Applications in Plant Sciences

Applications in Plant Sciences 2013 1(6): 1300024

Software Note

A new R package, exsic, to assist taxonomists in creating indices[1]

Reinhard Simon[2,4] and David M. Spooner[3]

[2]Integrated IT and Computational Research Unit, International Potato Center, Avenida La Molina 1895, La Molina, Lima, Peru; and [3]USDA–Agricultural Research Service, Vegetable Crops Research Unit, University of Wisconsin, 1575 Linden Drive, Madison, Wisconsin 53706-1590 USA

- *Premise of the study:* Taxonomists manage large amounts of specimen data. This is usually initiated in spreadsheets and then converted for publication into locality lists and indices to associate collectors and collector numbers from herbarium sheets to identifications (*exsiccatae*). This conversion process is mostly done by hand and is time-consuming, cumbersome, and error-prone.
- *Methods and Results:* We constructed a tool, 'exsic,' based on the statistical software R. The exsic function is part of the R package 'exsic' and produces specimen citations and exsiccatae conforming to four related formats.
- *Conclusions:* The tool increases speed, efficiency, and accuracy to convert raw spreadsheet tables to publication-ready content.

Key words: exsic; exsiccatae; R; reproducible research; software.

A tedious part of preparing a taxonomic monograph is collating specimen data for (1) specimen citations and (2) indices to associate collectors and collector numbers to taxonomic identifications (*exsiccatae*). While preparing a taxonomic monograph of wild potatoes for northern South America, we designed tools to efficiently and accurately convert specimen data into the formats specified by *Systematic Botany Monographs*. One of the tools is a conversion tool to reformat tabulated records into specimen citations and the other for exsiccatae. It is a specific implementation of the reproducible research approach (Gentleman and Temple Lang, 2004). A version of our software tool is freely available for R (R Core Team, 2012) at http://cran.r-project.org/web/packages/exsic/.

METHODS AND RESULTS

***Overall design of the package*—**The package was designed to facilitate very specific steps in the processing of tabular specimen data: the conversion from table-oriented to index-based or list-based formats. Other necessary steps, such as data cleansing or data conversions, are not part of the main functions as there are other tools available for that purpose. That is, date conversions or geographical coordinate transformations into the final desired format must be done separately by the user and before using the exsic function. The main function assumes, therefore, correct structure and content, but has been designed to be robust and to gracefully handle missing columns or content. In the case of essential columns (for sorting or filtering, e.g., "species", "country", "collcite", "number", "majorarea", "minorarea"), these will be added and filled with meaningful values for missing content (e.g., "Anonymous" for missing collector information). However, a few convenience functions have been added to handle specific transformations, and these are explained in more detail below.

***Tabular data format*—**As a starting point for preparing the specimen data, we used a format based on conventions defined by the BRAHMS (Botanical Research and Herbarium Management Software) package (Filer, 2001), as detailed in an online manual (Filer, 2010). The column names used by exsic are mostly the same as those used by BRAHMS, with a few modifications and additions to facilitate both the concurrent use of the BRAHMS conventions and this package (Table 1). One is the addition of the "colldate" field for the "date of collection". The content of this field should usually have the format "1 Jan 2013"; the corresponding BRAHMS field names are: "colldd", "collmm", and "collyy" for day, month, and year, respectively. Functions in the package "date" can be used to convert from these three columns to the desired format and stored in the column "colldate". The other addition is the column or field "collcite". The information on citing collectors is stored in two BRAHMS fields: "collector" and "addcoll". A custom function ("coll.cite") is provided that helps address most formatting issues, including whether or not to use initials, whether or not to use periods after initials, and whether to cite both names for collector pairs or cite only the first and add "et al." when there are more than two. The variations can be set using "coll.cite" function parameters. Another convenience function is called "strip.last.dot"; this is meant to process the column "locnotes" for location notes because they are separated by a comma in most specimen citations. A last variation is the use of a separate column for "phenology", with expected content to be "fl" for flowering stage, "fr" for fruiting stage, or empty for neither stage.

