University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

USDA Systematic Entomology Laboratory | Entomology Collections, Miscellaneous

1-1-2009

Data publication and dissemination of interactive keys under the open access model

Lyubomir Penev
Bulgarian Academy of Sciences and Pensoft Publishers, info@pensoft.net

Michael Sharkey
University of Kentucky

Terry Erwin
Smithsonian Institution

Simon van Noort
South African Museum

Matthew Buffi ngton
Smithsonian Institution

See next page for additional authors

Follow this and additional works at: http://digitalcommons.unl.edu/systentomologyusda

Part of the Entomology Commons

Penev, Lyubomir; Sharkey, Michael; Erwin, Terry; Noort, Simon van; Buffi ngton, Matthew; Seltmann, Katja; Johnson, Norman; Taylor, Matt; Th ompson, F. Christian; and Dallwitz, Michael J., "Data publication and dissemination of interactive keys under the open access model" (2009). *USDA Systematic Entomology Laboratory*. Paper 66.
http://digitalcommons.unl.edu/systentomologyusda/66

This Article is brought to you for free and open access by the Entomology Collections, Miscellaneous at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in USDA Systematic Entomology Laboratory by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Authors
Lyubomir Penev, Michael Sharkey, Terry Erwin, Simon van Noort, Matthew Buffi ngton, Katja Seltmann, Norman Johnson, Matt Taylor, F. Christian Th ompson, and Michael J. Dallwitz

This article is available at DigitalCommons@University of Nebraska - Lincoln: http://digitalcommons.unl.edu/systentomologyusda/66

ZooKeys 21: 1–17 (2009)
doi: 10.3897/zookeys.21.274
www.pensoftonline.net/zookeys

FORUM PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

Data publication and dissemination of interactive keys under the open access model

ZooKeys working example

Lyubomir Penev[1], Michael Sharkey[2], Terry Erwin[3], Simon van Noort[4], Matthew Buffington[5], Katja Seltmann[6], Norman Johnson[7], Matt Taylor[8], F. Christian Thompson[9], Michael J Dallwitz[10]

1 *Central Laboratory of General Ecology, Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria* **2** *Department of Entomology, University of Kentucky, Lexington, Kentucky 40546–0091, USA* **3,9** *Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013–7012 USA* **4** *Natural History Division, South African Museum, Iziko Museums of Cape Town, PO Box 61, Cape Town 8000, South Africa* **5** *Systematic Entomology Lab, USDA/ARS c/o NMNH, Smithsonian Institution 10th & Constitution Ave NW PO Box 37012 MRC-168 Washington DC 20013 USA* **6** *Department of Entomology, North Carolina State University, Campus Box 7613, 3212 Gardner Hall, Raleigh, NC 27695–7613, USA* **7** *Department of Entomology, The Ohio State University, 1315 Kinnear Road, Columbus, Ohio 43212, USA* **8** *Lucidcentral, Centre for Biological Information Technology, Level 6 Hartley-Teakle Building, The University of Queensland, Brisbane Qld 4072 Australia* **10** *13 Warrambool Close, Giralang ACT 2617, Australia*

Corresponding author: *Lyubomir Penev* (info@pensoft.net)

Received 20 August 2009 | Accepted 14 September 2009 | Published 23 September 2009

Citation: Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ (2009) Data publication and dissemination of interactive keys under the open access model: ZooKeys working example. ZooKeys 21: 1–17. doi: 10.3897/zookeys.21.274

Abstract

The concepts of publication, citation and dissemination of interactive keys and other online keys are discussed and illustrated by a sample paper published in the present issue (doi: 10.3897/zookeys.21.271). The present model is based on previous experience with several existing examples of publishing online keys. However, this model also suggests ways to publish, cite, preserve, disseminate and reuse the original data files to the benefit of future workers, the authors, and society in general. To be regarded as a "formal scientific publication," an online key should satisfy the same criteria of peer review, registration, persistence, bibliographic description, etc., as with conventional publications. Keys can be published as either "static" or "dynamic" publications. We define a "static" publication as a discrete unit of information preserved in a persistent and unchangeable way on the publisher's Web site and/or on paper and consequently in con-

Copyright *Lyubomir Penev* et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ventional/electronic libraries and archives. This contrasts with the nature of the Internet, which allows and tends to encourage updating and improvement on a continuing basis. We call "dynamic" a publication of an interactive key on a Web site where its contents can be continuously updated. "Dynamic" publications meet some of the criteria of a "formal scientific publication" (identification, citation and location), while they lack other important features of it (persistence, archiving, indexing, science metric and citation metric services). Dynamic Web-based interactive keys may benefit from publishing the first version of their underlying datasets in a form of "formal scientific publication". We define here the minimum set of data files to be published for several different platforms (Intkey, Lucid2, Lucid3, MX) to ensure both (1) priority, identification, location and citation of the first published work and (2) future use and re-use of the keys.

Introduction

"What is it?" is the fundamental question which begins most biological inquiries. The answer is provided by the identification process, that is, a sequence of observations of a specimen until a matching taxon is selected on the basis of a unique combination of attributes (character states). The identification process has been an integral part of biology since Linnaeus first developed his *Systema Naturae* (1758, 10th edition). In his *Systema*, the attributes were directly embedded in his hierarchical classification. Lamarck, in 1788, was the first to separate the identification process from the classification by making keys: hierarchical sets of choices between character states (Walter and Winterton 2007). These are usually called "dichotomous keys" (although strictly speaking this term should be restricted to keys having only two choices at each branching point), and sometimes "conventional keys", "single-access keys", or "pathway keys". They are still the mainstay of taxonomy today, although they have many deficiencies (Dallwitz et al. 2000), the most fundamental of which is that the sequence of attributes used is predetermined by the author.

