Tạp chí Khoa học và Công nghệ 53 (2) (2015) 127-138
DOI: 10.15625/0866-708X/53/2/3684

BIOKEYS – AN INTEGRATED SYSTEM FOR WORKING WITH DATABASE AND POLYCLAVE IDENTIFICATION KEYS OF VARIOUS TAXONOMIC LEVELS

Nguyen Van Sinh

Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Nghia Do, Cau Giay, Hanoi, Vietnam

Email: vansinh.nguyen@iebr.ac.vn

Received: 21 February 2014; Accepted for publication: 4 November 2014

ABSTRACT

Unlike the other currently available identification software, BioKeys allows the users to create and use own polyclave identification keys of various taxonomic levels. The information of dedicated taxonomic level is stored in the head line of the key file so that BioKeys can understand and search in the right field of the database for the records of matched taxa. The results of database referencing or specimen identification will be displayed in a form of web page. Besides, utilities are available for creating and managing the database, for analyzing the polyclave identification key. An example of using BioKeys is provided with polyclave key of plant families of Magnoliophyta that has been created based on the punched-card system of Betel Hansen and Knud Rahn (1969).

Keywords: polyclave key, identification, database, key analysis.

1. INTRODUCTION

Currently available identification systems allow the user only to use integrated keys to identify organisms at one taxonomic level [1, 2] or to create own keys but without automatic recognition of taxonomic level by the software (Ch. Oliver Coleman, James K. Lowry, Terry Macfarlane 2010 [3]). BioKeys is an integrated tool that allows the user to create and manage biological database, and to use and create own polyclave identification keys of various taxonomic levels. The result of an identification process is illustrated by displaying the records of matched taxa from the database in the form of a web page. BioKeys provides also several statistics of the database and of the polyclave identification key.

2. MATERIAL AND METHODS

Delphi XE Professional Workstation ESD (item number: 2010111885211109) of the Embarcadero Technologies company [4] has been used to create BioKeys. The Installation

package of BioKeys is available at the web site of the Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology [5].

3. DESCRIPTION OF BIOKEYS

3.1. BioKeys User Interface

BioKeys main window (Fig. 1) is a multiple document interface that can display an image component, a web browser component or a key file child window. BioKeys main window has a toolbar and a menu bar. Image component is dedicated for browsing database records. Web browser displays the results of database referencing or of specimen identification in a form of webpage that allows to use the links to open pictures or informtaion of the taxa. Key file child window serves as device for editing polyclave identification key. It can also display the BioKeys simple stream files that contain results of the key analysis.

Dialog boxes of BioKeys are the utilities for interactive working. The 'Reference /Specimen Identification' dialog box (Fig. 2) is dedicated for selecting characters in case of database referencing or specimen identification.

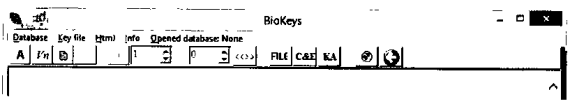

Figure 1. BioKeys main window.

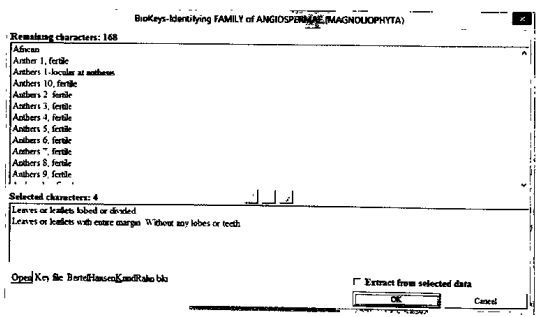

Figure 2. Dialog box for database referencing or specimen identification.

