Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 83-88.
ISBN 978-88-8303-295-0. EUT, 2010.

Simple matrix keys from Excel spreadsheets

Gregor Hagedorn, Mircea Giurgiu, Andrei Homodi

Abstract — An innovative workflow is presented, leading from a simple character and taxon matrix, prepared in an Excel spreadsheet, to a functioning free-access matrix key, stored in SDD (Structured Descriptive Data) format and presented in the flash-based free-access key player IBIS-ID directly on the web. The subset of matrix key functionality supported by this workflow include categorical and quantitative characters, multiple character states for a given taxon, taxon, character, and state illustrations, and a single character grouping level. Advanced features such as complex character or taxon trees, character dependency, or taxon-specific state images are not supported. The workflow aims to attract new contributors for which the learning curve of special purpose matrix-key editing software is too steep.

Index Terms — identification tools, free-access key player IBIS-ID, DELTA, SDD.

1 Introduction

Identification tools such as free-access (= multi-access) or multi-entry keys that are based on a character × taxon data matrix (i. e. a table with the taxa in one dimension and characters in the other) have certain advantages over single-access keys [1, 2, 3, 4]. However, creating a computer-aided matrix key typically requires learning a special purpose application. This limits the number of matrix-based keys produced by biologists, who tend to produce keys similar to the single-access keys typically encountered in the printed literature.

Many biologists are acquainted with spreadsheet applications, especially Microsoft Excel, to edit tabular data. Unfortunately, the visualization of a simple character × taxon table – for which spreadsheet applications are ideal – is a simplified idealization of a more complex data model. All of taxa, characters, states, and the descriptive matrix cells may have further structure:

1. Taxa may require common and scientific names, web links to taxon pages, images, brief diagnostic text, etc.
2. Characters may require a data type, a list of supported states (i. e.

G. Hagedorn is with the Julius Kühn-Institute, Federal Research Centre for Cultivated Plants, Institute for Epidemiology and Pathogen Diagnostics, Königin-Luise-Str. 19, D-14195 Berlin, Germany, E-mail: gregor.hagedorn@jki.bund.de – M. Giurgiu and A. Homodi are with the Telecommunications Department, Technical University of Cluj-Napoca, Cluj 400027, Romania. E-mail: Mircea.Giurgiu@com.utcluj.ro.

constraining the vocabulary), illustrations, explanatory notes.
3. States may provide illustrations or explanatory notes.

The cells of the matrix may contain multiple values, modifiers, notes, and taxon-specific state or character illustrations. For a character: “flower colour” the cell content may be: “usually pink, sometimes red, or blue (immediately after opening)”; for a character: “stem hairiness” it may be “long (2-5 mm) or medium long (1-2 mm) hairs”. A character having more than one state in a taxa is called “polymorphic” in biology. It may occur as a result of a true genetic polymorphism in a population, environmentally induced phenotypic variation (e. g., occurring within the set of flowers on a single plant), or relatively minor quantitative variation that happens – in the present taxon – to cross the artificially drawn borders of a continuously varying character (such as hairiness).

However, when relatively simple rules are followed it is possible to support a subset of the potential complexity of matrix keys within spreadsheets nevertheless.