"Multi-access keys" or "polyclaves" were developed to overcome these deficiencies. They allow users to select, from a set of attributes, the ones that they think are best for their particular location on situation. For example, if one is identifying an angiosperm but the specimen has no flowers or fruits, other characters, such as those relating to habit and leaves, can be used. These keys may use punched cards (e.g. Archbald 1967) or computer programs (e.g. Goodall 1968; see also Pankhurst 1991). We shall refer to computerized, multi-access keys as "interactive keys". Note that this definition excludes computerized dichotomous keys – e.g., keys in which each character is on a separate HTML page, and the paths to subsequent characters are implemented as hyperlinks.

Interactive keys are usually (but not always – see Dallwitz 1992) based on character-taxon matrices. These matrices can also be the basis of other taxonomic products, provided that the matrix format is sufficiently general, and suitable software is available. For example, Morse (1974) describes programs for interactive identification, construction of dichotomous keys, and description printing from matrices. DELTA (Dallwitz 1983; TDWG 2007) is a more powerful data-matrix format, which is also comparatively easy for humans (as opposed to computers) to read. DELTA data can be used to interact with cladistic software, as well as the products mentioned above

(DELTA Programs and Documentation: http://delta-intkey.com/www/programs.htm; Dallwitz 1993; Dallwitz et al. 1993). The latest TDWG standard for data matrices is Structured Descriptive Data (SDD) (TDWG 2007). Software is available to produce interactive keys and descriptions from SDD (e.g. Lucidcentral 2009).

Along with improvements in the identification process, improvements have been made in dissemination media. For most of the history of taxonomy, the medium was ink on paper, i.e., the traditional printed publication. The first multi-access keys were distributed on punched cards – either edge-punched cards or IBM 80-column punched cards. Computer programs and data were distributed on punched cards, punched paper tape, magnetic tape, magnetic (floppy) disks, optical disks (e.g. CD-ROM), and computer networks. With the development of a world-wide network, the Internet, this method of dissemination began to supplant the others, not only for computer programs and data, but also for traditionally paper-based publications. In particular, both dichotomous and interactive keys could now be made available online (Dallwitz et al. 2002). With this new medium, the traditional methods of scientific publication are becoming obsolete. The aims, which date back to the Royal Society and Henry Oldenberg (1660s), of recognition, priority, and peer review remain valid today, but the mechanisms need to be modified to adapt to the new online medium. Such modifications are outlined herein for online identification systems.

The present paper does not aim toward a thorough overview of interactive keys or at analyzing the advantages/shortcomings of the different software platforms; these are available elsewhere (e.g. Dallwitz 2000; Dallwitz et al. 2000; Farr 2006; Walter and Winterton 2007). Rather, we discuss and offer a model for *open data publishing* that allows a wider dissemination and reuse of data and interactive keys, and proper publication credit. The model adheres to the principles of open access to biodiversity information proclaimed in the opening editorial paper of ZooKeys (Penev et al. 2008), and recently illustrated by publishing occurrence datasets underlying taxonomic revisions (Miller et al. 2009; Penev et al. 2009).

"Formal scientific", "static", and "dynamic" publication

In this time of revolutionary changes in the way scientific information is being disseminated, the concepts of publication must be clearly defined. The process for "formal scientific" publication involves several steps, the most distinct of which are:

(i) peer-review (not mandatory but desirable) and editorial quality control;
(ii) registration of the publishing media in international registries through ISBN or ISSN numbers;
(iii) long-stranding persistence and supply of the published information (on paper or in digital form) through libraries, trade, digital repositories (all recently facilitated by the presence of digital object identifiers (DOIs));
(iv) accurate bibliographic description (metadata) of each scientific publication, which facilitates a proper identification, location and citation;

(v) registration, distribution and dissemination of relevant parts of the published information through specialized indexing and harvesting aggregators (e.g., Global Biodiversity Information Facilities (GBIF), Encyclopedia of Life (EOL), Catalogue of Life, ITIS and Species 2000, Zoological Record, CABI Abstracts, Current Contents and many others);
(vi) registration in citation, bibliometric and science metric indices (e.g., Thompson-Reuters Institute of Scientific Information (ISI), Google Scholar, Scopus);
(vii) distribution through established supply channels, i.e., book trade, journal subscriptions, or in a form of free open access publications.

Online keys displayed on a Web site are not formal scientific publications unless the above requirements are met. This does not mean that such keys are without merit; rather, authors generously provide data for other researchers, typically without obtaining publication credit from their superiors for their efforts.

A traditional definition of "publication" includes the concept of "a discrete unit of information published in a persistent and unchangeable way"; usually such publications are called "static" or "fixed." This contrasts with the nature of the Internet, which allows and tends to encourage updating and improvement of the published information on a continuing basis; such publications can be termed "dynamic." The contradiction between the traditionally static nature of scientific publications and the new opportunities to change and update their contents online may be partially overcome by citing the version and date of access of a website. However, this does not fully identify the information being cited unless the Web site is systematically archived with version numbers (in which case the versions become static publications). The possibility of accurately citing a Web resource, however, does not transform it automatically into a formal scientific publication because other steps of the publishing process (see i–vii above) must also be applied during creation of the particular Web resource.

