Published by Associazione Teriologica Italiana
Hystrix, the Italian Journal of Mammalogy
Volume 29 (2): 155–161, 2018
OPEN ACCESS
Available online at:
http://www.italian-journal-of-mammalogy.it
doi:10.4404/hystrix-00133-2018

Commentary

Good practices for sharing analysis-ready data in mammalogy and biodiversity research

Luis D. Verde Arregoitia[1,*], Natalie Cooper[2], Guillermo D'Elía[1]

[1] *Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Campus Isla Teja s/n, Valdivia, Chile*
[2] *Department of Life Sciences, Natural History Museum London, Cromwell Road, London, UK*

Keywords:
biodiversity informatics
data rectangling
specimen-based research
spreadsheets

Article history:
Received: 20/09/2018
Accepted: 13/12/2018

Acknowledgements
Two anonymous reviewers and the editorial team provided insightful feedback that greatly improved this manuscript. We also wish to thank Jenny Bryan and Duncan Garmonsway for their open-source work on data organisation. LDVA received support from Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECyT) Project 3170246. GD received support from FONDECyT Project 1180366.

Abstract

As both producers and consumers of data, scientists play an important role in defining how accessible their research outputs are to others. First by deciding to share, but also through the choice of file formats and data structures used to share data. Steps taken by authors, editors, and typesetters to format and store data often complicate the ability of future users to work with these data. At late stages of the scientific workflow, making analysis-ready versions of the data takes relatively little time and effort in exchange for a significant increase in usability and, potentially, other well-known benefits of data sharing such as more citations and potential collaborations. Well-structured and analysis-ready data also reduces the risk of unintended alterations introduced while cleaning and rearranging published data. We wish to reconcile what is easy to read and intuitive with machine-readable data that does not need extensive processing or advanced programming skills for inclusion in new analyses. For those who use and report biodiversity data and the results of specimen-based research, we wish to create awareness of the major differences in structure between data at the analysis stage compared with data arranged and formatted for reporting. We hope that the reader might apply these practices when sharing data with other scientists and with the public.

Introduction

For many areas of research in mammalogy and biodiversity science as a whole, computational analysis plays a central role and large amounts of data are continuously being collected and analysed. Journals, funders, and researchers in general are recognizing that sharing information facilitates science and that published research should include its associated data (Wallis et al., 2013). Even without added incentives or mandates to share, a wealth of data already gathered from field studies, literature reviews, natural history collections, and laboratory analyses are published and available (Reichman et al., 2011). This is evident in the number of published datasets that collate large amounts of biodiversity data (Tab. 1). There is also a growing practice of reusing, aggregating, and repurposing heterogeneous data from multiple studies (Lowndes et al., 2017). Meta-analyses are more frequently conducted (e.g. Auer et al., 2017), and new developments in molecular and computational tools allow us to revisit existing morphological data, ask new questions, and analyse trait evolution along phylogenetic trees using modern comparative methods (e.g. Schweizer et al., 2014). With this increase in the potential for data reuse, it is vital to make data available, easier to parse, and less prone to import errors. When data can be easily imported and manipulated using familiar software (either using scripting languages, or any program that can import common file formats), it becomes much easier to reuse and will have a greater impact (Hart et al., 2016).

Some barriers to effective data sharing are deeply rooted in the practices and culture of the research process as well as the researchers themselves (Tenopir et al., 2011). However, once scientists are willing to share their data, there are practical barriers to overcome that relate to common conventions and practices in data sharing. Popular ways of structuring and presenting data can inadvertently place important content beyond the immediate reach of those who want to collate or reanalyse it. This ultimately requires a substantial amount of time and effort invested in importing and re-organizing these data manually (Dasu and Johnson, 2003). Manual input and structuring of data can also lead to nonreplicable user-generated errors such as inadvertently omitting, altering, or duplicating data (Reschenhofer and Matthes, 2015).

Here we describe some common issues, along with simple suggestions authors can follow to make their data more reusable, and thus increase the impact of their work. We address specific practices that often create bottlenecks and reduce the reuse potential of data. All the examples are drawn from our experience in mammalogy, but these principles can apply to other types of biodiversity research. For more technical general guides on sharing data, see Ellis and Leek, 2018 and White et al., 2013. We wish to bring attention to simple low-effort practices that would help propagate analysis-ready data, and to help others structure information in such a way that it can be coerced easily into a structure and format that facilitates downstream analyses and synthesis. These recommended practices apply to data shared as appendices, supplementary files, or uploaded separately to repositories (e.g. DataDryad, figshare, GBIF), but also to relatively small tables shared as part of theses, reports, or publications. These tables often contain useful information that others may want to reuse, despite not typically showing raw, unprocessed values.

Data structures

Working with data from multiple different publications is a much smoother process if they share a common structure and format. To this end, the practices presented here aim (whenever possible) to make shared data compatible with 'tidy data' principles (Wickham, 2014). In tidy data, each variable must have its own column, each observation must have its own row, and each value must have its own cell (Fig. 1). Note that a tidy data structure is often not ideal for reporting. Cross-tabulations and other frequency tables may be far better choices for the presentation of summaries, relationships, and patterns, and they

*Corresponding author
Email address: luis@liomys.mx (Luis D. Verde Arregoitia)

doi:10.4404/hystrix-00133-2018

a)

levels of the 'Habitat' variable

	Forest	Mangrove
Pseudomys sp.	5	13
Hydromys sp.	1	3
Melomys sp.	8	2

values of the 'Captures' variable

b)

variables

Rodent	Habitat	Captures
Pseudomys sp.	Forest	5
Hydromys sp.	Forest	1
Melomys sp.	Forest	8
Pseudomys sp.	Mangrove	13
Hydromys sp.	Mangrove	3
Melomys sp.	Mangrove	2

observations

Figure 1 – Hypothetical example of rodent trapping data in two habitat types. The two data models convey the same information. The tabulated habitat vs. species format (a) is concise and compact, but the tidy structure (b) is more versatile for manipulation and visualization.

can help us process and digest information better than tidy data. What makes 'untidy' data problematic for further analysis is when there is no easy way of importing or coercing it to a tidy structure. Because of its versatility, tidy data is ready for immediate use. Filtering, grouping, transforming, sorting, aggregating, visualizing, and modeling are greatly simplified when working with tidy data. We can easily change tidy data into numerous useful formats and structures. Tidy data is more repetitive and takes up more space than other more condensed representations, so it may not be ideal for data entry or when preparing tables that will be embedded in a text. The ideas presented here are meant to guide those sharing data towards knowing when it is best to provide tidy data, and also knowing how to share succinct tables that can be easily 'tidied' afterwards.