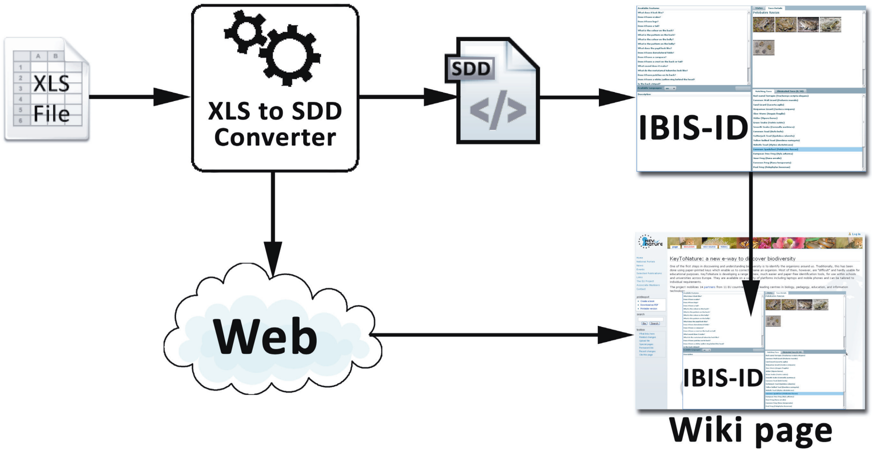
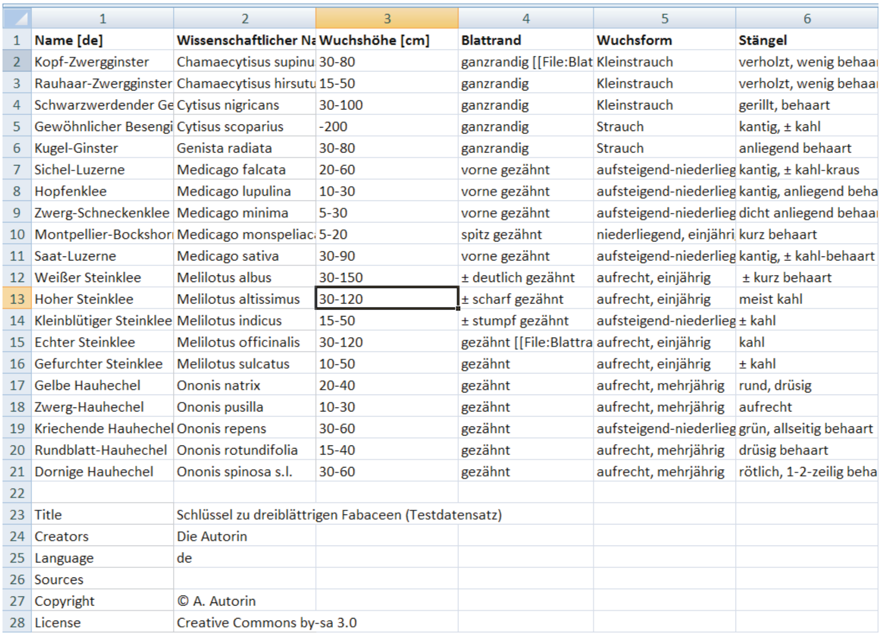


Fig. 1 – The workflow from a Microsoft Excel spreadsheet to SDD conversion and presentation with the free-access key player IBIS-ID inside and automatically published inside a MediaWiki web page.

2 The spreadsheet

The workflow (Fig. 1) starts with the creation of a character × taxon matrix by the biologist, following either instructions on the web [5] or supported by a downloadable template. The simplest layout is indeed one with characters being named in the first row, taxa in the first column and the remainder filled with the taxon × character data (Fig. 2).

For categorical characters (ordinal or nominal scale) the categorical value or state is expressed directly using its label or “name” (e. g., “red”) rather than using a code. In contrast, the DELTA [6] or SDD formats use numeric character and state codes to enforce higher consistency. Multiple states are supported by separating the state names with a semicolon, slash or ampersand (example: “red; blue“). The semicolon is provided as an intuitive delimiter for most

