Revista Mexicana de Biodiversidad

Revista Mexicana de Biodiversidad 92 (2021): e923592

Taxonomy and systematics

Taxonomic identification keys on the web: tools for better knowledge of biodiversity

Claves de identificación taxonómica en la web: herramientas para un mejor conocimiento de la biodiversidad

Miguel Murguía-Romero [a], Bernardo Serrano-Estrada [b], Enrique Ortiz [a], José Luis Villaseñor [a, *]

[a] *Instituto de Biología, Universidad Nacional Autónoma de México, Tercer circuito s/n, Ciudad Universitaria, Coyoacán, 04510 Ciudad de México, Mexico*

[b] *SERES Sistemas Especializados, Membrillos Mz. 120 Lt. 38. Col. Ojo de Agua, Tecámac, 55770 Estado de México, Mexico*

*Corresponding author: vrios@ib.unam.mx (J.L. Villaseñor)

Received: 18 June 2020; accepted: 22 December 2020

Abstract

Difficulty in correctly identifying species in biological collections is an important impediment in confronting the current biodiversity crisis. The development of tools to improve taxonomic knowledge would help reverse this deficiency. Here, we propose an informatics system for the creation and use of polykeys on the web as tools for the identification of taxa (species, genera, families, etc.). The design is based on 4 actions: the ease of use of the software (usability), polythetic identification, a theoretical model of dynamic identification, and the use of relational databases. A system that applies this design is presented and exemplified using the FAMEX polykey, a tool for identifying the families of flowering plants (Magnoliophyta) of Mexico. The AbaTax system (www.abatax.abaco2.org) allows the creation of polykeys that are published and made available to all web users. The system considers the use of responsive web design, which is adapted in real time so that the interface is properly displayed to the type of device from which it is accessed, be it a desktop computer, laptop, tablet or cell phone.

Keywords: AbaTax; Biodiversity informatics; Dynamic identification; FAMEX; Magnoliophyta; Polykeys

Resumen

El desconocimiento de ubicar gran parte de la biodiversidad en la jerarquía taxonómica es una limitante para enfrentar su crisis actual. El desarrollo de herramientas para avanzar en un mejor conocimiento taxonómico ayudará a revertir esta deficiencia. Aquí se propone un esquema de sistema para la creación y uso de policlaves en la web, con la finalidad de proporcionar herramientas para la identificación de los taxones (especies, géneros, familias, etc.). El diseño se fundamenta en 4 acciones: la facilidad de uso del software (usabilidad), la identificación politética, un modelo

ISSN versión electrónica: 2007-8706; Universidad Nacional Autónoma de México, Instituto de Biología. https://doi.org/10.22201/ib.20078706e.2021.92.3592

teórico de identificación dinámica y el uso de bases de datos relacionales. Se presenta el sistema que aplica dicho diseño, ejemplificándolo mediante la policlave FAMEX, una herramienta para identificar a las familias de plantas con flores (Magnoliophyta) de México. El sistema AbaTax (www.abatax.abaco2.org) permite la creación de policlaves que son publicadas y puestas a disposición de todos los usuarios de la web. El sistema considera el uso de diseño de páginas web responsivas, que se adaptan en tiempo real para que la interfaz se muestre de forma adecuada al tipo de dispositivo desde el que se accede, ya sea una computadora de escritorio, una laptop, una tableta o un teléfono celular.

Palabras clave: AbaTax; Informática de la biodiversidad; Identificación dinámica; FAMEX, Magnoliophyta; Policlaves

Introduction

In Mexico there are more than 23,000 native species of vascular plants (Villaseñor, 2016), but knowledge about their geographical distribution is still deficient. The number of species in many areas of the country is underestimated and the current records of distribution do not cover their entire range (Gómez-Pompa et al., 2010; Koleff et al., 2008; Sosa & Dávila, 1994; Villaseñor et al., 2005). Floristic knowledge can improve with new explorations and collections in poorly explored regions of the country. Another avenue for improvement is filling in gaps in information, especially the taxonomic identification of material that has already been collected and stored in herbaria but has not been curated at species level (Villaseñor, 2015).

The biodiversity crisis, where many species are becoming extinct mainly due to the loss of their natural habitat, is aggravated by the lack of knowledge of many species. The correct taxonomic identification of organisms is essential to accelerate knowledge of biodiversity and reduce the negative effects of this crisis (Villaseñor, 2015). Biological information derived from taxonomic studies is used as source of information in evolutionary work and the quality of taxonomic information used in phylogenetic studies is determinant of the quality of the results found. Misidentification of organisms whose sequences are published in molecular databases —such as GenBank— can lead to erroneous results and inferences (Nilsson et al., 2006). The study of diversity patterns at different space-time scales requires inventories based on taxonomic units delimited and correctly identified under a system in which comparisons can be made between them, allowing the study of variations in diversity (Gotelli, 2004). Proper species identification is also key to the study of biodiversity distribution. Inaccurate identification can not only provide erroneous estimates of species ranges of distribution, but also on the diversity and composition of communities, committing both biogeography studies and the identification of priority conservation areas (Bortolus, 2008).

Van Regenmortel (2016) considers that a classification of viruses based only on nucleotide sequences is a classification of genome sequences and not of viruses. Molecular techniques can support the construction of genome-based classifications, and will be useful for the identification of microorganisms, cryptic species or when only organic fragments are available; it should also be noted that these techniques are not without problems or criticism (Nilsson et al., 2006; Will & Rubinoff, 2004). On the other hand, most consider still necessary to maintain a taxonomy focused on morphology and its information is still valid in many areas of biological research (Dunn, 2003; Gotelli, 2004). Integrative taxonomy, on the other hand, incorporates multiple sources of evidence, morphological, molecular, ecological, biogeographical, etc., into the analysis of taxonomic-nomenclatural decisions (Rajpoot et al., 2016; Sheth & Thaker, 2017). Still in this modern approach, identification keys based on morphological characters are the tools that make taxonomic classifications operational and are fundamental for the knowledge of biodiversity, however this 21st century is the "era of molecular biology and genomics" (Dunn, 2003; Scotland et al., 2003; Walter & Winterton, 2007).

Strategies for automatic taxonomic identification systems can be classified into 2 large groups: identification supervised by a human and unsupervised identification. In the former group, there are computer programs such as IntKey (Dallwitz et al., 1995, 1998) or Lucid (www.lucidcentral.com), in which the interface allows the user to indicate the observable characteristics of the specimen, usually through the character-character state scheme. Supervised identification keys can be classified into 2 main types: *a)* monothetic keys, which follow a predefined sequence of questions about the character states of the specimen, such as traditional dichotomous keys, and *b)* polythetic keys, in which the user can select the answer to a question from more than 1 of the character states as they are observed in the specimen being identified (Murguía-Romero & Villaseñor, 1992).