We offer a solution to formalize publication of dynamic interactive keys. The first version of datasets underlying a Web-based interactive key may be published in an online journal or book as data files, complete with title, authors, abstract, possibly introduction and other relevant text material. Such a practice would bring benefit to the authors ensuring their priority, publication and citation credit, as well as a proper dissemination, archiving and indexing of the first published work through the established channels the scholarly publishing.

Publishing keys

In the context of identification, examples of reconciliation between the dynamic nature of the Internet, and the static nature of conventional publications, are the pioneering models of Flora Online (http://www.mobot.org/plantscience/ResBot/FO/FloraOnline.htm) and the Canadian Journal of Arthropod Identifications (CJAI) (Marshall 2008;

see journal's Web site at: http://www.biology.ualberta.ca/bsc/ejournal/ejournal.html). Flora Online was the first electronic journal to receive an ISSN number from the Library of Congress. It published several interactive keys, as well as checklists, etc., in the period 1987–1993. It was available online (but the keys could not be *used* online – they had to be downloaded, and used offline), and was also distributed on floppy disks. Once published, material could not be altered, although new versions could be (and were) published.

CJAI offers at least two versions of each key: a dichotomous key is embedded in a conventional PDF publication, and a separate and possibly enhanced HTML version of this key is made available online. A limited number of printed copies of the PDF file are distributed to libraries. Despite the advances the CJAI model has provided, one limitation is that CJAI does not include the raw data files that were used to generate the keys. Unless the data files are obtained from the original authors, future revisers will be forced to duplicate much of the work that went into the original publication. For example, adding a single species to an already published key often requires substantial time and effort.

Here we suggest criteria for a more complete static publication model for keys, utilizing the resources we have at the present time (see also Fig. 1):

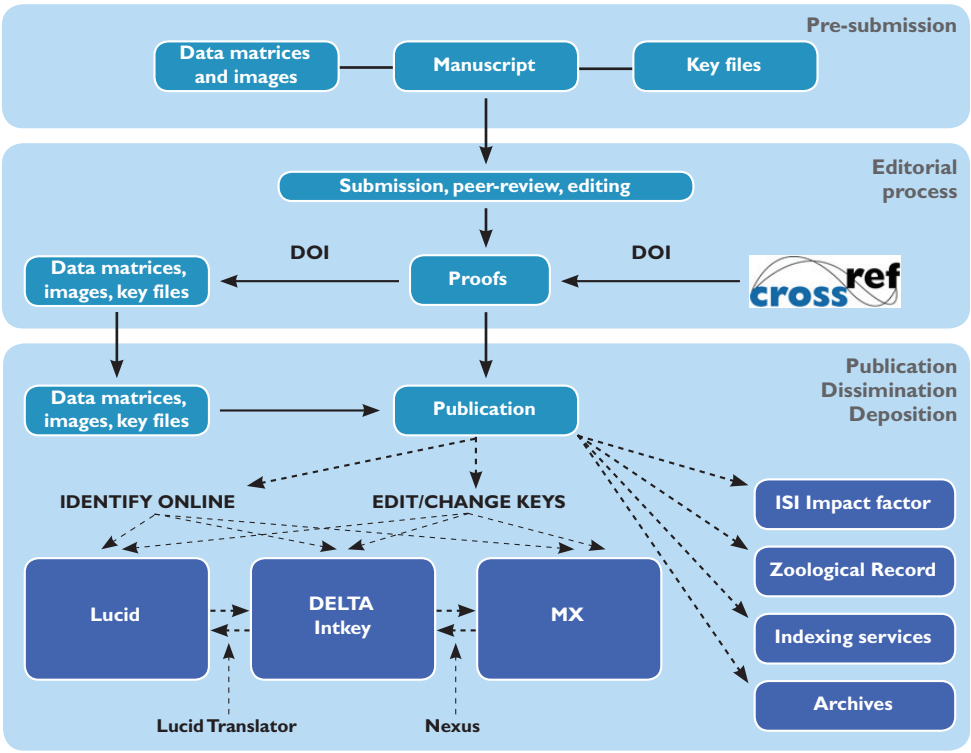

Fig. 1. The ZooKeys model for data publication and dissemination of interactive keys.