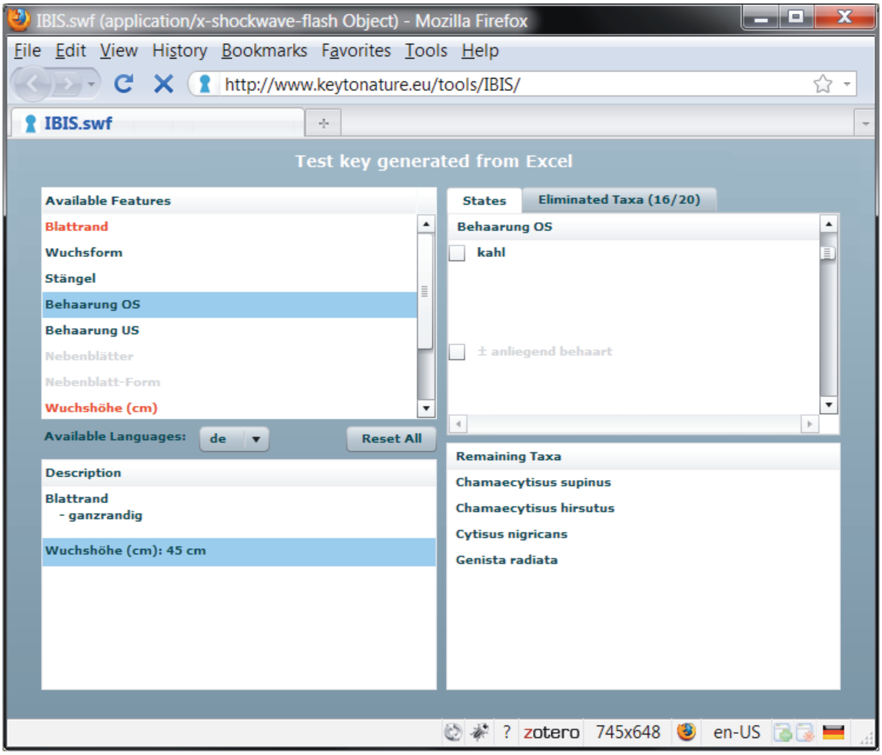


	1	2	3	4	5	6
1	**Name [de]**	**Wissenschaftlicher Na**	**Wuchshöhe [cm]**	**Blattrand**	**Wuchsform**	**Stängel**
2	Kopf-Zwergginster	Chamaecytisus supinu	30-80	ganzrandig [[File:Blat	Kleinstrauch	verholzt, wenig behaa
3	Rauhaar-Zwergginster	Chamaecytisus hirsutu	15-50	ganzrandig	Kleinstrauch	verholzt, wenig behaa
4	Schwarzwerdender Ge	Cytisus nigricans	30-100	ganzrandig	Kleinstrauch	gerillt, behaart
5	Gewöhnlicher Besengi	Cytisus scoparius	-200	ganzrandig	Strauch	kantig, ± kahl
6	Kugel-Ginster	Genista radiata	30-80	ganzrandig	Strauch	anliegend behaart
7	Sichel-Luzerne	Medicago falcata	20-60	vorne gezähnt	aufsteigend-niederlieg	kantig, ± kahl-kraus
8	Hopfenklee	Medicago lupulina	10-30	vorne gezähnt	aufsteigend-niederlieg	kantig, anliegend beha
9	Zwerg-Schneckenklee	Medicago minima	5-30	vorne gezähnt	aufsteigend-niederlieg	dicht anliegend behaa
10	Montpellier-Bockshor	Medicago monspeliac	5-20	spitz gezähnt	niederliegend, einjähri	kurz behaart
11	Saat-Luzerne	Medicago sativa	30-90	vorne gezähnt	aufsteigend-niederlieg	kantig, ± kahl-behaart
12	Weißer Steinklee	Melilotus albus	30-150	± deutlich gezähnt	aufrecht, einjährig	± kurz behaart
13	Hoher Steinklee	Melilotus altissimus	30-120	± scharf gezähnt	aufrecht, einjährig	meist kahl
14	Kleinblütiger Steinklee	Melilotus indicus	15-50	± stumpf gezähnt	aufsteigend-niederlieg	± kahl
15	Echter Steinklee	Melilotus officinalis	30-120	gezähnt [[File:Blattra	aufrecht, einjährig	kahl
16	Gefurchter Steinklee	Melilotus sulcatus	10-50	gezähnt	aufrecht, einjährig	± kahl
17	Gelbe Hauhechel	Ononis natrix	20-40	gezähnt	aufrecht, mehrjährig	rund, drüsig
18	Zwerg-Hauhechel	Ononis pusilla	10-30	gezähnt	aufrecht, mehrjährig	aufrecht
19	Kriechende Hauhechel	Ononis repens	30-60	gezähnt	aufsteigend-niederlieg	grün, allseitig behaart
20	Rundblatt-Hauhechel	Ononis rotundifolia	15-40	gezähnt	aufrecht, mehrjährig	drüsig behaart
21	Dornige Hauhechel	Ononis spinosa s.l.	30-60	gezähnt	aufrecht, mehrjährig	rötlich, 1-2-zeilig beha
22						
23	Title	Schlüssel zu dreiblättrigen Fabaceen (Testdatensatz)				
24	Creators	Die Autorin				
25	Language	de				
26	Sources					
27	Copyright	© A. Autorin				
28	License	Creative Commons by-sa 3.0				