***Target formats for specimen citations of numbered collections*—**We researched format variations using a Google search and found three additional formats to that used in *Systematic Botany Monographs*. The four formats (*Systematic Botany Monographs* [format.SBMG], American Society of Plant Taxonomists [format.ASPT], New York Botanical Garden [format.NYBG], and *PhytoKeys* [format.PK]) are summarized in Table 2. The main differences between these specimen citation formats relate to the formatting of species and country information as well as to differences in formatting the collector information. New York Botanical Garden journals include phenology information after the date of observation. Numbered collections formats differed with respect to using index numbers to point to the species or directly using a species. Another difference is whether consecutive numbers of the same species for a

[1]Manuscript received 28 March 2013; revision accepted 22 May 2013.

This research was supported by the National Science Foundation (DEB 0316614). The authors thank two anonymous reviewers for their feedback on earlier versions, which greatly helped to improve the scope and usability of the tool.

[4]Author for correspondence: r.simon@cgiar.org

doi:10.3732/apps.1300024

Table 1. The data dictionary for the primary table of specimens. Field or column names correspond largely to BRAHMS standard; sample content data are invented.

Exsic column name	BRAHMS field name	Exsic required	Description	Type	Sample content
id	NA	Obligatory	Continuous number from 1 to n	Integer	1
genus	genus	Obligatory	Genus	Text	Solanum
species	sp1	Obligatory	Species	Text	tuberosum
collector	collector	Obligatory	Collector name	Text	Linne, C
number	number	Obligatory	Collector number identifying the specimen	Integer	1111
addcoll	addcoll	Recommended	Additional collector names	Text	Author, A
collcite	NA	Obligatory	Final citation of collector(s)	Text	Author
dups	dups	Recommended	Duplicated herbaria	Text	PAR
majorarea	majorarea	Recommended	Major subnational level	Text	Puno
minorarea	minorarea	Recommended	Second subnational level	Text	Puno
locnotes	locnotes	Recommended	Location	Text	at the shore of lake Titicaca
altitude	alt	Optional	Elevation in meters	Text	4000 m
latitude	lat	Optional	Latitude	Text	12°1′23″N
longitude	long	Optional	Longitude	Text	12°1′23″E
colldate	NA	Recommended	Collection date	Text	1 Jan 2013
country	country	Obligatory	Country of origin	Text	Peru

Note: NA = not applicable.

collector citation are grouped or not. The report created by the command example (exsic) creates a Web page as shown in Fig. 1. The HTML page can then be further edited and formatted using a word processor. A user's guide showing the principal usage along with a precise description of the input format can be found in the package documentation (http://cran.r-project.org/web/packages/exsic/exsic.pdf).

***Sorting and Filtering*—**The specimen citation records are, by default, not filtered from the given table, and species and countries are sorted alphabetically. However, a taxonomist may want to order species by taxonomic relationships, and countries may be ordered from north to south and west to east. Also, the same table may be used for checking subsets. Therefore, to allow both filtering and custom sorting "on-the-fly," an additional table (an R data.frame) can be defined. It has two columns, "country" and "species", where the desired countries and species may be listed separated by semicolons without spaces. To list all countries or all species, the word "all" may be used. The table may be added to the exsic function as a parameter "sortfilter". If this parameter is omitted, default options apply. A helper function to check format compliance is available (is.sortfilter). A sample "sortfilter" table is also available (sort.specs).