1. **Publication of keys in the traditional, dichotomous form on paper in a journal, report, or book, normally with an ISSN or ISBN.** The journal article or book might include, or be closely associated with, supplementary material such as full descriptions, interactive keys, images, and the data from which the keys, descriptions, etc. were generated. This material might be on microfiche, floppy disk, CD, or DVD. Some examples are: Watson and Dallwitz (1981) (supplementary material on microfiche); Britton (1986) (DELTA data included in the paper publication); Watson and Dallwitz (1988) (supplementary material, including interactive key, on microfiche and floppy disks; Pirchegger (1998) (supplementary material on CD-ROM).
2. **Publication of keys in the traditional, dichotomous form in an online (or both online and print) journal or book.** The article might be closely associated with supplementary online material such as full descriptions, interactive keys, images, and the data from which the keys, descriptions, etc. were generated. The article and the associated material are associated with a single DOI, if a key is published stand-alone. In case a key or keys is/are published as a part of larger paper, e.g., taxonomic revision, their datasets can be published as supplementary files under separate DOIs (usually semantically related to the DOI of the article). Examples that include such supplementary material under a single DOI are the first 8 issues of CJAI (http://www.biology.ualberta.ca/bsc/ejournal/ejournal.html, accessed 6 September 2009). The DOI of each issue resolves to a HTML file containing the author, title, and abstract, and also containing links to the article in PDF form, and links, direct and indirect, to the supplementary material.
3. **Stand-alone publication, under separate DOIs, of interactive keys or dichotomous HTML keys.** This type of publication should comprise the necessary program and data files and/or HTML pages; title, authors, date, and abstract; and possibly associated material such as an introduction. The interactive keys would usually work online, but might also (or sometimes only) be available for downloading and use offline.
4. **Stand-alone publication, as a downloadable set of files under a separate DOI, of all primary data files related to the keys (e.g., data matrices and images).** This type of publication should contain title, authors, date, and abstract; and possibly associated material such as an introduction. The data files will be preserved primarily on the publisher's server and secondarily through leading international e-publication archives, such as LOCKSS, Portico, e-depot and others. This option will allow the updating of the data, and generation from the data of new keys (to be republished statically or dynamically) by researchers other than the original authors (naturally, provided that the original author and source are credited). For the data files, adoption of standards developed and supported by the Biodiversity Information Standards (TDWG) (http://www.tdwg.org) community is recommended to ensure maximum interchangeability and persistence. The latest TDWG standard is Structured

Descriptive Data (SDD); however, the earlier DELTA standard is still widely used, and neXML (NEXUS file format, http://www.nexml.org) may reach maturity in the near future.

Any or all of criteria 1–4, applied together with criteria i–vii for formal publication, will ensure that authors of the publication(s) get credit through peer review, registration, citation, distribution, dissemination, long-term preservation, and bibliometry. This will incorporate all of the advantages of the traditional and open access models, while allowing wide dissemination and reuse of original data, multiplying the benefits to both authors and the end users.

Well-referenced primary data files are encouraged, with literature citations and information on voucher specimen material provided for each couplet, Operational Taxonomic Unit (OTU), image, or matrix coding

To provide maximum dissemination and use, a key potentially may be published as a combination of some or all 4 forms listed above, as it is partly illustrated in the paper of Sharkey et al. (2009) in the present issue.

Key data files, key program files, primary data files

The different software packages for implementing interactive keys have different folder and file structures for their data and programs. As examples, we illustrate these structures using the fruit-fly package at http://delta-intkey.com/ffa/ (Carroll et al. 2002). The package is generated from DELTA data files, and consists of natural-language descriptions, with links to images of the taxa, and interactive keys using Intkey, Lucid2, and Lucid3.

The contents of the folder http://delta-intkey.com/ffa/ is shown in the listing ffadir.txt, which is published with the present paper (Appendix 1, doi: 10.3897/zookeys.21.274.app.1.ik.).

Publishing of MX data files is illustrated on the example of the paper of Sharkey et al. (2009, doi of the data files: 10.3897/zookeys.21.271.app.4.ik)

"Key data files" and "key program files" are the files required for running an interactive key. The "basic" key data files are the minimum set of files required to run the key, such as the character list, taxon names, and data matrix. The "supplementary" key data files are the other files used by the key, such as images and descriptions. As published, the keys would normally be run online, that is, via an Internet connection. Some keys can *only* be run online, because the key program is designed to run on a Web server rather than on a personal computer. In other cases, it may be possible, at least in principle, to download the keys and run them offline. However, such keys may not be downloadable in practice because of restrictions imposed by the Web site. Some kinds of interactive key may work *only* by downloading and running them offline, but this is rare nowadays.

"Primary data files" are the data files from which the key data files and other products (e.g., dichotomous keys, descriptions, classifications) are generated. The "basic" primary data files contain data such as the character list, taxon names, and data matrix.

The "supplementary" primary data files are associated files such as images and descriptions (if the latter are not generated from the basic data). The basic primary data files are usually different from the basic key data files. However, some key programs use the primary data files directly, e.g. NaviKey (NaviKey 2009), which runs directly from DELTA data files (plus other files).

Intkey

The ***Intkey program files*** must be downloaded and installed on the user's computer before the interactive keys can be run. The program requires MS-Windows. Later versions of Intkey have more features, and these features have often required corresponding enhancements of the Intkey data files. Nevertheless, all recent program versions and data versions are intended to be compatible, and new program versions have bug fixes, as well as enhancements. It is therefore recommended that the latest version of Intkey (see http://delta-intkey.com/www/programs.htm) should always be used, rather than storing older versions of the program on journal Web sites.

The ***basic key data files*** are ffa.ink and ffai.zip; ffa.ink is a "startup file", containing information about the locations of various files:

```
: To run the Intkey identification package from this file
; you need Intkey5 and Windows 95/NT or later
;
InkFile=http://delta-intkey.com/ffa/ffa.ink
   ; Name of this file
DataFile=http://delta-intkey.com/ffa/ffai.zip
   ; Name of the compressed data file
InitializationFile=intkey.ink
   ; Name of the Intkey initialization file within the compressed data file
ImagePath=http://delta-intkey.com/ffa/images
   ; Image path
InfoPath=http://delta-intkey.com/ffa/info
   ; Information path
```

Two folders, "images" and "info", contain images and RTF descriptions that may be accessed by the key. ("www" contains the same descriptions in HTML format. These are not used by Intkey; they are linked from the main contents page (http://delta-intkey.com/ffa/index.htm) so that users can view them *without* using Intkey.)