Among unsupervised identification systems are image recognition systems by automatic vision (Bonnet et al.,

2015; Watson et al., 2004; Weeks et al., 1997), in which a photograph of the specimen being identified is analyzed by the system and as a result it proposes an identification. These types of systems are designed for use by non-experts, but they are still far from the effectiveness of expert taxonomists (Bonnet et al., 2015; Gaston & O'Neill, 2004). On the contrary, users of supervised identification systems can be non-experts or experts.

The polythetic condition can be defined as a particularity of a class or group, for example a taxon, which is defined by a variable set, and is unique to the class of properties, none of which is necessarily present in each member of the class (Dubois, 2017) (Fig. 1). Specifically, in taxonomic identification, the polythetic condition is when a specimen can be associated with a taxon, not by a group of unique diagnostic characteristics, but by a set whose combination is unique. This polythetic condition, which refers mainly to the classification process, can also be applied to the identification process, in which this condition is more likely, since in many cases some of the diagnostic characters used in the classification are not present or not observable in the identification process. Many dichotomous keys are monothetic; however, when more than 1 route is provided for some taxa, they can be considered polythetic. In the unsupervised identification, the specimen can be identified as member of a taxon even without having information on its key or diagnostic characteristics (Morse, 1975).

Properties

Individuals	A	B	C	D	E	F	G	H
1	1	1	1					
2	1	1		1				
3	1		1	1				
4		1	1	1				
5					1	1	1	
6					1	1	1	
7					1	1		1
8					1	1		1

Figure 1. Schematization of the difference between 'polythetic' and 'monothetic' concepts. The presence of a property is indicated by the number 1. Individuals 1-4 constitute a polythetic group, where each individual records 3 of 4 properties and no property is common to all individuals. Individuals 5-6, 7-8 and 5-6-7-8 form 3 monothetic classes with 3, 3 and 2 properties, respectively, present in all members. Modified from Van Rijsbergen (1979) and Van Regenmortel (2016).

State of the art of online taxonomic identification programs

A web search of tools for creating online keys makes evident several features of the state of the art that are useful to guide the development of these tools. The list shown in Table 1 is far from exhaustive, but it is effective to illustrate the current situation of development of interactive identification keys. The following issues can be identified: *1)* there is no clear classification or single dominant paradigm that guides the future developments of online keys; *2)* no description of the computer development methodology used to develop the software is given; *3)* neither the model nor the design criteria for the user interface is described; *4)* the use of proprietary files is preponderant, whereas the use of relational databases is scarce, so effort expended in the development of one key cannot be easily harnessed for others; *5)* features that have been technologically available for more than a decade continue to be underutilized, such as apps for cell phones, voice recognition, use of colors in the users interfaces as an important frame of communication, among others. For example, these systems often do not use colors to communicate system states to the user, do not use responsive interfaces that adapt to the different types of devices according to the shape and size of the screen, or do not consider internet and cell phones as the predominant means of access to software and information. Another important situation is that most of the points discussed above are referenced only on the web and are not described in scientific publications.

In the development of software for taxonomic identification there is a delay in the incorporation of relational databases as a model of information representation. Although relational databases were widely used in the early 1970s, the field of taxonomic identification took almost 30 years to incorporate this technology. For example, one of the first programs that explicitly refers to the use of relational databases as a model for internal representation of information is the PANDORA program (Pankhurst, 1998).

The goal of this work is to present an identification tool, built as a web page that facilitates identification using already accessible keys, creates taxonomic identification keys, and publishes them immediately on the web for universal use. This enterprise takes into account the current situation of the development of interactive identification keys, which includes various aspects that have not allowed the consolidation of solid paradigms of this type of tools. Also, the development process of the tool presented considers the most important features that informatics technology offers today, which have been underutilized or ignored in the multiple efforts to build interactive identification keys.

Table 1

Examples of online keys and computer programs to create them. 'No reference' = bibliographic reference not available or not identified for the tool described.

Tool	Description
DAISY	Watson et al. (2004) Identification of Lepidoptera by image recognition (35 species with 20 images each). Natural History Museum, London and University of Costa Rica.
INTKEY	Dallwitz (1980); Dallwitz et al. (1995, 1998, 2000) Program to generate interactive keys from files in DELTA format. SCIRO, Canberra, Australia - dmitz, M. http://csiropedia.csiro.au/delta-taxonomic-computer-programs/
INTKEY	Seltmann (2004) Key created in INTKEY for 145 Hymenoptera taxa. University of Kentucky, USA.
LEASYS	Abdulrahaman et al. (2010) System to identify savanna trees in Nigeria based on leaf morphology. The identification interface is mainly composed of 2 "simple sheets" and "composite sheets" windows, which the user chooses according to the specimen being identified. Each window contains 7 and 8 characters respectively, with 2 possible states per character. Ilorin University, Ilorin, Nigeria.
LucID	Norton et al. (2012) Commercial software to produce interactive keys, either stand-alone or public on the web, on the company's site. Queensland, Australia Identic Pty. Ltd. http://www.lucidcentral.com
LucID	Bittrich et al. (2012) Key created in Lucid for 649 angiosperm genera of the Ducke reserve, Brazil. State University of Campinas, Brazil. http://www.ib.unicamp.br/plantkeys
MEKA	No reference System to build interactive keys; the first version for MS-DOS is from 1986, the latest version from 2005 is for an obsolete Windows operating system. Meacham, C. A. University and Jepson Herbarium - University of California, Berkeley, USA. http://ucjeps.berkeley.edu/meacham/meka/
MKey Xper^3	No reference Interactive key generator, part of the Xper^3 platform, a web development with broader purposes than taxonomic identification. The data matrix is imported using comma separated files. Laboratory of Informatics and Systematics of the Pierre et Marie Curie University, France. http://www.xper3.fr
MKey Xper^3	Chrétiennot-Dinet et al. (2014). Key for 58 species of *Chrysochromulina* (phytoplankton) with 9 characters built in MKey. University of Paris, France. http://wwwphp.obs-banyuls.fr/chrysochromulina
MKey Xper^3	Rojas-Cortés (2017) Use MKey to create a key of 251 tree species from the Los Tuxtlas Tropical Biology Station of the UNAM.
NaviKey	No reference Applet created in 1999 to build interactive keys based on the DELTA format, the latest version is presented as a downloadable * .jar file to be installed on a web server. Neubacher, D. and Rambold, G. University of Bayreuth, Germany. http://www.navikey.net
SLIKS	No reference Program in Javascript language that can be downloaded and installed locally or on a web server. The data matrix is imported from a proprietary text format file in the form of lists of items delimited by square brackets and separated by commas. Created in 2004 and maintained until 2012 by Guala, G.F. United States Geological Survey, USA. http://www.stingersplace.com/SLIKS/