'Create/Edit key file' (Fig. 3) is the dialog box for creating template of a new polyclave identification key file or editing an existing one. The 'Key analysis' dialog box (Fig. 4) is

dedicated for key analysis. In this dialog box we can let BioKeys calculate statistics to answer several questions: What is the size of the key (total number of characters)? How many characters each taxon owns in this key? In how many taxa each character is available? For a certain size of character set: How many total possible successful character sets and how many possible successful character sets of each taxon or character.

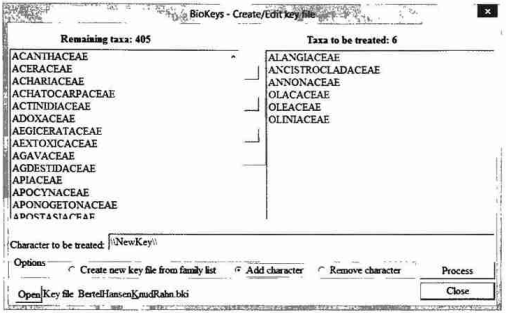

Figure 3 Dialog box for creating and editing key file.

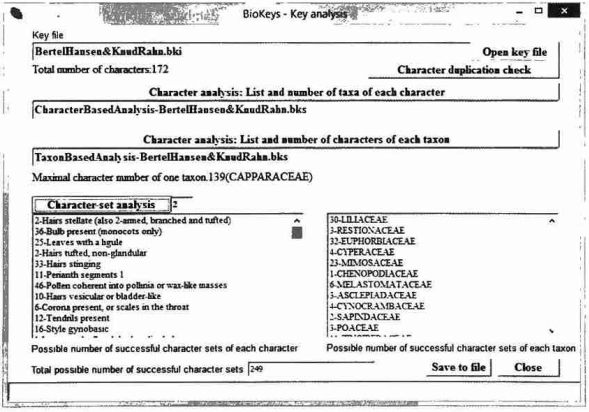

Figure 4. Dialog box for key analysis.

Figure 5. Declared own file formats of BioKey.

3.2. BioKeys' Declared File Formats

BioKeys declared three own file formats (Fig. 5):

Database file (the file extention is '*.bkd').

Polyclave identification key file (the file extention is '*.bki').

Simple stream file (the file extention is '*.bks').

Because being declared, once the BioKeys has been installed with installation package, double clicking on these files will cause BioKeys to be started with opening the clicked file.

3.3. Create and Manage Database

Database in BioKeys is image oriented. One species can be represented in the database by one or more images. The database file consists of the records that contain information of images with following fields: 1) Division/Phylum, 2) Class, 3) Order, 4) Family, 5) Genus, 6) Species, 7) Name of image file.

To create a new database, we click on the "Create new database" submenu and set the name of the automatically created one record template in the appeared save dialogbox to save it. After that we can open the newly created database and edit, add, delete records as needed. To view the opened database in web page form (Fig. 6), we click on the 'Create and show html' submenu. If we have text files with information on taxa (the file must be named after scientific name of the corresponding taxon with '_E' suffix, e.g., 'scientific name_E.txt' file for the 'species scientific name' taxon), we can click on the 'Info' link beside this taxon name in the web page to load and view this file.

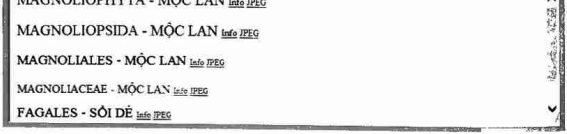

Figure 6. Database view in web page form.

When we add a record to the database, the picture will be copied to the folder of BioKeys with original file name. We can change the various file names of the pictures to standard file names of BioKeys by clicking on the submenu item 'Change name of picture file'. After changing, the name of a picture file of a database record will have following formula:

's' + record number + '.jpg'

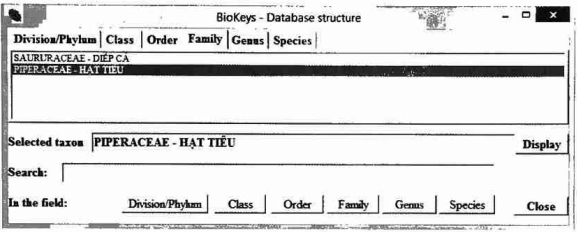

Figure 7. Displaying the database structure in tab view mode.