When the user follows the link to ffa.ink (from the Web page "Pest Fruit Flies of the World - Interactive Keys": www/_wident.htm), this file is read by Intkey (which must already be installed on the user's PC). Intkey then uses this information to download (to a temporary file) and read the file ffai.zip, which contains all of the information necessary for the basic key. If and when the user asks to see an image or descrip-

tion, that image or description is downloaded to a temporary file and displayed (i.e., the same way that image and HTML files are handled when viewed in a browser).

When the user exits from the key, Intkey asks "*Do you want to save the data set you have just downloaded and used on your own disk?*" If the user clicks "*Yes*", the temporary copies of ffa.ink and ffai.zip are copied to permanent files in a folder of the user's choosing, and ffa.ink is modified to point to this copy of ffai.zip. Thus, when the user later opens this local ffa.ink, Intkey gets the basic key files from the local copy of ffai.zip, and doesn't need to access the Web. However, the Web will still be accessed if and when the user asks for an image or description, because ffa.ink is still pointing to the Web folders for those files. The user can, if he wishes, download the complete "images" and "info" folders (e.g., by ftp), and modify ffa.ink to point to the local copies. Then the key will never access the Web (unless, of course, it happens to contain links to other Web resources, e.g., in citations of publications).

If the user clicks "*No*" in response to the "*save the dataset?*" prompt, Intkey asks "*Do you want to add the URL for the data set to the index file?*". If the user clicks "*Yes*", Intkey adds the URL to its own index file. This index is displayed whenever Intkey is started, and the user can choose any of the datasets in the index (which may include local keys, as well as ones on the Web).

To summarize, all the user normally has to do is click on the link to the key, and then choose whether to save and/or index the key on exit.

The ***basic primary data files*** from which the key and other products are generated comprise (normally) three DELTA text files: specs, chars, and items. In addition, there are other "directives" files which describe the links from characters and taxa to the images, and how the data are to be processed into other forms (e.g. where to put new paragraphs in descriptions), and subsidiary files such as the introduction and references. The author would normally keep all of these in a single folder. The ***primary data files*** comprise a zipped version of this folder, and the folder containing the images. (The descriptions that accompany Intkey packages are not normally part of the primary data files; they are generated from those files.)

The DELTA Editor has its own binary data files, but these are not recommended for data exchange. The Editor imports and exports DELTA text files.

Lucid2

The ***Lucid2 key program files*** must be downloaded and installed on the user's computer before the interactive keys can be run. The program requires MS-Windows. There may be incompatibility between program and data versions, but any newly published key will presumably use the latest version, which is available at http://lucidcentral.org, or Lucid3 (see below).

The ***basic key data file***, to which the Web page links, is "pest fruit flies.lcd". When the user follows the link to this file, it is read by the Lucid2 Player (which must already be installed on the user's PC). The key links to images in the folder "images" (shared

with Intkey), descriptions in the folder "html", and macros in the folder "macros". The descriptions in the "html" folder are almost the same as the ones in the "www" folder, except that the latter contain a link back to the contents page, which is not appropriate in the Lucid context. Both forms are generated automatically from the DELTA data.

The key file can be downloaded by the usual browser mechanism, and run locally. It will access images and descriptions from the Web as required.

Although the "images" and "html" folders can be downloaded (e.g., by ftp), the downloaded key won't be able to access them. This is because the paths are built into the binary key file (unlike Intkey, where they are in easily modified text files). To access the images etc. locally, it would be necessary to download the LIF data file (see below), and rebuild the key.

The ***basic primary data file*** from which the key is generated is a Version 2 LIF file. The ***primary data files*** comprise this file and folders containing the images, descriptions, and macros.

The Lucid Builder Version 2 has its own binary data files, which are not normally used for data exchange. The Builder imports and exports LIF files.

Lucid3 applet keys

Lucid3 applet keys will run under MS-Windows, Solaris, Linux, and Mac OS X. All of the ***Lucid3 key files***, comprising the key data files and the key program files, are packaged for Web deployment by the Lucid Builder, as follows.

l3\ffa.html
l3\key\ffa.lkc3
l3\key\ffa\ffa\ffa.* (i.e. ffa.dep, .fil, .sco, and .xml)
l3\key\ffa\media\images* (the images)
l3\key\ffa\media\html* (the descriptions)
l3\lucid_player* (the program)

This directory structure is rigidly imposed by the Lucid3 Builder. It is a little awkward to share the images with Intkey and/or Lucid2, but it could be done by using the greater flexibility of the latter programs.

Before using the key, the Java Runtime Environment (JRE) must be installed on the user's local (client) computer. When the user follows a link to "l3\ffa.html", the applet in the "l3\lucid_player" folder is downloaded and run. The applet then downloads the rest of the key files.

The basic key will work locally if the directory tree containing the basic key files is downloaded, but it won't access the images etc. from the Web. The key will work locally if the *whole* directory tree is downloaded (i.e., including the "media" tree); however some Web sites may not support downloading.