Data rectangling

For most of the data that we collect, analyse, and report, the best layout is often a single dataset with rows corresponding to subjects and columns corresponding to variables, or a 'data rectangle' (Broman and Woo, 2018). The first row in a data rectangle should contain the variable names (the header row), and header rows should only appear once (Broman and Woo, 2018). Fig. 2 shows one of many ways in which data are often presented using a non-rectangular structure. In Fig. 2a: empty cells, subheaders, and table formatting are used to organise data into an array of smaller non-rectangular datasets. See Broman and Woo, 2018 for more examples of non-rectangular data. Other common examples of non-rectangular data include unstructured text, spreadsheets holding multiple disparate tables, nested lists, or more complex data structures such as JavaScript Object Notation (JSON) files.

Good practices

The following recommended practices for making data easier to reuse are presented in no particular order as they are equally important. Guides for better data entry and organization are now available (e.g. Michener, 2015; Wilson et al., 2017), and we suggest following best practices in general since they streamline research and lead to better data for the reporting stage. Our recommended good practices are mostly meant for implementation in the final stages of data-sharing just prior to publication, but earlier implementation during the research workflow will make data more reusable even when it is not meant for publication. During early stages of data collection and analysis, any format that is readable and works for our specific purposes may be used, as long as we remain aware of tidy data principles. To avoid errors and preserve the integrity of existing datasets, we suggest adopting these principles for future projects, rather than restructuring existing data before sharing. Although it is not our main objective, we suggest functions from different R packages to address some of these specific issues when importing and manipulating existing datasets. Although we focus on R, other scripting languages are also well-suited for cleaning and transforming data, such as Julia and Python. We recognise the learning curve involved, but wish to emphasise that rearranging data is best accomplished with code-based routines that can be run many times, and keep a record of what we did to the data. Errors can also happen with code-based data manipulation, the difference is that errors are reproducible when they come from scripts so they can be identified and fixed. Managing data via programming allows us to also work with massive datasets efficiently (for example: thousands or hundreds of thousands of records), a difficult task when having to click or scroll through the data using spreadsheet software (Baumer et al., 2017).

Table 1 – Examples of recent aggregate datasets describing multiple dimensions of biodiversity.

Dataset name	Content	Reference
BioTIME	biodiversity time series data with ~12 million records	Dornelas et al., 2018
TetraDENSITY	population density estimates for terrestrial vertebrates with >18k records	Santini et al., 2018
ATLANTIC MAMMAL TRAITS	morphological traits of mammals in the Atlantic Forest of South America	Gonçalves et al., 2018
PHYLACINE	phylogeny, distribution, and trait data for 5,831 species of mammals (living and extinct)	Faurby et al., 2018

a)

Summer 2012				
hedgehog				
	Lakeside	Urban	Forest	Park
ear tag	5	2	8	3
	Lakeside	Urban	Forest	Park
mass (kg)	0.45	0.61	0.36	0.47
	Lakeside	Urban	Forest	Park
age class	adult	adult	juvenile	adult
pine marten				
	Lakeside	Urban	Forest	Park
ear tag	1	1	7	4
	Lakeside	Urban	Forest	Park
mass (kg)	1.1	0.81	0.76	0.43
	Lakeside	Urban	Forest	Park
age class	adult	juvenile	adult	pup

b)

animal	site	ear tag	mass (kg)	age class	season	year
hedgehog	Lakeside	5	0.45	adult	Summer	2012
hedgehog	Urban	2	0.61	adult	Summer	2012
hedgehog	Forest	8	0.36	juvenile	Summer	2012
hedgehog	Park	3	0.47	adult	Summer	2012
pine marten	Lakeside	1	1.1	adult	Summer	2012
pine marten	Urban	1	0.81	juvenile	Summer	2012
pine marten	Forest	7	0.76	adult	Summer	2012
pine marten	Park	4	0.43	pup	Summer	2012

Figure 2 – Data for a hypothetical mammal survey in a non-rectangular structure (a) compared with the same data represented as a tidy two-dimensional rectangle (b).

Use explicit and consistent delimiters for text-based specimen lists

Collection materials (including fossils) can be assessed in many ways (e.g. photographed, scanned, x-rayed, measured, or sampled for genetic material), which leads to a significant amount of derived data which must be associated with their source. Therefore, reporting adequate identifying information and the depository of the material studied is critical. The practice of listing the collections material ex-

a)

Cavia aperea MLP 151, 523, 542.M15,573.3; Galea leucoblephara MLP 676, 738.4, 6.XII.35.2, MACN 34.193, 15324; Galea sp. MACN 31.30,36.419

Separates taxa

Separates specimens

Different collection not explicitly delimited

b)

Taxa	Specimen
Cavia aperea	MLP 151
Cavia aperea	MLP 523
Cavia aperea	MLP 542.M15
Cavia aperea	MLP 573.3
Galea leucoblephara	MLP 676
Galea leucoblephara	MLP 738.4
Galea leucoblephara	MLP 6.XII.35.2
Galea leucoblephara	MACN 34.193
Galea leucoblephara	MACN 15324
Galea sp.	MACN 31.30
Galea sp.	MACN 36.419

Figure 3 – a) Specimen list shown as inline text, compared with the same data redrawn in tabular format (b). Taxa and specimens are sampled from the full list provided in Álvarez et al., 2011.