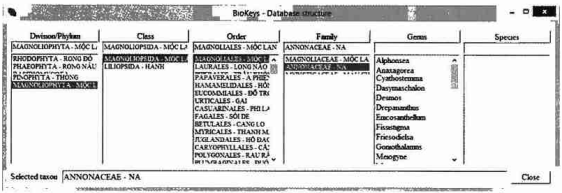

Figure 8 Displaying the database structure in column view mode.

To explore a database, we need just to click on the "Display database structure" menu. In the appeared dialogbox we can choose to display the database structure in tab view mode (Fig. 7) or in column view mode (Fig. 8). In these views, the next lower ordered taxa will be displayed when any taxon is clicked. Once a taxon is clicked, we can display all the lower ordered taxa in a web page form by clicking on the "Display" button (in tab view mode) or on the button above the recently chosen taxon (in column view mode).

During exploring the database, when we came to the species list of a genus, we can change to the picture browsing mode by double clicking on the name of the species. In this mode, we can use the 'Change picture of species' button to view all the pictures of the species. During viewing a picture, we can use the 'Change picture size' up-down bar in order to change the size of the picture, and we can keep the mouse down on the picture and move it in order to move the picture.

To browse the records of the database, we use the 'Switch between web/database browsing' button to change to the database browsing mode. In this mode, we use the 'Browse database' button to go through database. When the record number is changed, BioKeys loads the picture of the current record and displays it on the main window, the taxonomic information of the will be displayed in title bar. Here we can also change the size of the picture and move it around the screen.

BioKeys allows user to search in the fields of the database. In the tab view, we type the search string in the edit field after label 'Search', then click on a button in the line below that indicates a taxon level field of the database, where the searching should be conducted. In the column view, we type the search string directly into the edit field under the button with name of a taxon level field where the searching should be conducted. After searching, BioKeys displays the results in a web page form.

Finally, BioKeys can calculate and give statistics on taxa number of each taxonomic level of the database or of the selected records (Fig. 9). To get these statistics we click on the 'The whole database' or 'Selected items' submenu.

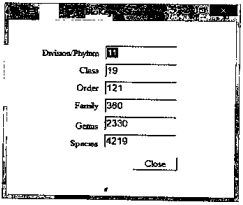

Figure 9. Statistics of the database or of the selected records.

3.4. Working with Polyclave Identification Key

3.4.1. Format of polyclave identification key file

BioKeys supports working with polyclave identification keys. A polyclave identification key file consist of a head line and key content. The head line has the following format:

[WORD FOR TAXONOMIC LEVEL] OF [higher ordered taxon]

The word in the first square brackets must be capital and can be one of the following: DIVISION or PHYLUM, CLASS, ORDER, FAMILY, GENUS, SPECIES. For example, the head line of a polyclave identification key for identifying family of specimen of the Magnoliophyta plant division will be:

[FAMILY] OF [Magnoliophyta]

Thus the content of the first square brackets is the name of the taxonomic level, the taxa of which should be identified by this key. The content of the second square brackets describes the higher ordered taxon for which the current polyclave identifiction key is dedicated.

Each line of the key content consists of the taxon name and the characteristics of taxon (character list) and has following format:

TaxonName=\\Character1\\Character2\\....\\CharacterN\\

Thus, each character is bound by two double backslash from the sides. In the case of a polyclave identification key for identifying family of specimen of the Magnoliophyta plant division, the TaxonName will be a name of plant family of this plant division: Magnoliace, Annonaceae,... and hence one line of the key content will be of following form:

Magnoliaceae=\\Character1\\Character2\\...