Table 1. Continued

Tool	Description
Symbiota	Gries et al. (2014) Web platform for administration and information consultation of specimens of biological collections. The system has a "Dynamic Key interface" module. http://symbiota.org
WEBiKEY	Attigala et al. (2016) Key to 7 species of the genus *Kuruna* (Poaceae). The database model for storing the data matrix is described. Iowa State University - Iowa Crop Improvement Association, Iowa, USA. http://webikey.agron.iastate.edu
3I	Dmitriev (2006). The data matrix is stored in MS Access 2000. The interface is based on forms created with the tools of the commercial package itself. The program was created in 2003, and the latest version is from 2011. Illinois Natural History Survey, USA. http://dmitriev.speciesfile.org

Materials and methods

Our strategy was to design a polythetic taxonomic identification system based on a model we call "Dynamic Identification" (Murguía-Romero, 1992; Murguía-Romero & Villaseñor, 1998). This system allows the user to operate in 2 directions: by introducing information on the character states present in the specimen being identified and waiting for the system to report the taxa as possible identities, or the user can explore the character states that occur in a taxon considered as a possible hypothesis to refute or accept. This model has been the result of past experiences in the creation of interactive online keys (Murguía-Romero, 1987; Murguía-Romero & Villaseñor, 1993).

The system was designed following a three-layer architecture (Fowler, 2002): *1)* the user interface, *2)* the business layer or algorithms of the application, and *3)* the data layer. The system design is based on: a) the usability of the software, *b)* polythetic identification, *c)* the theoretical model of dynamic identification, and *d)* the use of relational databases. The points b and c constitute the algorithms and methods that the system automates in layer 2.

Other important features considered during the design were the use of colors to indicate system states to the user, the construction of a responsive interface, that is, an interface that adapts to different devices' screen sizes, the use of cell phones and the possibility that the user may or may not be connected to the internet, the use of open source software for its construction as much as possible, and when open source software was unavailable, the use of free software, thus avoiding the payment of rights.

Usability

The usability of a system refers to the extent to which it can be used by users to achieve specific objectives with effectiveness, efficiency and satisfaction. It is important to underline that all this is within a specific context of use (Bevan et al., 2016). One of the main features of the system's design in terms of usability was that it could be used from the web and on different types of devices. Therefore, we decided to use responsive web design technology (Marcotte, 2010): pages that detect the type of device from which the website is viewed so that decisions about the layout of the interface elements, such as buttons, menus, and windows, can be automatically optimized. For example, on a desktop computer screen, the application can be presented with 2 windows that are displayed simultaneously, whereas on a cell phone, the application can show only 1 window at a time, with a link that allows the change of window to display.

Another feature implemented is that the information on characters and character states is presented to the user as monolithic statements, joining both in a single sentence. For example, the character state 'Tree or shrub' and its character 'Lifeform' are presented together with the statement 'Woody plants (trees or shrubs)'. Storage in the database is done in a differentiated way, i.e., there is one table for the characters and another for the character states, which includes a field called 'statement'.

Polythetic identification

The identification algorithm used is polythetic; the taxa that remain as possible identities of the specimen under determination are those in which the presence of character states in the data matrix are recorded, constituting a subset

of the presences indicated by the user as observable in the specimen. Polythetic identification is based on the fact that a particular combination of character states present in a specimen is only compatible with 1 or few taxa represented in the data matrix by the union of the presences of character states of a large set of specimens of the same species (or taxonomic group) and which must include all the taxa in the lower category. In this work, the various ways to refer to the type of taxonomic tools discussed here, such as multi-access keys, polykeys, interactive keys and online keys are considered synonyms.

Dynamic identification model

The theoretical framework of computer development was the Dynamic Identification Model (Murguía-Romero & Villaseñor, 1998), whose main characteristics are its simplicity and the possibility of indicating a hypothesis or name of the taxon suspected of being the identity of the specimen (Fig. 2). Regarding its first characteristic (simplicity), it is implemented considering 4 aspects: *a)* the type of data of the taxonomic data matrix is Boolean, that is to say 'true' or 'false'; *b)* the concept of 'statement' is created, which is a sentence that specifies a character state along with the character to which it belongs; for example, the statement 'Woody plants (trees or shrubs)' represents the pair 'character - character state' (character: life form; character state: trees or shrubs); *c)* system with few windows, minimizing the need for navigation, and *d)* scanning in 2 directions through the same interface; in one direction you can find out which taxa present a certain set of character states and in the other the character states that are present in a given taxon or set of taxa.

The possibility of indicating a hypothesis, that is, the name of the taxon suspected of being the identity of the specimen, makes explicit use of the supervised identification process. Since it is a human being who is interacting with the system (rather than automatic image recognition), the user refutes their own suspicions about the possible identities of the specimen, making an interactive user-system feedback process.

Currently the system does not implement denial, that is the possibility that the user indicates that a certain character state is not present, or the denial of a hypothesis (indicating that he suspects that a taxon is not the identity of the specimen). Thus, only the quadrants in the corners of figure 2 are implemented in the interface. This decision was made not because denial could not be programmed or implemented, but because of the complexity that it would add to the user interface, making it less understandable and intuitive.