The ***basic primary data file*** from which the key is generated is a Version 3.x LIF3 file. These are XML files, and the later versions at least are supposed to conform

to SDD. The ***primary data files*** of a static publication comprise this file and folders containing the images and descriptions.

The Lucid Builder Version 3.x has its own binary data files, which are not normally used for data exchange. The Builder imports LIF and LIF3 files, and exports LIF3 files (export is only in the latest LIF3 format, which can't be read by earlier 3.x Builders).

Lucid3 keys can now be implemented by a program running on the server (instead of by an applet on the client) (Lucidcentral 2009).

MX

MX (Yoder et al. 2006) keys are Web based. To build new keys a user must either obtain an account with an existing instance of MX or a new installation on a machine that runs a Web server. Basic installation instructions are found on the MX wiki (http://purl.oclc.org/net/mx-installation) and presently require fairly advanced technical experience. The project is open source with the code base available on SourceForge (http://sourceforge.net/projects/mx-database/develop). MX is coded in Ruby on Rails with a MySQL database.

There is not a specific **key file** for MX, however, **data files** are downloadable as text. Present examples of these files are linked in the Revision of the Oriental genera of Agathidinae (Sharkey et al. 2009) and contain database-generated output describing components of the key. These include: a character/character state list, and NEXUS file. Images are also made available and linked to character state and Operational Taxonomic Unit (OTU) descriptions. These output files are adequate to properly recreate the key, but the format of these files (except the NEXUS file) are presently under review and should not be considered standardized output at the present time.

To run the key a user needs Internet access and the key URL (http://purl.oclc.org/NET/oriental-agathidinae). MX is optimized for Firefox Web browser.

To summarize, Table 1 provides a list of minimum set of files to be published for each platform.

Table 1. Minimum set of files to publish for Intkey, Lucid and MX to ensure a proper data use and re-use

Platform	Key data files	Primary data files	Software
Intkey	.ink startup file, .zip data file, images, descriptions	specs, chars, items, other directives files, images	Intkey5
Lucid2	.lcd file, images, descriptions, macros	Version2x .LIF, images, descriptions, macros	Lucid2 Player
Lucid3	.html, .lkc3, .dep, .fil, .sco,and .xml; images, descriptions	Version3x .LIF3, images, descriptions	Lucid3, Java Runtime Environment (JRE)
MX	none	NEXUS, Character list with image ids, OTU list with image ids, images	MX, Web-based

Flexibility of data use

It is desirable to have software to convert between the different formats for primary data, so that users can use take advantage of the capabilities of a variety of programs.

The program Confor (a component of the DELTA software package available at http://delta-intkey.com/www/programs.htm) translates basic DELTA data (the "specifications," character list, taxon names, and matrix, i.e. the "specs," "chars," and "items" files) into Nexus format, but important information may be lost because of the limitations of the Nexus format, at least in the version targeted by this conversion (see Dallwitz 1999). The program NDE (http://taxonomy.zoology.gla.ac.uk/rod/NDE/nde.html) can convert NEXUS data to DELTA.

The Lucid Translator (http://www.lucidcentral.org/Software/Lucid3/LucidBuilder/LucidTranslator/tabid/189/Default.aspx, accessed 30th of June 2009) translates between basic DELTA data and Lucid Interchange Format Version 2 (.LIF files), in both directions. However, the capabilities of these formats differ, and important information may be lost in both directions of translation (see Dallwitz 1999). For example, LIF does not support text characters or characters with more than 15 states, and DELTA does not directly support the Lucid "by misinterpretation" coding (though a workaround is available – see DELTA Programs - Test Versions and Utilities: http://delta-intkey.com/testp/). Later versions of Lucid Interchange Format support more of the DELTA features, but there is no direct route between these versions and DELTA. Version 2 LIF files can be imported into later versions of Lucid, and thus provide a pathway from DELTA (with the associated information loss already mentioned), but conversion in the other direction is currently not possible.

The key to Oriental agathidine genera (Sharkey et al. 2009) in the present issue was originally produced using DELTA and subsequently converted to Lucid. The Lucid Translator was used to import the basic DELTA data into the Lucid Builder, as described above. However, the taxon and character images had to be linked manually to the Lucid key.

A complete translator is not at present available to convert DELTA to MX. The basic DELTA data was acquired from a Nexus file produced by Confor, as described above. Images were associated with character states using information in the DELTA cimages file. The images were manually associated with character states and needed reformatting before entering them into MX. Multiple images illustrating character states were combined into one image in the Oriental Agathidinae DELTA data. These had to be broken apart back into individual images before importing them to MX for illustrating character states. Import scripts for images, OTU names and characters were used after the data was manipulated into a format MX handles.

Previously publications also allowed comparison of interactive keys using different software with the same data: Carroll et al. (2002) for Intkey, Lucid2 and Lucid3 and Watson and Dallwitz (2004) for Intkey and NaviKey; and the DELTA data are available from the authors on request. However, Sharkey et al. (2009), in addition to including a MX key, provide a solution for a formal, static publication of the keys.

Copyright

Copyright law is complex, varying between countries and content type. Different legal tools exist to attempt to simplify the process allowing easy application of licenses for creative and scientific products, ensuring future open access to these products.