3.4.2. Create and edit polyclave identification key file

The first way to create and edit polyclave identification key file is to use the key file child window (Fig. 10). To open the key file child window we click on the 'Create key file child window' submenu item under the 'Key file' menu item. In the appeared child window we just type in a new key or open a saved key file to edit.

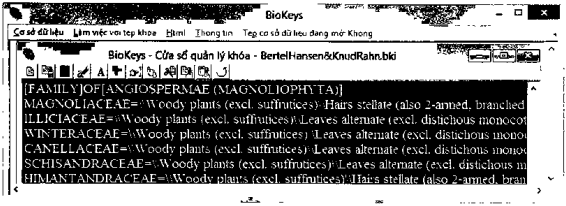

Figure 10. Key file child window.

The other way for creating and editing polyclave identification key file is to use the 'Create/Edit key file' dialog box. To open this dialog box we click on the 'Create/Edit key file' submenu item under the menu 'Key file' item. In the appeared dialog box, after we open the saved polyclave identification key file, the taxa from the key will be listed in the left list box. To create a new, empty polyclave identification key file with the same taxa as these of the opened key, we choose the 'Create new key file from family list' item in the options box, then click on the 'Process' button. To add or remove a character from character list of any taxa of the opened key, we move these taxa from the left list box to the right list box (Taxa to be treated) by marking them and clicking on the appropriate arrow. Then we type the character in the 'Character to be treated' edit field, choose the appropriate option ('Add character' or 'Remove character'), and click on the 'Process' button.

3.4.3. Analysis of a polyclave identification key

The interface for analysis of a polyclave identification key that is saved in a file is the 'Key analysis' dialog box. To open this dialog box we click on the 'Key analysis' submenu. In the appeared dialog box we click on the 'Open key file' button to open a polyclave identification key, before any analysis can be conducted. Results of analysis will be saved in files of 'BioKeys simple stream file' format with '*.bks' extention. After analysis we can use the key file child window to open the file with analysis results.

To check, if any character is duplicated in the character list of one taxon, we click on the 'Character duplication check' button. BioKeys will check the character list of all the taxa of the key. If a character is represented more than once, BioKeys will make needed correction.

To create a file with information on how many and which taxa have a defined character (Fig. 11), we click on the 'Character analysis: List and number of taxa of each character' button. A save dialog box will appear with an automatic provided file name that consist of the 'CharacterBasedAnalysis-' prefix and the name of the opened key file. We just need to click on save button to save the results.

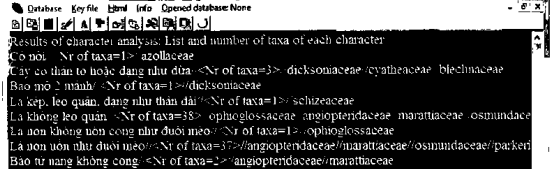

Figure 11. Results of the character based analysis.

To create a file with information on how many and what characters each taxon of the key has, we click on the 'Character analysis: List and number of characters of each taxon' button. A save dialog box will appear with an automatic provided file name that consists of the 'TaxonBasedAnalysis-' prefix and the name of the opened key file. We just need to click on save button to save the results (Fig. 12).

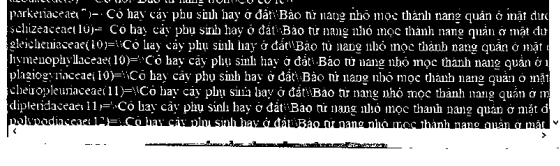

Figure 12. Results of the taxon based analysis.

BioKeys allows us also to analyse the key in different aspects of identification capability – called Character-Set-Analysis. For this analysis we set at first a size of character set by entering a number in the edit field (next to the 'Character-set analysis' button on the 'Key analysis'

dialog box. Then we click on the 'Character-set analysis' to begin the analysis. A save dialog box will appear with an automatic provided file name that consist of the 'CharacterSetBasedAnalysis-' prefix and the name of the opened key file. We just need to click on save button to accept the file name. During analysis, BioKeys will inform us through message boxes on the total character number of the key, the maximal possible number of character sets (calculated based on total character number of the key and the size of character set), the number of successful sets of each character (based on the trial identification with using all possible character sets), the number of successful sets of each taxon (also based on the trial identification with using all possible character sets). After the analysis has been completed, we can open the file to see the results (Figs. 13-14).