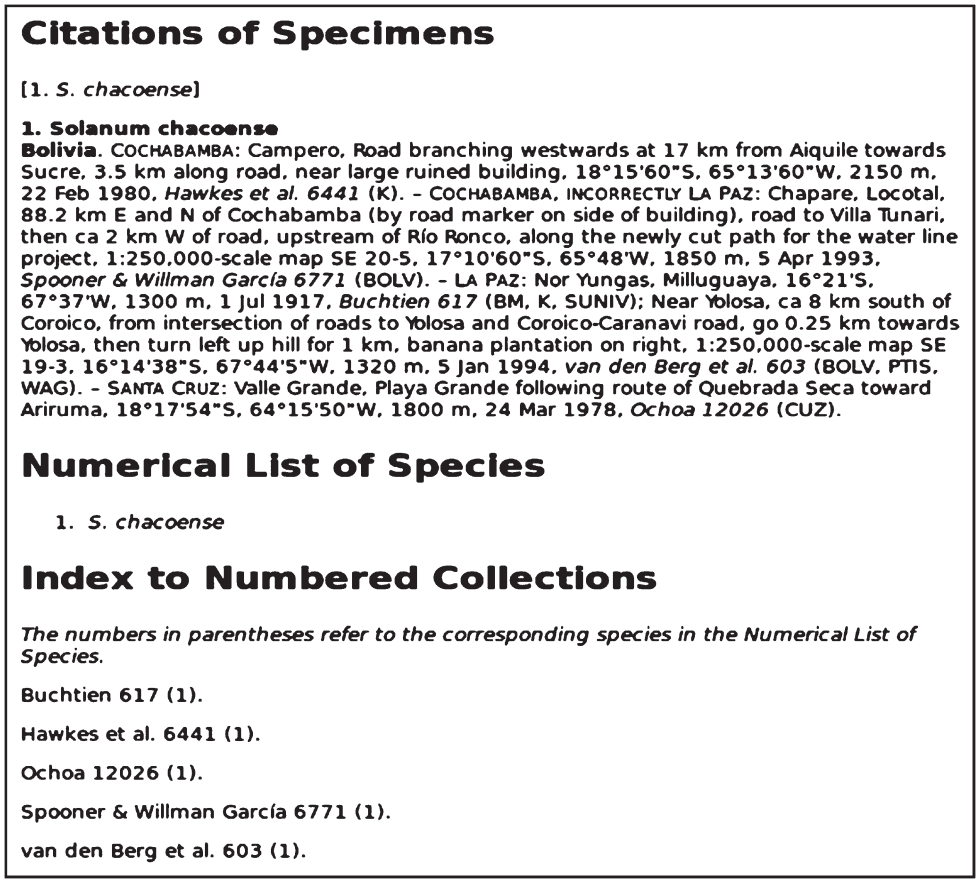
***Usage*—**The package can be downloaded from the central R repository CRAN using standard procedures in R and activated using the command "library(exsic)". A user would typically start using the main exsic function and either provide an in-memory table or a file path. For example: a table may be read using the custom function "data <- read.exsic(filepath)". This function ensures that all obligatory fields (see Table 1) are present; if not, they will be created and prefilled with placeholder text (e.g., missing collector citation in "collcite" will be replaced with "Anonymous"; missing collection numbers with "s.n."; missing dates with "s.d."; and missing minor area or major area information with "Unknown major/minor area"). An example table with 1000 records of wild potato specimens is included in the package and can be accessed using "system.file("samples/exsic.csv", packages="exsic")". This may take a minute to process. For quicker testing, a subset can be constructed using "pt = potato[1:10,]" which will use only the first 10 records. A complete minimal example using the provided sample table is given here:

```
afile <- system.file("samples/exsic.csv", package="exsic")
potato <- read.exsic(afile)
pt = potato[1:10,]
exsic(pt)
# or shorthand for the whole table:
exsic(file="afile")
```

Table 2. A comparison of the four index formats detailing the fields used in the tables format.SBMG, format.ASPT, format.NYBG, and format.PK.

	Systematic Botany Monographs		*Systematic Botany*		New York Botanical Garden publications		*PhytoKeys*		
Field	Style	Sept[a]	Style	Sept[a]	Style	Sept[a]	Style	Sept[a]	Required[b]
species	bold		capitals		bold		bold		obligatory
country	bold	.	none	.	bold;uppercase	.	uppercase	.	obligatory
majorarea	capitals	:	none	:	bold	:	none	:	obligatory
minorarea	none	,	none	,	none	,	none	,	obligatory
locnotes	none	,	none	,	none	,	none	,	optional
latitude	none	,	none	,	none	,	none	,	optional
longitude	none	,	none	,	none	,	none	,	optional
altitude	none	,	none	,	none	,	none	,	optional
colldate	none	,	none	,	none		none	,	optional
phenology[c]					()	,			optional
collcite	italics		italics		underline		none		obligatory
number	italics		italics		underline		none		obligatory
dups	()	;	()	;	()	;	()	;	optional
group.majorarea	yes	—	no	—	no	—	no	—	obligatory
species.referral	();number	,	();number	,	();number	;	();name	;	obligatory
group.specimens	yes	—	yes	—	yes	—	yes	—	obligatory

[a] Sept = separator punctuation.
[b] Required field in exsic.
[c] The "phenology" field is only used in one format.