The open access publishing model under the terms of the Creative Commons Attribution License (BY) (http://www.creativecommons.org) permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. These licenses are applied to the creative product or something created from raw data. Creative Commons licenses may be applied to images, Web sites and keys. Raw data is treated differently under law. Data associated to an open access publication, data in a database, or data published stand alone, are treated by a related set of legal tools called Open Data Commons (http://www.opendatacommons.org). Currently, Open Data Commons offers two types of licenses (and associated statements) to cover copyright issues concerning open data publishing:

(1) Public Domain Dedication and License, assuming that data may be used and reused in an unrestricted manner. The statement associated with this license is as follows:
This {DATA(BASE)-NAME} is made available under the Public Domain Dedication and License version v1.0 whose full text can be found at http://opendatacommons.org/licenses/pddl/
(2) Open Database License, which requires (similar to CC Share-Alike) public reusers of data to share back changes:
This {DATA(BASE)-NAME} is made available under Open Database License whose full text can be found at http://opendatacommons.org/licenses/odbl/. Any rights in individual contents of the database are licensed under the Database Contents License whose text can be found http://opendatacommons.org/licenses/dbcl/

Identification, location and citation

There is at least one good reason to identify data separately from the publication they are related to. Published data may be incorporated into larger datasets and should be identified properly so that the original author and source may be traced in consequent uses. The DOI number assigned to a published dataset serves as such an identifier; moreover, it holds the function not only of *identifier* but also as a bibliographic metadata *descriptor* (through CrossRef, http://www.crossref.org), as well as a locator (through CrossRef's domain http://dx.doi.org). Moreover, the number of downloads of data files can be tracked separately and automatically and further exposed on the site so that authors will gain additional credits for use of their data.

Another reason to provide separate citation details for datasets implies in cases where several interactive keys of different taxa are published as datasets within one and the same paper, e.g., in a large taxonomic revision.

The DOIs of the publication and datasets/keys related to it can be described in a form of a "parent-child" relationship. It seems convenient DOIs of datasets/keys to be semantically related to the DOI of the "parent" publication to facilitate harvesting and indexing through machine-generated methods. Datasets related to interactive keys can be identified with the acronym "ik" (from "interactive keys") within its DOI, which is a semantic extension of the paper's DOI (worked examples from the paper of Sharkey et al. (2009) in the present issue):

if a DOI of a paper is: 10.3897/zookeys.21.271

then the DOI of the dataset will have the form:

doi: 10.3897/zookeys.21.271.app.1.ik

which means that the downloadable file is published as Appendix 1 of the paper.

Naturally, in most cases of stand-alone publications of interactive keys, there is no need to imply a "child" DOI, because it can't identify the key as a separate digitally published object. In such cases, it would be enough to assign a DOI to the publication itself and a link to either associated data files or to the external Web platform where the key is hosted.

To summarize, we describe the following combinations to identify and cite online keys, in correspondence with criteria 1–4 listed in the "Publishing Keys" section above:

1. Interactive keys are published stand-alone and contain bibliographic metadata descriptors (author, title, publication date, publication details, abstract), possibly introduction, as well as supplementary files published and hosted on the journal's platform:

   <Author> <(Year)> <Title of the publication> <Journal, volume, issue, pages> <Doi of the publication>.

2. Interactive keys are published dynamically and hosted on an external Web platform; their first-version datasets are published stand-alone in an online journal or book and contain bibliographic metadata descriptors (author, title, publication date, publication details, abstract), possibly introduction, as well as a Web link to the key:

   <Author> <(Year)> <Title of the publication> <Journal, volume, issue, pages> <Doi of the publication>. Interactive key: <Interactive key format > <Link to the external Web site>. Version (if applicable), day of accession.

3. Interactive keys are published as datasets accompanying either a conventional dichotomous key or a paper of a wider scope (e.g., taxonomic revision):

   <Author> <(Year)> <Legend of the dataset>. <File format> DOI of the dataset <DOI> Dataset published in: <Journal, volume, issue, pages> < DOI of the publication>

Dissemination and use

Publishing the datasets underlying a taxonomic paper under the open access model benefits everyone, from the authors to other researchers, as well to publishers and society in general. The social and psychological constraints, benefits, and incentives are being actively discussed (i.e., Costello 2009; Smith 2009). We believe that publishing online keys in a variety of forms (conventional dichotomic, interactive dichotomy or matrix, static or dynamic) within one and the same publication will visibly enhance their use and citation rates. By opening the possibility to modify and update keys, a long-standing perpetuity of a key's life is facilitated, bringing all of the deserved credits to the original and subsequent authors.

References

Archbald D (1967) Quick-key guide to trees: trees of northeastern and central North America. Doubleday, Garden City, NY, 1–69 + 50 punched cards.

Britton EB (1986) A revision of the Australian chafers (Coleoptera: Scarabaeidea: Melolonthinae). Vol. 4. Tribe Liparetrini: genus *Colpochila*. Australian Journal of Zoology Suppl. ser. 118: 1–135.

Carroll LE, White IM, Freidberg A, Norrbom AL, Dallwitz MJ, Thompson FC (2002 onwards) Pest fruit flies of the world. http://delta-intkey.com/ffa/

Costello MJ (2009) Motivating online publication of data. BioScience 59: 418–427. doi:10.1525/bio.2009.59.5.9.

