. The manual to use MAKEY is accessible on line.

3 Description of MyKey

The final user of a key is the best person to know his observation constraints. So, the concept of the MyKey service was to offer to the final user the possibility to create identification keys customized to his needs.

A web interface gives an access to the different input parameters of the Makey software. We classify these parameters in four categories:

- parameters are related to the data coverage of the key; for example one can generate a key to all species of the knowledge base or a key restricted to a given geographical area or to a genus etc.
- parameters are related to the taxonomic domain (importance and easiness to observe a character),
- parameters have consequences on the topology of the key, like the criterion to select characters (by default it minimizes the mean number of questions to achieve an identification),
- parameters concern the format of the result: indentented or bracketed key (see [21] in this book), text or HTML format etc.

So the interface is divided in four parts according to these categories.

3.1 Goal or terminal nodes of the key

The user selects the goal of the discrimination, it means the terminal nodes of the key. The key can identify all or just a set of taxa, or any group of taxa defined by character state. So, considering a knowledge base describing species and a character “genus”, the user can then create a key discriminating the different genera and then keys to recognize species within each genus. In the same manner we can compute a key to identify the toxicity of mushrooms and not the species themselves.

Considering a knowledge base that covers a world distributed taxon, the user may need to consider only a subarea (hereafter called “sub-base”). The sub-base will then only include the taxa specified by the user. If the user can fill in a background (a specific region or country, a maximal bathymetric range etc.), a sub-base will be extracted, excluding taxa not compatible with the given conditions. The decision tree generated by MAKEY is then shorter than the key including all taxa, and so it minimizes the probability of error. Indeed, if two taxa are quite similar but are not present in the same altitude/country, using a key built on a sub-base reduces the risk of misidentification.

3.2 Background knowledge

Weights (or ponderation values), one for each descriptor, define a pre-order on the descriptor set (by default an equal weight is associated to all the descriptors or characters). MAKEY will respect this pre-order to select the character at each node. The characters can be ordered by the final user himself to force their choice in the key. So if flowers are absent the weight of all the flower's characters can be minimized or put to zero. At the contrary if some characters are easy to observe for the user he can associate to these characters a higher weight.

3.3 Topology of the identification graph

The topology section lets the user to choose some criteria to be used during the key construction (minimal number of branches at each node; to merge branches; to eliminate first some taxa etc.). Some statistics measurements help to compare the topology of the keys with different parameters and to choose the best decision tree.

3.4 Output format

The user can define the parameters to display the key: nested key (also called "yoked" or "indented") or parallel key (also called "bracketed" or "linked" key). Additional characters and states may be added if they are deduced at a step of the key.
The generated key is available in HTML format (including an option for a special layout for handheld devices) or in PDF for printing.

4 Architecture

Mykey is a server-side software implemented as a CGI script written in PYTHON; the system is easy to maintain and to upgrade, and it is compatible with any operating system.

According to the user selected parameters, (a) Mykey extracts a sub-base if necessary, (b) Mykey creates or modifies the file of character weights, (c) Mykey calls the software MAKEY which is then executed on the server with the selected parameters and (d) Mykey formats the MAKEY output and the result is sent to the client browser in the selected design. The key can also be saved on the server (in fact only the parameters will be saved), to restore it when needed, to modify it or to share it with other users.

5 Conclusion

Mykey is a running prototype. It is an efficient additional system to Xper^2, a midfield solution between single access key and free access key. Today a depository for Xper^2 knowledge bases is accessible at http://lis-upmc.snv.jussieu.fr/xper2/infosXper2Bases/en/index.php to any user. Then the data

can be accessible with Mykey. An option modifies the display for output on a personal pocket palm. Few similar options were encountered (URL: http://www.phylodiversity.net/palmkey/), and the one proposed by MyKey is perfectible.

Mykey is not a website to access to keys but an online service to produce keys [17]. In the European project EDIT the functions to create keys were implemented in the CDM library (see [20] in this book).