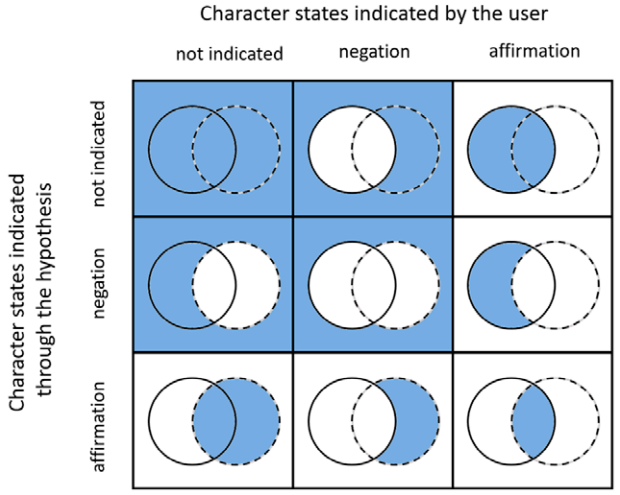

Figure 2. The 9 areas of the logical navigation space of dynamic identification (adapted from Murguía-Romero & Villaseñor, 1998). Both axes represent character states indicated by the user; on the horizontal axis character states are explicitly indicated (circles with solid lines); on the vertical axis, the character states are indicated by an identification hypothesis (the character states present in the hypothetical taxon, represented as circles with dashed lines). The shaded area logically represents the set of character states that may be present in the specimen being identified.

Database model

The information structure consists of a relational database that includes the tables specified by the user in the process of creating a polykey, such as the list of taxa, the list of statements (character-character state), and the data matrix of character states present in taxa. In addition, the database includes other tables that the system makes use for administrative purposes; for example, the catalog of polykeys in the system or the users or types of access. The general structure of the database is described in the user manual of the tool.

Results

The design was implemented on a web platform called AbaTax (www.abatax.abaco2.org). The website can be used on any device (cell phone, tablet or computer) and adapts to display optimally on that device. Users can create their own polykeys by preparing Excel files in specific formats with lists of taxa, characters and character states, as well as the presence-absence data matrix. This is explained in more detail in the section 'Creating polykeys

in AbaTax'. If the files comply with the structure required by the system, the construction of the polykey only requires importing the files and recording some administrative data, such as names of the authors, name of the taxonomic group and date of creation.

Several polykeys are currently available on the AbaTax platform, mainly for plants, such as the FAMEX polykey for families of flowering plants (Magnoliophyta) of Mexico and the GENCOMEX polykey for the genera of Compositae of Mexico; additionally, there are 20 additional available publicly and 80 privately, accessible only to the user who created them or whoever decides to share the corresponding link and password. The results of the system with the proposed design are shown below, exemplifying it with the FAMEX polykey and with the polykey for species of *Ageratina* (Asteraceae) of the State of Mexico, both available for use and consultation at www.abatax.abaco2.org

The dynamic identification interface was implemented using 2 lists (Fig. 3), one for character states (left) and one for taxa (right). The list of character states is displayed in a statement format composed of a single sentence that associates the character and character states to make the list more readable.

Figure 3 shows the interface when the user has selected 2 statements: herbaceous plants (annual or perennial, including subshrubs) and plants with thorns (on stems or leaves). At the top of the interface a message displays: selected character states: 2 of 150 and possible identities of the specimen: 49 of 264. Throughout the session the user can select more statements, following their observations on the specimen, in order to reduce the list of possible identities of the specimen to a single taxon.

The 'advanced view' button allows the user to access an alternative mode of identification by entering a hypothesis (figure 4, Acanthaceae as selected taxon). In this mode, statements corresponding to character states occurring in the family are highlighted with a green background, and the user will know that the identification hypothesis is contradicted if he selects statements that are not highlighted. In 'advanced view' mode, the number of buttons on the top bar is increased, since it is possible to filter the various areas of the logical navigation space (Fig. 2). For simplicity, the negation that is included in the dynamic identification model has not been implemented, so only positive assignments, both in the statements (selection of character states) and in the taxa (hypotheses), are considered. Thus, 4 possible elements are generated. The button with the trash icon is used to start a new session, deleting the selections previously made.

For each statement and taxon, it is possible to associate 1 or more images available to the user to support the taxonomic identification process (Fig. 5a). On the other hand, the polykey's information is stored in a relational database, from where the character and character state of each statement is displayed in the interface. That is useful

a

b

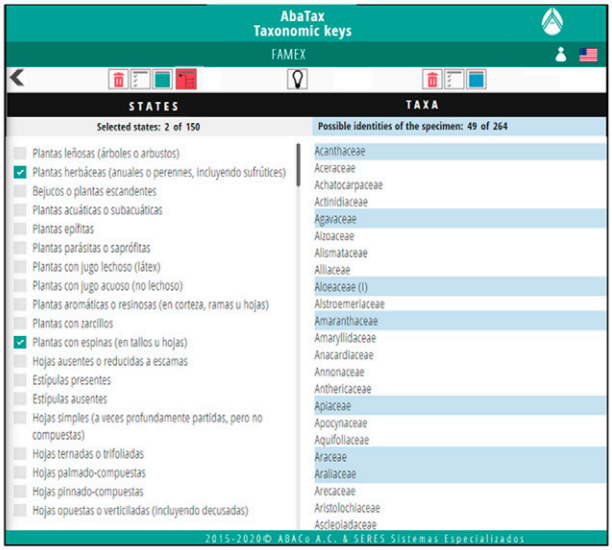

Figure 3. Basic view of the dynamic identification interface when displaying the FAMEX polykey (www.abatax.abaco2.org). a) Interface with the button "Show statements" off; b) interface with the button "Show statements" on. The list on the left shows the a) character - character states, or b) character state statements; the list on the right shows the list of possible taxa considered as identities of the specimen being determined. The symbol ✓ indicates the selected characters, and the taxa that comply with the selection are blue shaded.

Figure 4. Advanced view of the dynamic identification interface. The Acanthaceae family has been selected as an identification hypothesis (first line in the list on the right); statements congruent with this hypothesis are highlighted in green and with a ✓ the selected ones. Any selection by the user of a character state statement that is not highlighted would indicate that the state refutes the hypothesis, that is, that the specimen does not belong to the Acanthaceae family.

for the automatic generation of taxonomic descriptions. An example for a species included in the polykey 'Species of the genus *Ageratina* (Asteraceae) in the State of Mexico' is shown in figure 5b.

Creation of polykeys in AbaTax

There are 2 methods for creating a polykey in AbaTax: importing Excel files or using the web interface editor. The 'User Manual', available on the AbaTax page, explains the user interface in general and how to create polykeys in detail. On the same webpage, tutorial videos are available that explain step by step how to use each section. Below is a brief description of how to create polykeys in AbaTax.

Creation by importing Excel files

This method is recommended when a polykey is created for the first time. It is necessary to specify 5 Excel files, each with a single sheet: *cat_taxa*: list of taxa; *cat_caracter*: list of characters; *cat_estado*: list of character states associated with each character. For each one a statement is defined which is then displayed in the polykey interface; *cat_grupo*: groups to which each character belongs, used to organize the automatic generation of taxonomic descriptions; *dat_matriz*: presence-absence matrix of each character state for each taxon.