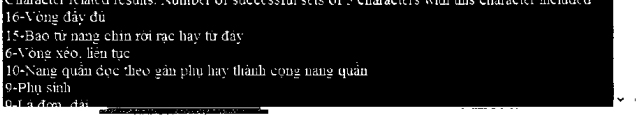

Figure 13. Character set based analysis – Character related results

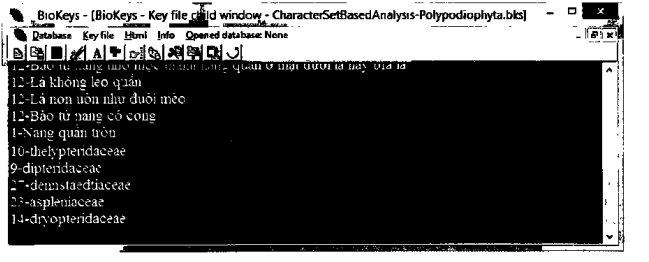

Figure 14. Character set based analysis – Taxon related results.

3.4.4. Reference/Specimen identification in BioKeys

Because after the identification process BioKeys will search all matched records in the database and display them on the main window in the form of a web page, before we begin with identification, we need to open the database. After clicking on the 'Reference/Specimen

identification' submenu under the 'Key file' menu, a dialog box will appear. To open a polyclave identification key file, just click on 'Open' button on this dialog box. After the key has been opened, BioKeys extracts all the characters of the key and list them in the first list box.

To reference on what taxa have some defined characters, we choose these characters from the first list box (of source characters) and move them to the second list box (of selected characters) by using arrow keys, then click on the 'OK' button, the BioKeys will identify what taxa have all these characters then search the records of these taxa in the database and display them in a form of web page. BioKeys will also list the taxa that are not represented in the database by any records at the end of the resulted web page.

To identify a specimen, we read the characters of the specimen and choose these, if available, in the list box of source characters and move them to the list box of selected characters. Then we click on the 'OK' button and the same procedure will be done as in the case above of referencing database.

4. AN EXAMPLE OF USING BIOKEYS

4.1. The Polyclave Identification Key for the Families of Magnoliophyta Plant Division

The polyclave identification key for the families of Magnoliophyta plant division is built based on the punched-card system of Hansen and Rahn (1969) [6]. There are 411 families from Hutchinson (1959), Engler Syllabus (1964), Bentham & Hooker (1867-1880), and Engler Pflanzenfamilien ed. 1. The polyclave key therfore consists of a headline and the key content that consists of 411 lines (Bertel Hansen and Knud Rahn 1969). The key is saved to the file named 'BertelHansen&KnudRahn.bki'.

4.2. Reference/Specimen Identification

To use these utilities, we first open the plant database ('367.bkd'). Then we open the 'Reference/Specimen identification' dialog box by using the submenu of the same name under the 'Key file' menu. In this dialog box, we open the key file named 'BertelHansen&KnudRahn.bki'. After opening the file, BioKeys will list all the characters (172 characters) in the upper list box of the dialog box.

To reference, for example, which families from Magnoliophyta plant division have contorted petals, we select this character from the upper list box and move it to the lower one by using arrow key. After clicking on the 'OK' button, BioKeys will define the taxa that have this character by using the opened key file (matched taxa). Then BioKeys will search in the opened database for the records of the matched taxa, create for them the '_BioKeys.htm' file and open this file in the integrated web browser. A list of 27 families that are not represented in this database will also be attached.