Fig. 2 – A Microsoft Excel spreadsheet with a simple data matrix. Visible are an extra column for scientific taxon names ("Wissenschaftlicher Name" in German), the addition of a measurement unit ("[cm]") in brackets after quantitative characters, state images in the column "Blattrand", and the metadata for the entire dataset or identification tool at the bottom.

biologists; the "/" and "&" to help those who also use DELTA tools.

The drawback of the direct use of state labels is that the vocabulary of available states is not controlled. This is a purposeful design decision. While it is possible to devise spreadsheet layouts that include separate state listings, we have noticed that our test users found all options to be too difficult and confusing and were unable to create them autonomously. The vocabulary control is therefore postponed to the publication of a first draft of the identification tool in the IBIS-ID player. In the implemented workflow, the IBIS-ID key player will make undesirable entries (combinations of states with modifiers or spelling variants of states) transparent and users can modify their data for the next revision. "Normalizing" the state labels is well supported by the typical search-and-replace functionality of spreadsheet software. While careful planning and control is essential for large matrix projects covering hundreds of characters and taxa, the workflow presented here aims at smaller datasets, where a post-data-entry-validation workflow may result in more agile contributions than a plan-ahead workflow.



Available Languages: de

Description

Blattrand
- ganzrandig

Wuchshöhe (cm): 45 cm

Fig. 3 – The resulting interactive matrix key running under IBIS-ID (here in stand-alone mode, not embedded in a web page).

For quantitative characters a DELTA-like encoding is supported. Various combinations of minimum-maximum (in parentheses), “typical range” and a mean are possible (example: “(1-) 3-6.4-8.2 (-15)”).

The characters themselves may be grouped into character groupings by adding the group name surrounded by “{{…}}” after the character label (i.e. in the first row). Presently only a single grouping level is supported.

Matrix cells further support modifiers (“usually”, “rarely”, “about”, “weakly”, etc.) and free-form text comments, if they are enclosed in DELTA-like “<…>” markup.

The spreadsheet rules further provide for the inclusion of multiple taxon, character, and state-specific images. Character and state images must be included in double square brackets after the respective label; in contrast taxon images may be placed in an extra column. Not supported are taxon-×-character and taxon-×-state-specific images.

For taxa, several additional column for images, common and scientific names, and web links can be provided. Finally, metadata for the entire key like creators, title, copyright, license, source may be added.