Citations of Specimens

[1. *S. chacoense*]

1. Solanum chacoense
Bolivia. COCHABAMBA: Campero, Road branching westwards at 17 km from Aiquile towards Sucre, 3.5 km along road, near large ruined building, 18°15'60"S, 65°13'60"W, 2150 m, 22 Feb 1980, *Hawkes et al. 6441* (K). – COCHABAMBA, INCORRECTLY LA PAZ: Chapare, Locotal, 88.2 km E and N of Cochabamba (by road marker on side of building), road to Villa Tunari, then ca 2 km W of road, upstream of Río Ronco, along the newly cut path for the water line project, 1:250.000-scale map SE 20-5, 17°10'60"S, 65°48'W, 1850 m, 5 Apr 1993, *Spooner & Willman García 6771* (BOLV). – LA PAZ: Nor Yungas, Milluguaya, 16°21'S, 67°37'W, 1300 m, 1 Jul 1917, *Buchtien 617* (BM, K, SUNIV); Near Yolosa, ca 8 km south of Coroico, from intersection of roads to Yolosa and Coroico-Caranavi road, go 0.25 km towards Yolosa, then turn left up hill for 1 km, banana plantation on right, 1:250,000-scale map SE 19-3, 16°14'38"S, 67°44'5"W, 1320 m, 5 Jan 1994, *van den Berg et al. 603* (BOLV, PTIS, WAG). – SANTA CRUZ: Valle Grande, Playa Grande following route of Quebrada Seca toward Ariruma, 18°17'54"S, 64°15'50"W, 1800 m, 24 Mar 1978, *Ochoa 12026* (CUZ).

Numerical List of Species

1. *S. chacoense*

Index to Numbered Collections

The numbers in parentheses refer to the corresponding species in the Numerical List of Species.

Buchtien 617 (1).

Hawkes et al. 6441 (1).

Ochoa 12026 (1).

Spooner & Willman García 6771 (1).

van den Berg et al. 603 (1).

Fig. 1. An example result for the exsic function using five records from a sample table.

The resulting Web page will be in the same directory called "exsic.html" (this name can be changed via a parameter).

By default, the whole table passed to exsic will be used, sorted alphabetically by species and country, and formatted according to the conventions of *Systematic Botany Monographs*. To see an example, one may also type "example(exsic)" in the R console. The user may modify the defaults by combinations of the following five options: (a) choose of one of four formats, (b) select a subset of species and countries, and order nonalphabetically by species or country, (c) define additional formats, (d) set section titles and the output file name, and (e) selectively execute only some of the subroutines. These options are explained below.

A. Choice of formats: The main exsic function has an additional parameter, "formats", that accepts tabulated parameters. The table must have only four columns named exactly as "field", "style", "sept", "comments" (for example contents see Table 2). The four formats are listed in Table 2. The default format is "format.SBMG".

B. The "sortfilter" is a parameter in the form of a table that must have only two columns, named "country" and "species". Within each field, countries or species must be written exactly as in the primary table and must be separated by semicolons without spaces. Only those countries or species recognized will be used in the final indices and will be sorted according to the provided sequence.

C. Advanced users may define their own format conventions using one of the "format.XXXX" (where XXXX is either SBMG, ASPT, NTBG, or PK) examples as a starting point. The formats are defined in simple tables so that they can be edited using spreadsheet software. The available formats for text include: bold, italic, underline or underscore (no difference), capitals, uppercase, parentheses, or square brackets. These formats can be combined and are applied from left to right; several options must be separated by a semicolon without a space. Unrecognized formatting words are ignored. For example, the formatting "();italics" will result in italicized parentheses whereas "italics;()" will not. More details can be found in Table 3.

D. The parameter "header" can be used to set other section titles; the parameter "out.file" is used to set the output file name.

TABLE 3. A comparison of the four index formats showing examples of how to create the format using the exsic function.