These Excel files can be created directly by the user or generated by exporting a taxonomic database created in AbaTaxEdit, a system based on Microsoft Access ® specifically for this purpose, also available on the AbaTax website.

The use of the web interface is recommended to make modifications to existing polykeys created by importing Excel files. Through this interface the user can modify the presence-absence data matrix, the list of characters, character states and taxa (Fig. 6).

AbaTax incorporates 2 options for access and visualization. The user can publish the taxonomic key within the system as follows. Private, only the user is able to view and access using their own password. This option is recommended when the specialist is still working and

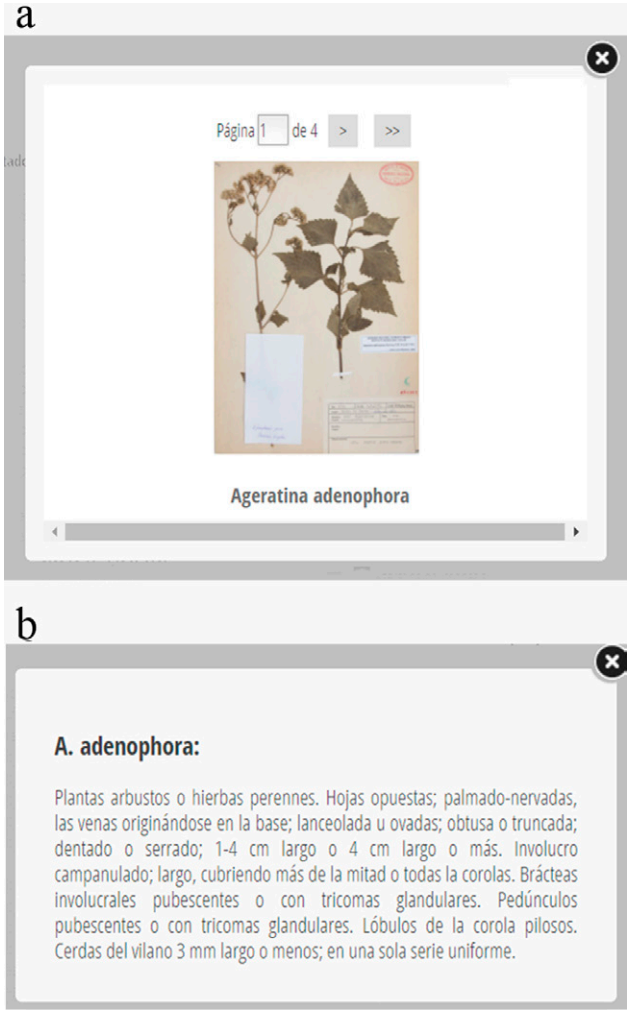

Figure 5. Views of attributes in the key "Species of the genus *Ageratina* (Asteraceae) in the State of Mexico". a) Displayed image of the species *A. adenophora*; b) taxonomic description generated automatically for the same species.

making modifications. Public, the key is made available to everyone, allowing any user to access it and use it freely.

Review and editing of the keys through collective work of multiple users

The importance of multiple specialists working collectively is an important point that AbaTax has considered within its functionalities. AbaTax offers the possibility of inviting more users to collaborate on the same polykey, under different types of access, such as only for review (read only) or also for editing.

Abatax has been migrated to a mobile application with iOS and Android operating systems (Fig. 7), which can be downloaded for free. Actually, the FAMEX key is installed by default, being able to install any other available public key, which can then be used without an internet connection. Cell phone applications have the advantage that they can be used in places where there is no Internet connection, so they are useful tools in the field.

Discussion

Building new informatics programs for taxonomic identification depends on development methodologies, as well as the use of new technologies. Today, interactive keys have been built without following or without documenting the process of their development, omitting many relevant technical aspects, such as the method of computer development, system architecture, how information is stored internally, and which algorithms are used to process it. The construction of increasingly efficient taxonomic identification tools depends on previous experiences; their documentation in an explicit and orderly manner will result in new tools being able to take advantage of those previous experiences, using them as true building blocks for more solid and useful technology for biodiversity research. Unlike previous keys, ours documents the development process, reports technical details including the method of computer development, system architecture, and how information is displayed to be used.

Useful future identification tools must be constructed considering 3 aspects: *1)* the model and algorithm underlying the tool, *2)* the data model for storing information, and *3)* the criteria applied in the design of the user interface. The dynamic identification model is both, a way of specifying the identification algorithm and a model that is familiar to users that carry out taxonomic identification. By allowing the user to propose an identification hypothesis, the tool approaches the way in which the taxonomist already goes about the identification process, making a fluid experience.

Penev et al. (2009) indicate that polykeys "are generally based on taxa matrices *vs* characters, and these matrices can also be the basis of other taxonomic products, as long as the matrix format is general enough and adequate software is available". The relational database model is currently the information storage tool with the most general format (Codd, 1970, 1990), which is why we propose storing the data matrix in a relational database rather than other formats that require translators to exchange information between different systems. Even for almost 2 decades, it had been anticipated that the DELTA format, over time, would be transformed into a relational database model (Pankhurst, 1991), but until now there is still no alternative proposal to translate it in a relational database. Regarding the condition that "adequate software be available" so that data matrices are useful for generating other taxonomic