amined within a manuscript dates back to the early stages of formal collections-based research. For example: Wilfred H. Osgood examined over 27,000 specimens for his work on the taxonomy of *Peromyscus* deer mice (Osgood, 1909) and included location data for the specimens examined. Much later, Ruedas et al., 2000 explicitly called for authors to list their specimens examined and provide a minimum set of details (scientific name, individual specimen identifier, name of collection, locality, and gene accession number for sequences obtained and used). This is particularly important, as Troudet et al., 2018 have documented a general decline in the connection between biodiversity data with tangible specimen material. Either through tradition or because of editorial policies in academic journals, authors usually provide lists of specimens examined to maintain the link between specimens and biological studies. However, to keep specimen lists within a limited amount of space, this information is often collapsed into continuous inline text and shortened using seemingly intuitive practices. For example: collapsing consecutive numbers in a series, or mentioning a collection only once with the implication that the specimen numbers that follow correspond to it until a different collection is mentioned (attribution by adjacency). The notation used by authors to separate collections, taxa, and specimen numbers, provenance, and the type of data gathered from the given specimen often varies across journals and from study to study. At the same time, explicit explanations of how the different list items are separated are only rarely provided. Fig. 3 is derived from a list of rodent specimens examined by Álvarez et al., 2011 and follows the notation used by the authors. As inline text, species are separated by semicolons and specimens are delimited by commas. When specimens from the same species come from more than one collection, there is no explicit separator. The contrasting redundancy and use of space on a page is evident when comparing inline text (Fig. 3a) with the same information as a data rectangle (Fig. 3b).

Recommendation

Lists of examined material tend to come from files or notes that authors keep on paper, relational databases, or spreadsheets in the first place. We suggest sharing these tables directly in appendices or as supplementary material in addition to the inline text embedded in a publication. Tabular specimen data allow us to easily generate summary statistics (e.g. number of specimens by taxa or by collection), or to combine specimen lists from multiple studies. Additionally, missing information is more easily detected in tabular format and this could help avoid accidental omissions or typing mistakes. We discourage the use of periods as delimiters, since they often appear within specimen IDs (e.g. MV14.481.1), when abbreviating genera (e.g. *S. lilium* instead of *Sturnira lilium*), when referring specimens to an unidentified species (e.g. Sturnira sp.), or as punctuation at the end of lists. We also discourage using special symbols to code male and female specimens, since these may be garbled when files get encoded and decoded and important information could be lost. If a specimen list is required as inline text, we suggest the following format:

- consistent delimiters with an explicit explanation;
- periods not used as delimiters;
- no interspersed grouping data (taxonomic, geographic);
- attention to series and consecutive numbers;
- avoid special unicode symbols (♂♀) for representing sex (unless there is a strict nomenclatural reason to use them), instead code sex as M/F, explaining what each letter represents.

In the following example built from data provided in Hoffmann and Baker, 2001, the list of examined specimens has consistent delimiters and separators that are explained in the preceding description of the specimen list. This text structure can easily be coerced into tabular form.

Example specimen list: TK numbers correspond to samples from the Natural Science Research Laboratory from Texas Tech University, Lubbock; FMNH numbers correspond to samples from the Field Museum of Natural History, Chicago, and NMNH numbers correspond to samples from the National Museum of Natural History, Washington D.C. Species and their respective specimens are separated by dashes (-), species are delimited by semicolons (;), specimens by commas (,) and collection abbreviations for the same taxon are delimited by forward slashes (/). We indicate series of consecutive specimen IDs with the preposition "to".

Glossophaga longirostris - TK 18501, TK 18585, TK 18613, TK 18667, TK 25150/ NMNH 580656, NMNH 580658; *Glossophaga morenoi* - TK 20563, TK 20564, TK 20579; *Glossophaga soricina* - TK 34707, TK 41573, TK 9251, TK 11040, TK 4728/ NMNH 578997, NMNH 579009, NMNH 579010/ FMNH 128675 to FMNH 128681

Studies that use specimen data in a spatial context (e.g. spatial variation in phenotypes, phylogeography, biogeography, species descriptions, notes on distributional records, etc.) tend to also include sex, country, and locality data in the specimen lists. Below is an excerpt of specimen data provided in Lim et al., 2003. Although this information could be delimited and ultimately wrangled into tabular form, we argue that too many variables presented as inline text makes for very dense data and risks running into inconsistencies with delimiters and nested information. Those who wish to share relevant information for each specimen could do so in tabular form, even if it is done separately. In tabular form, it is possible and straightforward to, among other things: plot spatial point occurrences on a map, subset by location or sex, or aggregate this information with other studies.

Example of a data-dense specimen list: *Vampyressa thyone* - Belize: Rockstone Pond, 17°46′ N, 88°22′ W, 1 ♀(ROM 33614). Colombia: Antioquia; Remedios, Finca San Martin, 7°2′ N, 74°41′ W, 1 ♂(ROM 84983). Cauca; Mechenguito, 2°40′ N, 77°12′ W, 1 ♀(ROM 63213).

a)

header

subheaders

Scientific name	Common name	Red List status
Leporidae		
Brachylagus idahoensis	Pygmy rabbit	Least Concern
Caprolagus hispidus	Hispid hare	Endangered
Ochotonidae		
Ochotona alpina	Alpine pika	Least Concern
Ochotona iliensis	Ili pika	Endangered
Ochotona princeps	American pika	Least Concern

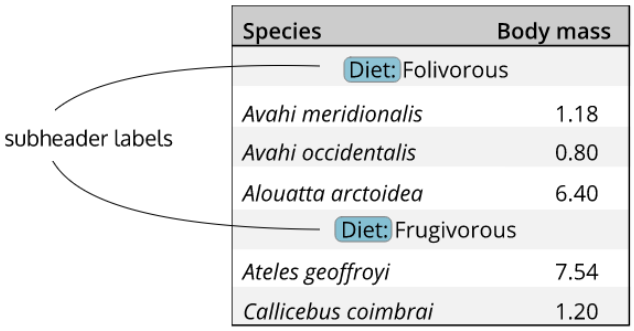
subgroups implied by adjacency

b)