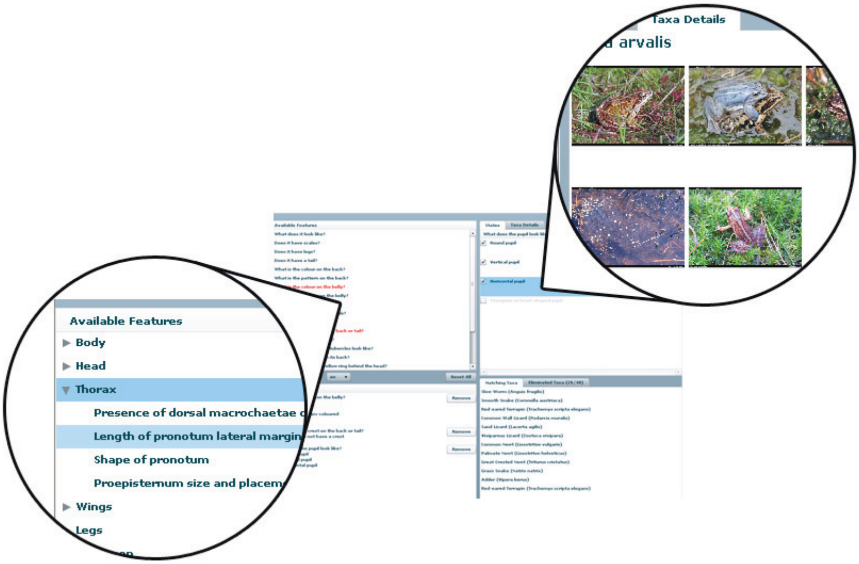

Taxa Details
a arvalis
Available Features
Body
Head
Thorax
Presence of dorsal macrochaetae
Length of pronotum lateral margin
Shape of pronotum
Proepisternum size and placem
Wings
Legs

Fig. 4 – Details of IBIS-ID key player, showing character grouping (left) and state images (right).

3 The converter

The converter is presently a downloadable Microsoft .NET for standalone applications (a web-based version is planned). The converter takes the spreadsheet in Microsoft Excel (XLS) format and converts it into SDD.

The converter supports both wiki-style and direct web image references. Uploading images to the wiki allows users to manage their images for both matrix keys, single-access keys and species pages. The simple wiki-style links are automatically translated by the converter into general web links as they are supported by IBIS-ID key player.

If the converter finds unexpected content, it will report this either as warnings (e. g., *“name contains opening double brackets (‘[[‘) but no closing ones, a malformed image may be present”*) or as errors. Error handling is considered important and efforts have been made to help biological users to understand minor errors. If no errors are encountered, the resulting SDD file will be uploaded to a web repository on the MediaWiki based biowikifarm [7].

Furthermore, to enrich the user experience, a wiki page containing the necessary statement to embed the IBIS-ID player [8] (Fig. 3 and 4) inside a wiki page is also generated.

4 Conclusions

It is possible to replace some features that Excel is missing to directly support matrix keys with rules that rely on simple text delimiters. The method is similar to that used by DELTA the special purpose Windows DELTA editor software.

However, the point of the workflow presented is to provide a simple functionality in an environment well known to most biologist, in order to attract new biologists and educators and increase the production of matrix-based identification tools. Although advanced rules may require some learning effort, it is possible to create useful matrix keys not using these features.

Acknowledgement

This work was supported by the *KeyToNature* Project, ECP-2006-EDU-410019, in the *e*Content*plus* Programme.

References

[1] R. J. Pankhurst, *Practical Taxonomic Computing*, 1991.

[2] J. Winston, *Describing Species*. Columbia University Press,1991.

[3] G. Hagedorn, *Structuring Descriptive Data of Organisms - Requirement Analysis and Information Models*. Ph. D. Thesis, Universität Bayreuth, 2007.

[4] G. Hagedorn, G. Rambold and S. Martellos, “Types of identification keys”. In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems,* pp. *59-64*, 2010.

[5] G. Hagedorn et al., The Excel to SDD converter, 2010. http://www.keytonature.eu/wiki/Excel_to_SDD_converter, 2010-07.

[6] DELTA – DEscription Language for TAxonomy http://delta-intkey.com/, 2010-07.

[7] G. Hagedorn, G. Weber, A. Plank, M. Giurgiu, A. Homodi, C. Veja, G. Schmidt, P. Mihnev, M. Roujinov, D. Triebel, R. A. Morris, B. Zelazny, E. van Spronsen, P. Schalk, C. Kittl, R. Brandner, S. Martellos and P. L. Nimis, “An online authoring and publishing platform for field guides and identification tools”. In: P. L. Nimis and R. Vignes Lebbe (eds.), *Tools for Identifying Biodiversity: Progress and Problems*, pp. 13-18, 2010.

[8] M. Giurgiu, G. Hagedorn and A. Homodi, “IBIS-ID, an Adobe FLEX based identification tool for SDD-encoded multi-access keys”. *Proc. of TDWG 2009*, 9-13 Nov. 2009, Montpellier, p. 90, 2009.