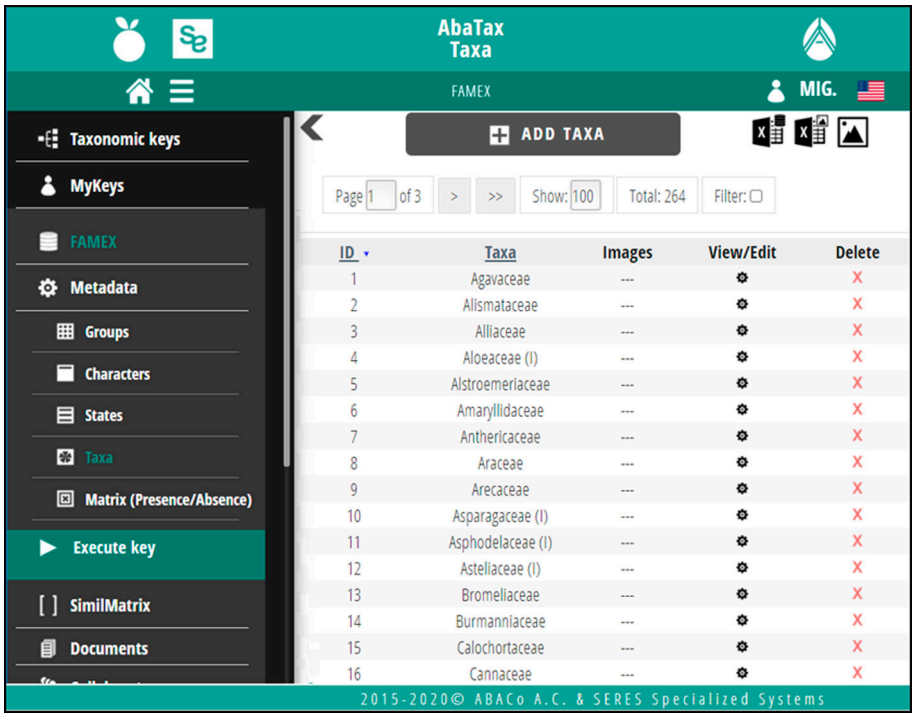

Figure 6. AbaTax interface used to modify or add taxa in the polykey.

products (as defined by Penev et al., 2009), we believe that 'relational database management systems' (RDBMS) should be exploited to the full extent of their possibilities before looking for alternatives. Currently, RDBMS are the most general tools for information management, and their use will result in better communication of information between different systems and less effort in programming algorithms to generate additional taxonomic products with the same information, for example, taxonomic descriptions.

The use of a relational database as a model to store information in computer programs to generate interactive keys facilitates the evolution of the architecture of these systems. The relational model is a universal, precise, and agile way of indicating the type of information contained in the software and its degree of normalization. In addition, it shows in a synoptic way the structure of the information through the different tools of the model, such as relationship diagrams or the catalog and dictionary of the database. All of this facilitates the understanding of the limits and possibilities of the software that makes use of a particular database model.

When designing computer identification tools, the available technological possibilities must be taken into account, as well as the usability criteria that allow an intuitive and direct user approach to the identification process. Taking advantage of the characteristics offered by computer technology (such as the use of colors or mobile applications) and incorporating them in a structured way (through formal models) into automated taxonomic identification keys will allow tools to be built in a more efficient way and increase their usefulness in improving knowledge of biodiversity.

The statement structure used in AbaTax is closer to the mental model of the taxonomist than the character-character state paradigm. An interface based on statements has several advantages over one based on character-character state. On the one hand, a logical sentence is presented to the user that unites concepts in a natural way, as is done, for example, in taxonomic descriptions, which taxonomists are already accustomed to, facilitating their reading and understanding. Secondly, a list of sentences usually requires fewer clicks in the machine-taxonomic interaction, since in the character-character state structure, character states are usually only displayed once the character is clicked, a step that is not necessary in the list of sentences. Thus, character-character state may be the information structure that underlies the tools, but is not necessarily the optimal format for the user who wishes to identify a specimen. The "statement" groups together in a single unit 2 concepts that must ultimately be related, while the duplex character-character state must be presented in the interface as such, occupying not only

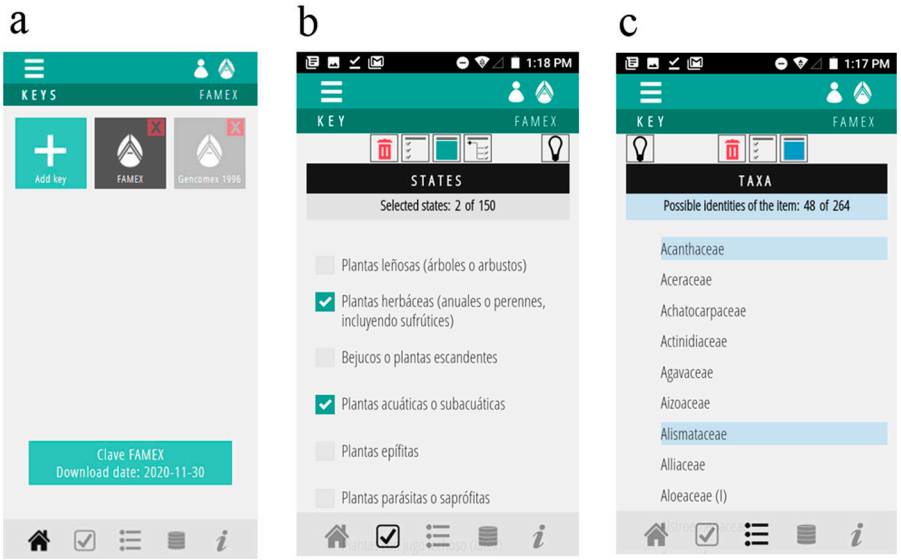

Figure 7. Representative screens of the AbaTax mobile application. a) Start´s menu screen showing that 2 keys has been installed: FAMEX (key to families of flowerings plants of Mexico) and GENCOMEX (key to genera of Asteraceae of Mexico); b) list of elegible character states of the FAMEX key; c) list of possible identities of the specimen given the list of selected character states.

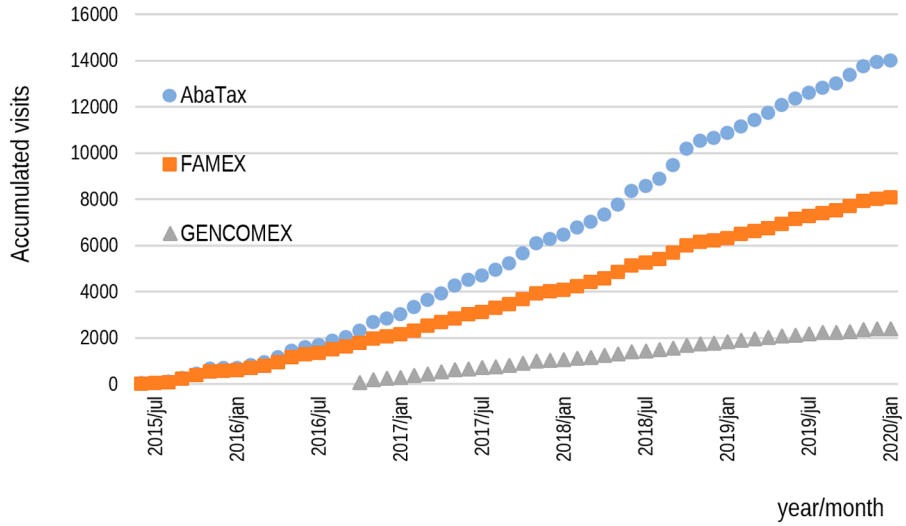

Figure 8. Accumulated visits to the keys on the AbaTax platform. Total visits are indicated as blue circles, visits to the FAMEX key as orange squares, and visits to the GENCOMEX key as gray triangles.

graphical space, but also mental space, since it leaves to the user the task of joining the 2 and associating their meaning. The "statement" already presents this association in a way that is logical from a taxonomic point of view, providing a direct and clear meaning, with the possibility of including additional details.