Scientific name	Common name	Red List status	Family
Brachylagus idahoensis	Pygmy rabbit	Least Concern	Leporidae
Caprolagus hispidus	Hispid hare	Endangered	Leporidae
Ochotona alpina	Alpine pika	Least Concern	Ochotonidae
Ochotona iliensis	Ili pika	Endangered	Ochotonidae
Ochotona princeps	American pika	Least Concern	Ochotonidae

Figure 4 – Data with values of a grouping variable embedded as subheaders (a), compared with data structured so that the grouping variable appears in its own column (b).

a)



b)

Species	Body mass	Diet
Avahi meridionalis	1.18	Folivorous
Avahi occidentalis	0.80	Folivorous
Alouatta arctoidea	6.40	Folivorous
Ateles geoffroyi	7.54	Frugivorous
Callicebus coimbrai	1.20	Frugivorous

Figure 5 – a) A small table using embedded subheaders for different diet categories, with explicit labels these rows with the variable they represent, compared with the tidy version of the same data (b).

Avoid embedding values from grouping variables into existing variables in the data rectangle

Embedded subheaders are commonly used to show hierarchical data such as membership in taxonomic or ecological groups. For example: Fig. 4a shows values from a grouping variable embedded in another existing variable, used to imply that the rows below belong to a given group, until the next embedded subheader implies otherwise. This practice of attribution by adjacency saves space and avoids repetition by having the value in one cell apply to cells below, in a way that is easy for humans to parse but not for computers. We can also view embedded subheaders as a way to show slices of the same dataset, as a stacked array of small tables (small multiples; Garmonsway, 2018) which can ultimately be combined into a single table. This approach is quite intuitive, and it is loosely-related with spreadsheet pivot tables and database normalization principles, so the use of embedded subheaders is widespread. This is not limited to tables meant for visual examination within the main text of an article, it is also common in spreadsheets, large tables, and supplementary materials that are not usually restricted in space or file size. The name of the embedded grouping variable is often left out of the data rectangle, and the cells in these rows are often merged to highlight the subgroups visually (Fig. 5a), which in turn creates additional problems for data reuse. Two particularly good examples of the common complications caused by embedded subheaders are: a) when we wish to subset and summarise data by group and we are forced to assign group membership visually, and b) when examining a table with more rows in a group than can fit on a page or computer screen, obscuring important information (Fig. 6). In both cases, the usual approach involves having to manually scroll through the rows to determine the start and end of each group. Additionally, merged and centered cells lead to non-rectangular and untidy data. Cell merging can make data unreadable to statistical software, or lead to unexpected results when interpreting the encoded information that recorded which cells were meant to be merged together.

Recommendation

Ideally, the information that is meant to "trickle down" should be put into its own variable, eliminating the need for embedded subheaders. However, the most suitable reporting structure will depend on how much data is being reported and shared. The compactness and readability of small tables in the main text of a document often increases when using subheaders to show groups, and for small tables this should not be problematic even in the worst-case scenario of having to transcribe the entire contents of a table if we wish to reuse the data. To facilitate data reuse, we suggest - even for small tables - labelling the subheaders with the variable they represent to disambiguate the subgroups from the data (Fig. 5a). This adds some repetition to the table content (but not as much as a tidy data structure) and should not create problems with space or word counts. Another major advantage of labelling the grouping rows is that the subheader labels can be matched in bulk and put into their own variable with a scripting approach. For example, using the *untangle2* function in the `unheadr` R package (Verde Arregoitia, 2018). An archived tutorial script for this package is available at doi:10.5281/zenodo.1724199. Fig. 5b shows the tidy version of the data in Fig. 5a. This structure is more versatile for analysis and there is no ambiguity as to the diet category of each species, but this arrangement is more repetitive and possibly less human-readable. For larger datasets, we suggest providing a tidy version of the same data as an appendix or supplementary file (or as an archived copy in a suitable repository). As a rule of thumb, any table that spans more than one page should be provided with all the grouping variables in their own column. This is to keep the group information visible while examining the data (Fig. 6) and to avoid accidental deletion of the embedded subheaders while working with such datasets (their intended use and reason for sharing).

Avoid broken values

When space is limited horizontally (e.g., table columns in a print or PDF page), the most common practice is to break up character strings within a cell into separate lines and display them together as using cell borders and line widths (Tab. 2). Sometimes this happens automatically in the software we use to make tables. We refer to these cases as wrapped or broken values because pieces of a single value are spread across more than one row. In formatted tables, multi-row character strings are easy to read because border formatting keeps the content cohesive. However, parsing these tables into a rectangular structure for further analyses often leads to an inconsistent number of empty cells within the different columns for each of the observational units (Fig. 7). In the data shown in Fig. 7, the main complication is that some of the observational units (scientific names for species in this case — specifically *Dipodomys merriami* and *Habromys simulatus*) are broken up across two rows, making it difficult to identify the observational units, match them with their values across the rest of the variables, and restructure the data.

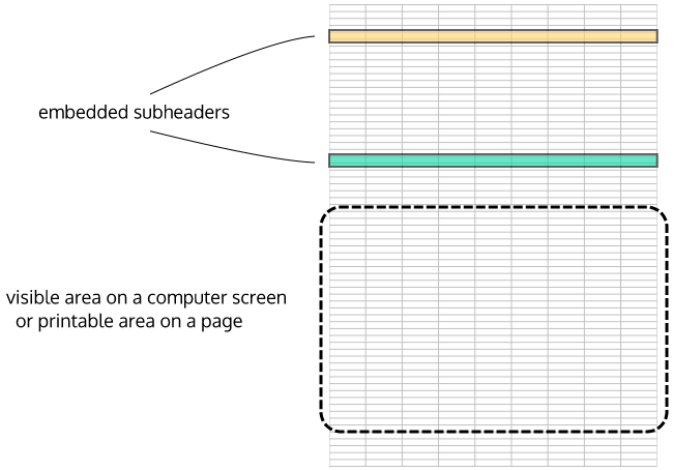
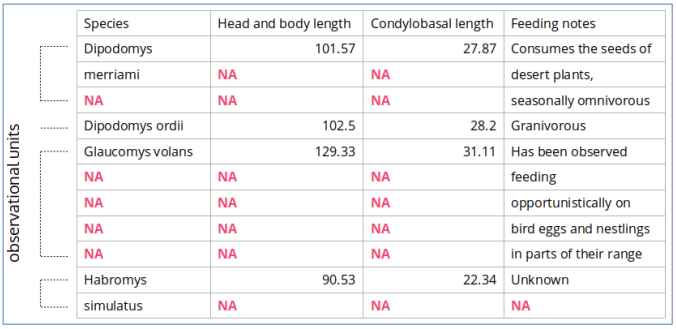