Dallwitz MJ (1993) DELTA and INTKEY. In Fortuner R (Ed) Advances in computer methods for systematic biology: artificial intelligence, databases, computer vision. The Johns Hopkins University Press, Baltimore, Maryland, 287–296. http://delta-intkey.com/www/dallwitz-1993-delta_intkey.htm

Dallwitz MJ (1996 onwards) Programs for interactive identification and information retrieval. http://delta-intkey.com/www/idprogs.htm

Dallwitz MJ (1999 onwards) A comparison of formats for descriptive data. http://delta-intkey.com/www/compdata.htm

Dallwitz MJ (2000 onwards) A comparison of interactive identification programs. http://delta-intkey.com/www/comparison.htm [accessed on 30th of June 2009]

Dallwitz MJ, Paine TA, Zurcher EJ (1993 onwards) User's guide to the DELTA System: a general system for processing taxonomic descriptions. 4th edition. http://delta-intkey.com/www/uguide.htm

Dallwitz MJ, Paine TA, Zurcher EJ (2000 onwards) Principles of interactive keys. http://delta-intkey.com/www/interactivekeys.htm [accessed on 21st of August 2009]

Dallwitz MJ, Paine TA, Zurcher EJ (2002 onwards) Interactive identification using the Internet. http://delta-intkey.com/www/netid.htm

Farr DF (2006) On-line keys: more than just paper on the Web. Taxon 55(3): 589–596.

Lucidcentral (2009) Welcome to Lucidcentral. http://lucidcentral.org

Marshall S (2008) Editor's Notebook. Canadian Journal of Arthropod Identification, http://www.biology.ualberta.ca/bsc/ejournal/notebook.html [accessed on 27th of June 2009]

Miller JA, Griswold CE, Yin CM (2009) The symphytognathoid spiders of the Gaoligongshan, Yunnan, China (Araneae, Araneoidea): Systematics and diversity of micro-orbweavers. ZooKeys 11: 9–195. doi: 10.3897/zookeys.11.160.

Morse LE (1974) Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. Publications of Michigan State University Museum biol. ser. 5: 1–128.

NaviKey (2009) NaviKey home. http://www.navikey.net

Pankhurst RJ (1991) Practical taxonomic computing. Cambridge University Press, Cambridge, 202 pp.

Penev L, Erwin T, Thompson FC, Sues H-D, Engel MS, Agosti D, Pyle R, Ivie M, Assmann T, Henry T, Miller J, Ananjeva NB, Casale A, Lourenço W, Golovatch S, Fagerholm H-P, Taiti S, Alonso-Zarazaga M, Nieukerken E van (2008) ZooKeys, unlocking Earth's incredible biodiversity and building a sustainable bridge into the public domain: From "print-based" to "web-based" taxonomy, systematics, and natural history. ZooKeys Editorial Opening Paper. ZooKeys 1: 1–7. doi: 10.3897/zookeys.1.11

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009) Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11: 1–8. doi: 10.3897/zookeys.11.210.

Pirchegger, H (1998) *Pardosa blanda* und *P. torrentum*, zwei Zwillingsarten der *P. monticola*-Gruppe in den Alpen (Araneae: Lycosidae) mit DELTA Datensatz zu den Männchen der Pardosinae Mitteleuropas (with CD-ROM). Diplomarbeit, Univiversität Innsbruck, 1–18.

Sharkey MJ, Yu DS, van Noort S, Seltmann K, Penev L (2009) Revision of the Oriental genera of Agathidinae (Hymenoptera: Braconidae) with an emphasis on Thailand including interactive keys to genera published in three different formats. ZooKeys 21: 19–54. doi: 10.3897/zookeys.21.271

Smith V (2009) Data publication: towards a database of everything. BMC research Notes 2: 113. doi: 10.1186/1756-0500-2-113.

TDWG (2007 onwards) TDWG: standards. Biodiversity Information Standards. http://www.tdwg.org/standards/ [accessed 31 August 2009]

Walter DE, Winterton, S (2007) Keys and the Crisis in Taxonomy: Extinction or Reinvention? Annual Review of Entomology 52: 193–208. doi: 10.1146/annurev.ento.51.110104.151054

Watson L, Dallwitz MJ (1981) An automated data bank for grass genera. Taxon 30: 424–429 + 2 microfiche.

Watson L, Dallwitz MJ (1988) Grass genera of the world: illustrations of characters, classification, interactive identification, information retrieval. With microfiches, and floppy disks for MS–DOS microcomputers. Research School of Biological Sciences, Australian National University, Canberra.

Watson L, Dallwitz MJ (2004 onwards) The ferns (Filicopsida) of the British Isles. http://delta-intkey.com/britfe/

Yoder MJ, Dole K, Seltmann K, Deans A (2006 onwards) MX, a collaborative Web based content management for biological systematists. http://hymenoptera.tamu.edu/wiki/ [accessed on 26th of August 2009]

Appendix 1.

Listing of directory containing key files for the fruit-fly interactive keys package at http://delta-intkey.com/ffa/ (Carroll et al. 2002). doi: 10.3897/zookeys.21.274.app.1.ik.

See section *Key data files, key program files, primary data files* for details.

Citation: Penev L, Sharkey M, Erwin T, van Noort S, Buffington M, Seltmann K, Johnson N, Taylor M, Thompson FC, Dallwitz MJ (2009) Listing of directory containing key files for the fruit-fly interactive keys package at http://delta-intkey.com/ffa/ (Carroll et al. 2002). doi: 10.3897/zookeys.21.274.app.1.ik. Dataset published in: Zookeys 21: 1–18. doi:10.3897/zookeys.21.274.