Field in exsic data table	*Systematic Botany Monographs*	*Systematic Botany*	New York Botanical Garden publications	*PhytoKeys*
Example exsic command[a]	exsic(rec, format = format.SBMG)	exsic(rec, format = format.ASPT)	exsic(rec, format = format.NYBG)	exsic(rec, format = format.PK)
Exsic formatted record	**Argentina.** CATAMARCA: Ambato, Sierra de Ambato, 27°42′S, 65°55′60″W, 3500 m, 22 Feb 1971, *Hunziker 20938* (CORD).	Argentina. Catamarca: Ambato, Sierra de Ambato, 27°42′S, 65°55′60″W, 3500 m, 22 Feb 1971, *Hunziker 20938* (CORD).	**ARGENTINA**. **Catamarca**: Ambato, Sierra de Ambato, 27°42′S, 65°55′60″W, 3500 m, 22 Feb 1971 (fl), Hunziker 20938 (CORD).	ARGENTINA. Catamarca: Ambato, Sierra de Ambato, 27°42′S, 65°55′60″W, 3500 m, 22 Feb 1971, Hunziker 20938 (CORD).

[a]The "rec" variable is the first record from the table. It can be created using: rec = read.exsic(system.file("samples/exsic.csv", package="exsic"))[1,]

E. The building blocks of the main exsic function are also available individually along with some helper functions. Each index function can be executed separately. Output format is in an intermediate format called "markdown" format and needs to be converted. This can be achieved using the function "write.exsic". Headers can be provided by using "exsic.header" and a combination of string concatenating commands such as "paste". For example, a custom index could be achieved using:

```
hdr = exsic.header("A header")
idx = index.collections(pt)
txt = paste(hdr, idx, sep="")
write.exsic(txt, "idx.html")
```

For more details, see the package documentation.

***Implementation and speed*—**We used R and two libraries: (1) stringr (Wickham, 2012) and (2) markdown package (Gruber and Swartz, 2004; Allaire et al., 2012) as a basis to implement the exsic package. The exsic package shows one application of R and reproducible research tools for taxonomists and botanists. The package is freely available under the open source GNU Public License (GPL) and for all platforms supported by R (currently Windows, Linux, and Mac OS). On a Dell T7400 precision PC with 4 GB of RAM, a 2.66-GHz Intel Xeon CPU E5430 processor, and running R 2.15.2 on top of Ubuntu 12.10, the sample table with 1000 records was processed in about 30 s. The function has also been tested on Windows XP and Mac OS Snow Leopard.

Important note—When working across operating systems with tables created in Excel, it is indispensable to make sure that data are saved not only in .csv format but also using the encoding standard UTF-8. This allows the use of accents or other alphabets in the indices. Excel does not use UTF-8 as a default, and this will result in formatting errors.

CONCLUSIONS

The tool primarily increases the speed of preparing specimen citations, numbered collections, and supporting indices. It also minimizes human errors in manual transcription from table to list formats as well as formatting errors. The possibility to quickly create indices in familiar formats also provides an opportunity to double check the consistency and completeness of the table before final publication, thereby increasing the final quality of the table and interpretation. We are not aware of a similar freely available tool except for the report module in the BRAHMS software version 7 that also facilitates the generation of these indices. While the BRAHMS software is Windows only, the exsic package works on Windows, Linux, and Mac OS.

LITERATURE CITED

Allaire, J. J., J. Horner, V. Marti, and N. Porte. 2012. Markdown: Markdown rendering for R. Website http://cran.r-project.org/web/packages/markdown/ [accessed 24 May 2013].

Filer, D. L. 2001. BRAHMS—Botanical research and herbarium management system. *BioNET News* 8: 6–7.

Filer, D. L. 2010. BRAHMS—Botanical Research and Management System: Training course guide. Website http://herbaria.plants.ox.ac.uk/bol/Content/Groups/brahms/Resources/BRAHMS%202010%20training%20guide.pdf [accessed 24 May 2013].

Gentleman, R., and D. Temple Lang. 2004. Statistical analyses and reproducible research. Bioconductor Project Working Papers. Working Paper 2. Website http://biostats.bepress.com/bioconductor/paper2 [accessed 24 May 2013].

Gruber, J., and A. Swartz. 2004. Markdown 1.0.1. Website http://daringfireball.net/projects/markdown/ [accessed 24 May 2013].

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website http://www.R-project.org/ [accessed 24 May 2013].

Wickham, H. 2012. stringr: Make it easier to work with strings. Website http://cran.r-project.org/web/packages/stringr/ [accessed 24 May 2013].