Figure 6 – Dataset with a large number (i.e., more than one page) of rows and in which the embedded subheaders that define the groups are not always visible.



observational units	Species	Head and body length	Condylobasal length	Feeding notes
	Dipodomys	101.57	27.87	Consumes the seeds of
	merriami	NA	NA	desert plants,
	NA	NA	NA	seasonally omnivorous
	Dipodomys ordii	102.5	28.2	Granivorous
	Glaucomys volans	129.33	31.11	Has been observed
	NA	NA	NA	feeding
	NA	NA	NA	opportunistically on
	NA	NA	NA	bird eggs and nestlings
	NA	NA	NA	in parts of their range
	Habromys	90.53	22.34	Unknown
	simulatus	NA	NA	NA

Figure 7 – Table without cell merging used to display multi-row strings, leading to missing values.

Recommendation

Table layouts and formatting in scientific journals and books usually follow strict guidelines, leaving less control for authors sharing data regarding multiline cells. Authors should be aware of the steep increase in complexity caused by wrapped values. Our recommendations to avoid this issue apply at different stages.

- When sharing a table as part of a manuscript, avoid breaking (wrapping) the values that represent the observational units across two or more rows. If the values for the observational units are not broken, the broken values in the remaining columns can be unwrapped (see the unwrap_cols and unbreak_values functions in the `unheadr` R package (Verde Arregoitia, 2018)). If it is still necessary to break up these values, use separators, indentations, or whitespace to show explicitly how the values are broken up.
- Otherwise, whenever important primary information is being shared in a report or publication (including relatively small tables), separate copies in which values are not split across rows could be made available.

Use adequate file formats

As scientific communications transitioned from print to electronic publishing, the portable document format (PDF) became the established file format for the electronic distribution of journal articles (Larivière et al., 2015). PDF files electronically replicate the appearance of pages in printed journal, and some publishers consider that the PDF file will remain as the preferred medium for the archiving, sharing, and offline use of scientific publications (Zudilova-Seinstra et al., 2014). Given the prevalence of PDF files, we wish to address the issues that arise when data is shared in this format. Despite its flexibility and portability, the PDF was not designed as a data format; it was designed as electronic paper. A PDF file contains a fixed layout and is meant to always look the same, regardless of the software or device used to open or print it. Even when content in a PDF page looks like a table or spreadsheet and was originally tabular, the format does not retain any sense of the unique cells that once contained the data. Tables in a PDF file are strategically-positioned lines and text, meaning that values cannot be easily copied and pasted into new aggregate datasets, or imported directly into statistical analysis programs. Although new tools and software routines are being developed to extract text and tables from PDF files (e.g. Tabula - https://tabula.technology/), the nature of PDFs mean that some degree of tedious and potentially error-prone hand-editing is still necessary to attain usable and structured data. Other output formats for academic publications exist (e.g. HTML and XML). Their use varies across journals and is typically only widespread for more 'recent' papers (e.g., after 1997 for the Journal of Biogeography, and after 1999 for the Journal of Mammalogy and The American Naturalist). These web-oriented formats (what we see when we read a full-text article in journal's website using a browser) can hold and display tables, links, references, and annotations reliably. However, these formats are almost exclusively an output medium generated by journals. A major advantage of this format is that tables retain their structure and are more readily accessible to users of spreadsheet software or scripting languages. A disadvantage is that accessing articles in this format often depends on an internet connection and a subscription to the respective journal. Additionally, these web formats can be technically challenging for many authors to create, manipulate, edit, and share.

Table 2 – Data with multi-line values shown inside vertically-merged cells.

species	Head and body length	Condylobasal length	Feeding notes
Dipodomys merriami	101.57	27.87	Consumes the seeds of desert plants, seasonally omnivorous
Dipodomys ordii	102.5	28.2	Granivorous
Glaucomys volans	129.33	31.11	Has been observed feeding opportunistically on bird eggs and nestlings in parts of their range
Habromys simulatus	90.53	22.34	Unknown

Recommendation

This issue with file formats is very important, yet relatively easy to solve. Broadly, we wish to raise awareness of how any data that is not shared separately will often get published exclusively in a PDF file, making it much less accessible and reusable. As a result of scientists and journals moving towards more transparency and reproducibility in data and methods, recent studies tend to provide tabular data as supporting information in downloadable delimited text or spreadsheet files. We encourage this practice, even when it is not an editorial requirement. Appendices are an important component of scholarly communications, and they have the advantage of often being bundled with the main text. In this case, appendices ultimately share the same PDF format of the publications they accompany, so appendix data should be shared separately as downloadable spreadsheet or delimited text files (e.g. `.csv`, `.txt`, or `.tsv` files). In addition to appendices, we often need to combine text, figures, and tables in supplementary information or extended methods files. If it is allowed by a journal's guidelines, we suggest sharing these files using XML-based word processor formats such as OpenDocument Text (`.odt`) or Office Open XML (`.docx`) files. These formats maintain the structure of tables, and represent a good option for reusability, interoperability, and long-term preservation. R users can extract tables from the popular and widely used `.docx` files using functions from the `docxtractr` package (Rudis, 2016).