AbaTax was launched in June 2015; since then, consultations of its identification keys exceed 14,000, with an average of 250 queries per month. The FAMEX key for families of flowering plants in Mexico has more than 8,000 queries to date, while the GENCOMEX key, for genera of Compositae of Mexico more than 2,000 (Fig.

8). The total views of the AbaTax page number around 18,000. More than 3 quarters of these total visits (14,000 or 77.8%), include consultations of polykeys. The FAMEX mobile application has been downloaded 1,220 times and has recently been made available for download from the Google Play and Apple's App Store stores, making it easy to install and distribute. Workshops, presentations at conferences and talks have been organized to publicize the AbaTax tool. All these experiences have been useful to evaluate the usability of the system, as well as to obtain user feedback and improve the interface.

The comments on the AbaTax tool that we have received from workshop attendees and teachers and students who have used it in bachelor's and graduate level courses have been useful to guide the improvement of the tool by correcting errors, adapting the interface and improving computational efficiency.

In the context of a particular classification, taxonomic identification keys are currently the most efficient tool for the assignment of a specimen's taxonomic name to a specimen. An essential step in curating any biological collection is the assignment of a taxonomic name, without which specimens lose value, and thus the resources invested in collecting the specimens are largely wasted. Identification keys that take advantage of computer automation can facilitate this often arduous task.

The creation of polykeys on the web also facilitates their universal use, not only because they can reach more users, but because their distribution extends to more types of devices, such as tablets or cell phones. The tools for the creation of identification polykeys must explicitly document the data model in which the information is stored internally, the model of the interaction between the user and the tool, and the criteria used to design the user interface. Nowadays, the web represents a mechanism that can be used to face the biodiversity crisis, in which the task of generating reliable floristic listings is essential and where the correct taxonomic identification of collected specimens is a necessary prerequisite.

Acknowledgments

Guadalupe Segura tested AbaTax on the web by building polykeys and providing suggestions and bug reports that have been corrected. Rosario Redonda and Rafael Torres have used the web system and the FAMEX mobile application in his courses on taxonomic identification, which has allowed the verification of its proper functioning.

References

Abdulrahaman, A. A., Asaju, I. B., Arigbede M. O., & Oladel, F. A. (2010). Computerized system for identification of some savanna tree species in Nigeria. *Journal of Horticulture and Forestry*, *2*, 112–116.

Attigala, L., De Silva, N. I., & Clark, L. G. (2016). Simple Web-Based Interactive Key Development Software (WEBiKEY) and an Example Key for Kuruna (Poaceae: Bambusoideae). *Applications in Plant Sciences*, *4*, 1500128. https://doi.org/10.3732/apps.1500128

Bevan, N., Carter, J., Earthy, J., Geis, T., & Harker, S. (2016). New ISO standards for usability, usability reports and usability measures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9731*, 268–278. https://doi.org/10.1007/978-3-319-39510-4_25

Bittrich, V., Souza, C. S. D., Coelho, R. L. G., Martins, M. V., Hopkins, M. J. G., & Amaral, M. C. E. (2012). An interactive key (Lucid) for the identifying of the genera of seed plants from the Ducke Reserve, Manaus, AM, Brazil. *Rodriguésia*, *63*, 55–64.

Bonnet, P., Joly, A., Goëau, H., Champ, J., Vignau, C., Molino, J. F. et al. (2015). Plant identification: man vs. machine: LifeCLEF 2014 plant identification challenge. *Multimedia Tools and Applications*, *75*, 1647–1665. https://doi.org/10.1007/s11042-015-2607-4

Bortolus, A. (2008). Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *AMBIO: A Journal of the Human Environment*, *37*, 114–118. https://doi.org/10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2

Chrétiennot-Dinet, M. J., Desreumaux, N., & Vignes-Lebbe, R. (2014). An interactive key to the *Chrysochromulina* species (Haptophyta) described in the literature. *Phytokeys*, *34*, 47–60. https://doi.org/10.3897/phytokeys.34.6242

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the Association for Computing Machinery*, *13*, 377–387.

Codd, E. F. (1990). *The relational model for database management: version 2*. Boston, Massachusetts: Addison-Wesley Longman Publishing Co., Inc.

Dallwitz, M. J. (1980). A general system for coding taxonomic descriptions. *Taxon*, *29*, 41–46.

Dallwitz, M. J., Paine, T. A., & Zurcher, E. J. (1995). User's guide to Intkey: a program for interactive identification and information retrieval. Available at: http://delta-intkey.com

Dallwitz, M. J., Paine, T. A., & Zurcher, E. J. (1998). Interactive keys. In P. R. Scott, P. Bridge, P. Jeffries, & D. Morse (Eds.), *Information technology, plant pathology and biodiversity* (pp. 201–212). Wallingford, United Kingdom: CAB International.

Dallwitz, M. J., Paine, T. A., & Zurcher, E. J. (2000). Principles of interactive keys. Available at: http://delta-intkey.com/www/interactivekeys.pdf

Dmitriev, D. A. (2006). 3I, a new program for creating Internet-accessible interactive keys and taxonomic databases and its application for taxonomy of *Cicadina* (Homoptera). *Russian Entomological Journal*, *15*, 263–268.

Dubois, A. (2017). Diagnoses in zoological taxonomy and nomenclature. *Bionomina*, *12*, 63–85.

Dunn, C. P. (2003). Keeping taxonomy based in morphology. *Trends in Ecology & Evolution*, *6*, 270–271. https://doi.org/10.1016/S0169-5347(03)00094-6

Fowler, M. (2002). *Patterns of enterprise application Architecture*. Boston, Massachusetts: Addison-Wesley Longman Publishing Co., Inc.

Gaston, K. J., & O'Neill, M. A. (2004). Automated species identification: why not? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*, 655–667.

Gómez-Pompa, A., Krömer, T., & Castro-Cortés, R. (2010). *Atlas de la flora de Veracruz: un patrimonio natural en peligro*. Veracruz: Universidad Veracruzana/ Gobierno del Estado de Veracruz.