Mykey has to be modified to become a web service able to be connected easily to other softwares. In the future ViBRANT project (Virtual Biodiversity Research and Access Network for Taxonomy http://vbrant.eu) such identification system (free access and single access key construction) will be available as a web service and will allow a more open and flexible use.

Acknowledgement

The authors wish to thank Amandine Sahl for her contribution to this work during her master PhD, and all the users of this prototype.

References

[1] J. D. Agosti, “Biodiversity data are out of local taxonomists’ reach”. *Nature*, p. 392, 2006.
[2] J. R. Quinlan, “Induction of decision trees”. *Machine learning,* vol. 1, pp. 81-106, 1986.
[3] D. W. Goodall, “Identification by computer”. *Bioscience,* vol. 18(6), pp. 485-488, 1968.
[4] R. J. Pankhurst, “Identification methods and the quality of taxonomic descriptions”. In: *Biological identification with computers*. Academic Press, London, 1975.
[5] P. M. Forget, J. Lebbe, H. Puig, R. Vignes and M. Hideux, “Microcomputer-aided identification / an application to trees from french Guiana”. *Bot. J. Linn. Soc*., vol. 93, pp. 205-223, 1986.
[6] M.J. Dallwitz, Overview of the DELTA System, 2009. http://delta-intkey.com/www/overview.htm, June 2010.
[7] M. J. Dallwitz, T. A. Paine and E.J. Zurcher, *User’s Guide to Intkey: a Program for Interactive Identification and Information Retrieval,* vol. 1, 1995.
[8] J. Lebbe R. Vignes and J.P. Dedet, “Computer-aided identification of insect vectors”. *Parasitology Today,* vol. 5 (9), pp. 301-304, 1989.
[9] A. R. Brach and H. Song, “eFloras: New directions for on-line floras exemplified by the Flora of China Project”. *Taxon*, vol. 55 (1), pp. 188-192, 2006.
[10] R. J. Pankhurst, *Practical Taxonomic Computing*. Cambridge Univ. Press, Cambridge, 1991.
[11] J. Lebbe and R. Vignes, “State of the art in computer-aided identification in biology”. *Oceanis*, vol. 24(4), pp. 305-317, 1998.
[12] K. Causse and J. Lebbe, “Modélisation des stratégies d’identification par la méthode MCC”. *JAVA-95,* (Conference proceedings), 1995.
[13] J. Lebbe and R. Vignes, “Génération de graphes d’identification à partir de description de concepts”. In: Y. Kodratoff and E. Diday (eds.), *Induction Symbolique et numérique à partir de données*, Cepadues, pp. 193-239, 1991.
[14] V. Ung, G. Dubus, R. Zaragüeta-Bagils and R. Vignes Lebbe, Xper²: introducing e-Taxonomy. *Bioinformatics,* vol. 26(5), pp.703-704, 2010.
[15] R. Vignes, *Caractérisation automatique de groupes biologiques*. Université Pierre et Marie Curie, 260 pp. (Thesis), 1991.
[16] J.C. Gower and R.W. Payne, “A comparison of different criteria for selecting binary tests in diagnostic keys”. *Biometrika,* vol. 62, pp. 665-672, 1975.
[17] N. Conruyt, D. Sébastien, S. Cosadia, R. Vignes Lebbe and Touraïvane, “Moving from biodiversity information systems to biodiversity information services”. In: L. Maurer, K. Tochtermann (eds.), *Information and Communication Technologies for Biodiversity Conservation and Agriculture*, Shaker, Aachen, (ISBN: 978-3-8322-8459-6), 2009.
[18] J. Nascimbene, S. Martellos and P. L. Nimis, “An integrated system for automatically producing

user-specific keys - A case study on Italian lichens". In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 151-156, 2010.
[19] E. van Spronsen, S. Martellos, D. Seijts, P. Schalk, and P. L. Nimis, "Modifiable digital identification keys". In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 127-131, 2010.
[20] W. G. Berendsohn, "Devising the EDIT Platform for Cybertaxonomy". In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 1-6, 2010.
[21] G. Hagedorn, G. Rambold and S. Martellos, "Types of identification keys". In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 59-64, 2010.