Provide metadata

A dataset structured following tidy data principles, stored using an appropriate file format, and deposited in a suitable repository will not be fully reusable if it lacks higher-level documentation (i.e., data about the data). Metadata typically includes the information that is necessary to understand the origin, organization and characteristics of a dataset (Michener, 2018). This includes details about units of measurement, contact information, abbreviations or codes, protocol information, version information and much more.

Suggestion

Formal step-by-step guidelines and templates for creating standardized and machine-readable metadata are now available (see https://www.dataone.org/best-practices/metadata/ and Michener, 2018). In the context of our recommended good practices, we suggest at the very least using clear variable names, avoiding unexplained acronyms of abbreviations, and making sure that the data can be correctly interpreted, by ourselves at a later date or by others working with our published data.

Table 3 – Recap of good practices for sharing analysis-ready data.

Good practice	Main recommendations
Use explicit and consistent delimiters for text-based specimen lists	Provide tabular versions of specimen lists; Avoid using periods as delimiters
Avoid embedded subheaders	Place grouping values into a separate variable
Avoid broken values	Avoid multi-line values for the variables holding the observational units
Use adequate file formats	Share tabular data as delimited text files, not PDFs
Provide metadata	Describe all shared datasets to ensure correct interpretation

Concluding remarks

Data gathered by others is vital to the work of mammalogists, so the ideas here should benefit both users and producers of data. The ideas presented here (summarized in Tab. 3) apply to a fairly narrow part of the scientific workflow, but various seemingly minor and often aesthetic decisions at the reporting and publishing stages can inadvertently create important obstacles to data sharing, which in turn may lead to less data reuse, limited contribution towards new avenues of future research, and less impact and citations for a given paper and the corresponding journal. Raising awareness about data structures and file formats can benefit researchers at all levels of technical expertise (in regard to statistical programming, data-entry software, and spreadsheet programs). For example, consider the time and effort needed to work with the actual content of a table that contains embedded subheaders and is also stored as a PDF - compared with the same information shared as a delimited text or spreadsheet file in a tidy or tidy-friendly format. Better data structures can increase the reproducibility of our results, and the possibility of using these data to explore new hypotheses, usually in combination with other publicly available datasets. This brief communication is aimed mainly at authors sharing data. However, journal editors and editorial staff should be aware of these issues, since they ultimately have the last word on the final appearance and formatting of data in the published version of a paper. We suggested good practices in such a way that they would not clash with journal policies relating to space (e.g. page or word limits), although in an era of increasing online-only publishing, page limits and file storage should not be a major obstacle (see Moore and Beckerman, 2016). Therefore, minor changes in journal policies or updates to author guidelines could potentially enhance data uptake and reuse. For instance, journals can move beyond only enforcing data accessibility, and also suggest that authors provide readily usable (i.e. tidy or tidy-compatible) versions of any multi-page tables as spreadsheets or delimited plain text files. Well-structured data are a part of the open science process, and the benefits of open data practices outweigh the potential costs. Papers that archive data publicly are cited more often than papers that withhold the data (Piwowar, 2013), and data shared in suitable repositories (e.g. Dryad) often include clear guidelines and licensing details for data reuse (Culina et al., 2018). With the suggestions presented here, we wish to contribute to a growing number of guides and publications on how to organise data during the entry and collaboration stage (Broman and Woo, 2018; Ellis and Leek, 2018), the best practices in archiving data relevant to ecology and evolution (Whitlock, 2011), and the ways of ensuring appropriate data attribution such as Digital Object Identifiers (DOIs) and data citations (Zhao et al., 2018). More than a byproduct or stepping stone, data are scientific legacy and efficient data sharing makes new types of research and insights possible.

References

Álvarez A., Perez S.I., Verzi D.H., 2011. Early evolutionary differentiation of morphological variation in the mandible of South American caviomorph rodents (Rodentia, Caviomorpha). J. Evol. Biol. 24:2687-2695. doi:10.1111/j.1420-9101.2011.02395.x

Auer S.K., Killen S.S., Rezende E.L., 2017. Resting vs. active: a meta-analysis of the intra- and inter-specific associations between minimum, sustained, and maximum metabolic rates in vertebrates. Funct. Ecol. 31:1728–1738.

Baumer B.S., Kaplan D.T., Horton N.J., 2017. Tidy data and Iteration. Pp. 91-130 in Modern data science with R, CRC Press.

Broman K.W., Woo K.H., 2018. Data Organization in Spreadsheets. The American Statistician 72:2-10. doi:10.1080/00031305.2017.1375989

Culina A., Baglioni M., Crowther T.W., Visser M. E., Woutersen-Windhouwer S., Manghi P., 2018. Navigating the unfolding open data landscape in ecology and evolution. Nature Ecology & Evolution 2:420-426. doi:10.1038/s41559-017-0458-2

Dasu T., Johnson T., 2003. Data Quality. Pp. 99-137 in Exploratory Data Mining and Data Cleaning (Shewhart W.A., Wilks S.S., Dasu T., and Johnson T. eds.), Wiley.