Gotelli, N. J. (2004). A taxonomic wish-list for community ecology. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*, 585–597. https://doi.org/10.1098/rstb.2003.1443

Gries, C., Gilbert, E. E., & Franz, N. M. (2014). Symbiota - A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal*, *2*, e1114. https://doi.org/10.3897/BDJ.2.e1114

Koleff, P., Soberón, J., Arita, H. T., Dávila, P., Flores-Villela. O., Golubov, J. et al (2008). Patrones de diversidad espacial en grupos selectos de especies. In Conabio (Ed.), *Capital natural de México, Vol. II: estado de conservación y tendencias de cambio* (pp. 323–364). México D.F.: Comisión Nacional para el Conocimiento y Uso de la Biodiversidad.

Marcotte, E. (2010). *Responsive web design*. París: A Book Apart. Available at: http://www.reprosol.be/sites/reprosol.beta.the-aim.be/files/responsive-webdesign(ethan-marcotte).pdf

Morse, L. E. (1975). Recent advances in the theory and practice of biological specimen identification. In R. J. Pankhurst (Ed.), *Biological identification with computers* (pp. 11–52). London and Orlando: Academic Press.

Murguía-Romero, M. (1987*). FAMEX, policlave para la identificación de Magnoliophyta de México (Bachelor Thesis)*. Facultad de Ciencias, Universidad Nacional Autónoma de México. México D.F.

Murguía-Romero, M. (1992). *Métodos en la identificación biológica automatizada (Master Thesis)*. Facultad de Ciencias, Universidad Nacional Autónoma de México. México D.F. http://132.248.9.195/pmig2017/0187269/Index.html

Murguía-Romero, M., & Villaseñor, J. L. (1992). La computadora en la identificación botánica. *Ciencia y Desarrollo (México)*, *18*, 130–137.

Murguía-Romero, M., & Villaseñor, J. L. (1993). *FAMEX: Policlave para familias de plantas con flores (Magnoliophyta) de México*. México D.F.: Asociación de Biólogos Amigos de la Computación, A.C.

Murguía-Romero, M., & Villaseñor, J. L. (1998). GENCOMEX: a computerized key for identify the genera of Asteraceae of Mexico. In P. R. Scott, P. Bridge, P. Jeffries, & D. Morse (Eds.), *Information technology, plant pathology and biodiversity* (pp. 305–314). Wallingford, United Kingdom: CAB International.

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. H., & Kõljalg, U. (2006). Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *Plos One*, *1*, e59. https://doi.org/10.1371/journal.pone.0000059

Norton, G. A, Patterson, D. J., & Schneider, M. (2012). LucID: a multimedia educational tool for identification and diagnostics. *International Journal of Innovation in Science and Mathematics Education*, *4*, 2000.

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

Pankhurst, R. J. (1998). A historical review of identification by computer. In P. R. Scott, P. Bridge, P. Jeffries, & D. Morse (Eds.), *Information technology, plant pathology and biodiversity* (pp. 305–314). Wallingford, United Kingdom: CAB International.

Penev, L., Sharkey, M., Erwin, T., van Noort, S., Buffington, M., Seltmann, K. et al. (2009). Data publication and dissemination of interactive keys under the open access model. *Zookeys*, *21*, 1–17. https://doi.org/10.3897/zookeys.21.274

Rajpoot, A., Kumar, V. P., Bahuguna, A., & Kumar, D. (2016). DNA barcoding and traditional taxonomy: an integrative approach. *International Journal of Current Research*, *8*, 42025–42031.

Rojas-Cortés, Á. P. (2017). *Manual de identificación con base en atributos foliares de los árboles de la estación de biología tropical Los Tuxtlas, Veracruz (Master Thesis)*. Instituto de Investigaciones en Ecosistemas y Sustentabilidad, Universidad Nacional Autónoma de México. Morelia, Michoacán, México.

Scotland, R., Hughes, C., Bailey, D., & Wortley, A. (2003). The *big machine* and the much-maligned taxonomist DNA taxonomy and the web. *Systematics and Biodiversity*, *1*, 139–143. https://doi.org/10.1017/S1477200003001178

Seltmann, K. (2004). *Building Web-based interactive keys to the Hymenopteran families and superfamilies (Master Thesis)*. College of Agriculture, University of Kentucky. Lexington, Kentucky.

Sheth, B. P., & Thaker, V. S. (2017). DNA barcoding and traditional taxonomy: an integrated approach for biodiversity conservation. *Genome*, *60*, 618–628. https://doi.org/10.1139/gen-2015-0167

Sosa, V., & Dávila, P. (1994). Una evaluación del conocimiento florístico de México. *Annals of the Missouri Botanical Garden*, *81*, 749–757.

Van Regenmortel, M. H. V. (2016). Classes, taxa and categories in hierarchical virus classification: a review of current debates

on definitions and names of virus species. *Bionomina*, *10*, 1–21.

Van Rijsbergen, K. (1979). *Information retrieval*. London: Butterworths.

Villaseñor, J. L. (2015). ¿La crisis de la biodiversidad es la crisis de la taxonomía? *Botanical Sciences*, *93*, 3–14. https://doi.org/10.17129/botsci.456

Villaseñor, J. L. (2016). Checklist of the native vascular plants of Mexico. *Revista Mexicana de Biodiversidad*, *87*, 559–902. https://doi.org/10.1016/j.rmb.2016.06.017

Villaseñor, J. L., Maeda, P., López, J. J. C., & Ortiz, E. (2005). Estimación de la riqueza de especies de Asteraceae mediante extrapolación a partir de datos de presencia-ausencia. *Boletín de la Sociedad Botánica de México*, *76*, 5–18.

Walter, D. E., & Winterton, S. (2007). Keys and the crisis in taxonomy: extinction or reinvention? *Annual Review of Entomology*, *52*, 193–208. https://doi.org/10.1146/annurev.ento.51.110104.151054

Watson, A. T., O'Neill, M. A., & Kitching, I. J. (2004). A qualitative study investigating automated identification of living macrolepidoptera using the Digital Automated Identification SYstem (DAISY). *Systematics and Biodiversity*, *1*, 287–300.

Weeks, P., Gauld, I., Gaston, K., & O'Neill, M. (1997). Automating the identification of insects: A new solution to an old problem. *Bulletin of Entomological Research*, *87*, 203–211. https://doi.org/10.1017/S000748530002736X

Will, K. W., & Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, *20*, 47–55. https://doi.org/10.1111/j.1096-0031.2003.00008.x