4.3. Key Analysis

To analyze the polyclave identification key for the families of the Magnoliophyta plant division, we open the 'Key analysis' dialog box (Fig. 4). In the appeared dialog box we open the key from the 'BertelHansen&KnudRahn.bki' file. There are 172 characters used in the key. The results of taxon based analysis has shown that character lists of the Capparaceae and

Euphorbaceae plant families are largest and consist of 139 characters, while the character list of the Hydrostachyaceae plant family is smallest and consists of only 32 characters. The character based analysis has shown that there are 3 characters that are most rare characters and are presented each only in 4 taxa. These characters are 'Pollen coherent into pollinia or wax-like masses', 'Hairs stinging', and 'Bulb present (monocots only)'. The most common character is 'Filaments not connate' and is presented in 395 of 411 families.

Identification is based on the character sets. The bigger character set is, the larger the number of maximal possible character sets becomes. When the size of character set is 2 characters, the number of maximal possible character sets is 14706, and the possible successful character sets is 251. When the size of character set is 3 characters, the number of maximal possible character sets became 833340, and the possible successful character sets is 1554. But when the size of character set is 7 characters, the number of maximal possible character sets became 780842580024, and the possible successful character sets became 2487.

5. CONCLUSIONS

BioKeys is a fully functioned tool that allows the users to create and manage database and polyclave key, to reference database and to identify specimen. The current version of BioKeys can provide certain statistics of a database and polyclave key. In the future, it is desired to develop criteria or indices for quantitative evaluation of polyclave keys.

Acknowledgement. BioKeys has been developed within the research project (VAST04.08/13-14) funded by Vietnam Academy of Science and Technology.

REFERENCES

1. http://www.colby.edu/info.tech/BI211 /PlantFamilyID.html, last accessed February 20, 2014.
2. http://www.stingersplace.com/SLIKS/, last accessed February 20, 2014.
3. Coleman O. Ch., Lowry J. K., Macfarlane T. DELTA for Beginners: An introduction into the taxonomy software package DELTA. ZooKeys 45: 1–75.
4. http://www.embarcadero.com/, last accessed February 20, 2014.
5. http://www.iebr.ac.vn, last accessed February 20, 2014.
6. Hansen B., Rahn K. Determination of Angiosperm Families by Means of a Punched-Card System. Dansk Botanisk Arkiv 26. Kobenhavn, 1969.

TÓM TẮT

BIOKEYS - MỘT HỆ THỐNG LIÊN KẾT CHO PHÉP LÀM VIỆC VỚI CƠ SỞ DỮ LIỆU VÀ KHÓA ĐỊNH LOẠI ĐA TRUY Ở CÁC BẬC PHÂN LOẠI KHÁC NHAU

Nguyễn Văn Sinh

Viện Sinh thái & Tài nguyên sinh học, Viện HLKHCNVN,
18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội

Email: vansinh.nguyen@iebr.ac.vn

Khác với các phần mềm phân loại hiện có, BIOKEYS cho phép người dùng tạo và sử dụng khóa định loại đa truy của riêng mình ở các bậc phân loại khác nhau. Thông tin về bậc phân loại của khóa được lưu tại dòng đầu của tệp khóa nên BIOKEYS có thể hiểu và tìm ở đúng trường cần thiết của cơ sở dữ liệu để lọc các bản ghi của các taxon phù hợp. Kết quả tra cứu cơ sở dữ liệu hoặc định loại được hiển thị dưới dạng web. Ngoài ra, còn có các tiện ích để tạo và quản lý cơ sở dữ liệu, phân tích khóa định loại đa truy. Một ví dụ về sử dụng BIOKEYS được cung cấp với khóa đa truy định họ thực vật ngành Mộc lan (Magnoliophyta) là khóa được xây dựng dựa trên hệ thống khóa đục lỗ của Betel Hansen và Knud Rahn (1969).

Từ khóa: khóa đa truy, định loại, cơ sở dữ liệu, phân tích khóa.