Dornelas M., Antão L.H., Moyes F., Bates A.E., Magurran A.E., Adam D., Akhmetzhanova A.A., Appeltans W., Arcos J., Arnold H., Ayyappan N., Badihi G., Baird A.H., Barbosa M., Barreto T., Bässler C., Bellgrove A., Belmaker J., Benedetti-Cecchi L., Bett B.J., Bjorkman A.D., Błażewicz M., Blowes S.A., Bloch C.P., Bonebrake T.C., Boyd S., Bradford M., Brooks A.J., Brown J.H., Bruelheide H., Budy P., Carvalho F., Castañeda-Moya E., Chen C.Allen, Chamblee J.F., Chase T.J., Siegwart C.L., Collinge S.K., Condit R., Cooper E.J., Cornelissen J.H.C., Cotano U., Kyle Crow S., Damasceno G., Davies C.H., Davis R.A., Day F.P., Degraer S., Doherty T.S., Dunn T.E., Durigan G., Duffy J.E., Edelist D., Edgar G.J., Elahi R., Elmendorf S.C., Enemar A., Ernest S.K.M., Escribano R., Estiarte M., Evans B.S., Fan T., Turini F.F., Loureiro F.L., Farneda F.Z., Fidelis A., Fitt R., Fosaa A.M., Daher C.F.G.A., Frank G.E., Fraser W.R., García H., Cazzolla G.R., Givan O., Gorgone-Barbosa E., Gould W.A., Gries C., Grossman G.D., Gutierréz J.R., Hale S., Harmon M.E., Harte J., Haskins G., Henshaw D.L., Hermanutz L., Hidalgo P., Higuchi P., Hoey A., Van Hoey G., Hofgaard A., Holeck K., Hollister R.D., Holmes R., Hoogenboom M., Hsieh C.-hao, Hubbell S.P., Huettmann F., Huffard C.L., Hurlbert A.H., Macedo I.N., Janík D., Jandt U., Jażdżewska A., Johannessen T., Johnstone J., Jones J., Jones F.A.M., Kang J., Kartawijaya T., Keeley E.C., Kelt D.A., Kinnear R., Klanderud K., Knutsen H., Koenig C.C., Kortz A.R., Král K., Kuhnz L.A., Kuo C.-Yang, Kushner D.J., Laguionie-Marchais C., Lancaster L.T., Min Lee C., Lefcheck J.S., Lévesque E., Lightfoot D., Lloret F., Lloyd J.D., López-Baucells A., Louzao M., Madin J.S., Magnússon B., Malamud S., Matthews I., McFarland K.P., McGill B., McKnight D., McLarney W.O., Meador J., Meserve P.L., Metcalfe D.J., Meyer C.F. J., Michelsen A., Milchakova N., Moens T., Moland E., Moore J., Mathias Moreira C., Müller J., Murphy G., Myers-Smith I.H., Myster R.W., Naumov A., Neat F., Nelson J.A., Paul Nelson M., Newton S.F., Norden N., Oliver J.C., Olsen E.M., Onipchenko V.G., Pabis K., Pabst R.J., Paquette A., Pardede S., Paterson D.M., Pélissier R., Peñuelas J., Pérez-Matus A., Pizarro O., Pomati F., Post E., Prins H.H.T., Priscu J.C., Provoost P., Prudic K.L., Pulliainen E., Ramesh B.R., Mendivil R.O., Rassweiler A., Rebelo J.E., Reed D.C., Reich P.B., Remillard S.M., Richardson A.J., Richardson J.P., van Rijn I., Rocha R., Rivera-Monroy V.H., Rixen C., Robinson K.P., Ribeiro Rodrigues R., de Cerqueira R.F.D., Rudstam L., Ruhl H., Ruz C.S., Sampaio E.M., Rybicki N., Rypel A., Sal S., Salgado B., Santos F.A. M., Savassi-Coutinho A.P., Scanga S., Schmidt J., Schooley R., Setiawan F., Shao K., Shaver G.R., Sherman S., Sherry T.W., Siciński J., Sievers C., da Silva A.C., da Silva R.F., Silveira F.L., Slingsby J., Smart T., Snell S.J., Soudzilovskaia N.A., Souza G.B. G., Maluf Souza F., Castro Souza V., Stallings C.D., Stanforth R., Stanley E.H., Mauro Sterza J.é, Stevens M., Stuart-Smith R., Rondon Suarez Y., Supp S., Yoshio T.J., Tarigan S., Thiede G.P., Thorn S., Tolvanen A., Teresa Z.T.M., Totland Ø., Twilley R.R., Vaitkus G., Valdivia N., Vallejo M.I., Valone T.J., Van Colen C., Vanaverbeke J., Venturoli F., Verheye H.M., Vianna M., Vieira R.P., Vrška T., Quang Vu C., Van V.L., Waide R.B., Waldock C., Watts D., Webb S., Wesołowski T., White E.P., Widdicombe C.E., Wilgers D., Williams R., Williams S.B., Williamson M., Willig M.R., Willis T.J., Wipf S., Woods K.D., Woehler E.J., Zawada K., Zettler M.L., 2018. BioTIME: A database of biodiversity time series for the Anthropocene. Global Ecol. Biogeogr. 27:760-786. doi:10.1111/geb.12729

Ellis, S. E., Leek, J. T. 2018. How to Share Data for Collaboration. The American Statistician 72:53-57. doi:10.1080/00031305.2017.1375987

Faurby S., Davis M., Pedersen R.Ø., Schowanek S.D., Antonelli A., Svenning J.-C., 2018. PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. Ecology. doi: 10.1002/ecy.2443

Garmonsway, D. 2018. Spreadsheet Munging Strategies. https://nacnudus.github.io/spreadsheet-munging-strategies/

Gonçalves F., Bovendorp R.S., Beca G., Bello C., Costa-Pereira R., Muylaert R.L., Rodarte R.R., Villar N., Souza R., Graipel M.E., Cherem J.J., Faria D., Baumgarten J., Alvarez M.R., Vieira E.M., Cáceres N., Pardini R., Leite Y.L.R., Costa L.P., Mello M.A.R., Fischer E., Passos F.C., Varzinczak L.H., Prevedello J.A., Cruz-Neto A.P., Carvalho F., Percequillo A.R., Paviolo A., Nava A., Duarte J.M.B., de la Sancha N.U., Bernard E., Morato R.G., Ribeiro J.F., Becker R.G., Paise G., Tomasi P.S., Vélez-Garcia F., Melo G.L., Sponchiado J., Cerezer F., Barros M.A.S., de Souza A.Q.S., dos Santos C.C., Giné G.A.F., Kerches-Rogeri P., Weber M.M., Ambar G., Cabrera-Martinez L.V., Eriksson A., Silveira M., Santos C.F., Alves L., Barbier E., Rezende G.C., Garbino G.S.T., Rios E.O., Silva A., Nascimento A.T.A., de Carvalho R.S., Feijó A., Arrabal J., Agostini I., Lamattina D., Costa S., Vanderhoeven E., de Melo F.R., de Oliveira P., Jerusalinsky L.L, Valença-Montenegro M.M., Martins A.B., Ludwig G., de Azevedo R.B., Anzóategui A., da Silva M.X., Moraes M.F.D., Vogliotti A., Gatti A., Püttker T., Barros C.S., Martins T.K., Keuroghlian A., Eaton D.P., Neves C.L., Nardi M.S., Braga C., Gonçalves P.R., Srbek-Araujo A.C., Mendes P., de Oliveira J.A., Soares F.A.M., Rocha P.A., Crawshaw Jr. P., Ribeiro M.C., Galetti M., 2018. ATLANTIC MAMMAL TRAITS: a data set of morphological traits of mammals in the Atlantic Forest of South America. Ecology 99:498–498. doi:10.1002/ecy.2106

Hart E.M., Barmby P., LeBauer D., Michonneau F., Mount S., Mulrooney P., Poisot T., Woo K.H., Zimmerman N.B., Hollister J.W., 2016. Ten Simple Rules for Digital Data Storage. PLoS Comp Biol 12:e1005097. doi:10.1371/journal.pcbi.1005097

Hoffmann F.G., Baker R.J., 2001. Systematics of bats of the genus *Glossophaga* (Chiroptera: Phyllostomidae) and phylogeography in G. soricina based on the Cytochrome-b gene. J. Mammal. 82:1092-1101. doi:10.1644/1545-1542(2001)082<1092:SOBOTG>2.0.CO;2

Larivière V., Haustein S., Mongeon P., 2015. The Oligopoly of Academic Publishers in the Digital Era. PLoS ONE 10:e0127502. doi:10.1371/journal.pone.0127502

Lim B.K., Pedro W.A., Passos F.C., 2003. Differentiation and Species Status of the Neotropical Yellow-Eared Bats *Vampyressa pusilla* and *V. thyone* (Phyllostomidae) with a Molecular Phylogeny and Review of the Genus. Acta Chiropt. 5:15-29. doi:10.3161/001.005.0102

Lowndes J.S.S., Best B.D., Scarborough C., Afflerbach J.C., Frazier M.R., O'Hara C.C., Jiang N., Halpern B.S., 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1:0160. doi:10.1038/s41559-017-0160

Michener W.K., 2015. Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comp Biol 11:e1004525. doi:10.1371/journal.pcbi.1004525

Michener W. K., 2018. Creating and Managing Metadata. Pp. 71-88 in Ecological Informatics: Data Management and Knowledge Discovery (Recknagel F and Michener WK eds.), Springer International Publishing, Cham.

Moore A.J., Beckerman A., 2016. Ecology and Evolution in an Open World (or: why supplementary data are evil). Ecology and Evolution 6:2655-2656. doi:10.1002/ece3.2101

Osgood W.H., 1909. Revision of the Mice of the American Genus *Peromyscus*. US Government Printing Office.

Piwowar H.A., Vision T.J., 2013. Data reuse and the open data citation advantage. PeerJ 1:e175. doi:10.7717/peerj.175

Reichman O.J., Jones M.B., Schildhauer M.P., 2011. Challenges and opportunities of open data in ecology. Science 331:703-705.

Reschenhofer T., Matthes F., 2015. An Empirical Study on Spreadsheet Shortcomings from an Information Systems Perspective. Pp. 50-61, Springer International Publishing, Cham.

Rudis B., 2016. `docxtractr`: Extract Data Tables and Comments from Microsoft Word Documents. R package version 0.2.0. https://CRAN.R-project.org/package=docxtractr

Ruedas L.A., Salazar-Bravo J., Dragoo J.W., Yates T.L., 2000. The Importance of Being Earnest: What, if Anything, Constitutes a "Specimen Examined?". Mol. Phylogen. Evol. 17:129-132. doi:10.1006/mpev.2000.0737

Santini L., Isaac N.J.B., Ficetola G.F., 2018. TetraDENSITY: A database of population density estimates in terrestrial vertebrates. Global. Ecol. Biogeogr. 27:787-791. doi:doi:10.1111/geb.12756

Schweizer M., Güntert M., Seehausen O., Leuenberger C., Hertwig S.T., 2014. Parallel adaptations to nectarivory in parrots, key innovations and the diversification of the Lorinae. Ecology and Evolution 4:2867-2883. doi:10.1002/ece3.1131

Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L., Read E., Manoff M., Frame M., 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6:e21101. doi:10.1371/journal.pone.0021101

Troudet J., Vignes-Lebbe R., Grandcolas P., Legendre F., 2018. The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it? Syst. Biol.

Verde Arregoitia L.D., 2018. `unheadr`: Handle Data With Embedded Subheaders. R package version 0.1.0. https://github.com/luisDVA/unheadr

Wallis J.C., Rolando E., Borgman C.L., 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8:e67332. doi:10.1371/journal.pone.0067332

White E.P., Baldridge E., Brym Z.T., Locey K.J., McGlinn D.J., Supp S.R., 2013. Nine simple ways to make it easier to (re) use your data. Ideas in Ecology and Evolution 6:1-10. doi:10.4033/iee.2013.6b.6.f

Whitlock M.C., 2011. Data archiving in ecology and evolution: best practices. Trends. Ecol. Evol. 26:61-65. doi:10.1016/j.tree.2010.11.006

Wickham H., 2014. Tidy Data. Journal of Statistical Software 59:23. doi:10.18637/jss.v059.i10

Wilson G., Bryan J., Cranston K., Kitzes J., Nederbragt L., Teal T.K., 2017. Good enough practices in scientific computing. PLoS Comp Biol 13:e1005510. doi:10.1371/journal.pcbi.1005510

Zhao M., Yan E., Li K., 2018. Data set mentions and citations: A content analysis of full-text publications. Journal of the Association for Information Science and Technology 69:32-46. doi:doi:10.1002/asi.23919

Zudilova-Seinstra E., Klompenhouwer M., Heeman F., Aalbersberg I.J., 2014. 15 - The Elsevier Article of the Future project: a novel experience of online reading. Pp. 357-377 in The Future of the Academic Journal (Second Edition) (Cope B and Phillips A eds.), Chandos Publishing.

Associate Editor: D. Preatoni