BarnebyLives: an R package to create herbarium specimen labels and clean spreadsheets

Reed Clark Benkendorf[1]*, Jeremie B. Fant[1,2]

[1]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022, USA

[2]Plant Biology and Conservation, Northwestern University, Evanston, Illinois 60208, USA

Abstract

Premise: Depositing specimens to herbaria is a time consuming task. Many institutions have reduced the amount of funding for herbaria, and universities have reduced the amount of education dedicated to curatorial tasks and specimen deposition. Despite this, the continual generation of herbaria specimens are essential for current and future research in evolution and ecology. In order to faciliate the continued growth of herbaria BarnebyLives was developed as tool to supplement collection notes, perform geographic and, taxonomic informatic processes, enact spell checks, produce labels, and submit digital data.

Methods and Results: BarnebyLives uses geospatial data from the U.S. Census Bureau to provide political jurisdiction information, and data from other sources, including the United States Geological Survey, to supplement collection notes by providing information on abiotic site conditions. It uses inhouse spell checks to verify the spelling of a collection at all taxonomic ranks, the IPNI standard author database to check standard author abbreviations, and the Royal Botanic Garden Kews 'Plants of the World Online' to check for nomenclatural innovations. Optionally the package writes driving directions to sites using Google Maps. Finally the package outputs data in a tabular format for review by the user to accept or confirm changes,

Conclusions: BarnebyLives provides accurate political and physical information, reduces typos, provides users the most current taxonomic opinions, generates driving directions to sites, and produces aesthetically appealing labels and shipping manifests in a matter of minutes.

Nearly 400 million specimens are housed in herbaria around the world (Thiers (2021)). These specimens, collected with the goal of describing the plant kingdoms taxonomic diversity, and documenting the worlds floristic diversity, have found myriad new applications in several adjacent fields such as conservation biology

*Author for Correspondence: rbenkendorf@chicagobotanic.org

and ecology (Greve et al. (2016), James et al. (2018), Brewer et al. (2019), Rønsted et al. (2020)). However, The rate of accessioning new collections to herbaria diminished in the 20th century as priorities in biology shifted away from describing, documenting, and understanding earths diversity towards understanding cellular processes (Prather et al. (2004), Pyke and Ehrlich (2010), Daru et al. (2018)). Which, among other factors, lead to a decline in the amount of funding allocated to collections based research, and the number of staff maintaining and accessioning new collections (Funk (2014)). Fortunately, renewed interest in collections have brought herbaria of all sizes back to the forefront of plant sciences (Rønsted et al. (2020), Marsico et al. (2020)).

In fact recent innovations in computing, specimen digitization, data sharing, DNA sequencing, and statistics have likely brought about greater use of herbarium specimens than ever before (Greve et al. (2016), James et al. (2018), Brewer et al. (2019), Rønsted et al. (2020)). These current uses of specimen based data extend far beyond their traditional roles in systematics and floristics, and studies utilizing collections are regularly carried out to better understand the ecological niches, phenological processes, and interactions of plants (Rønsted et al. (2020)). Further we anticipate that collections will gain their most widespread utilization as natural history is being revitalized in ecology, via novel approaches, such as remote sensing, meta-barcoding, community science, and electronic sensing (Tosa et al. (2021)). Even within systematics, the ability to collect and analyze micro-morphological data has fostered revisionary studies in many groups providing synapomorphies which can be correlated to molecular results to improve the interpretation of monophyletic clades.

However, while there is recognized need for more specimens the skills of collecting and processing specimens, and time allocated for collecting, have declined among young persons (Daru et al. (2018), Mishler et al. (2020)). The submittal of specimens to herbaria is a, well documented albeit time consuming process, especially for younger collectors with limited experience.

While many young collectors, who are capable of using dichotomous keys to reliably identify their collections, exist we have observed that they face difficulties navigating several aspects of data collection and preparing specimens for herbaria. This scenario results in not only the delay in the deposition of many specimens, but undoubtedly in a failure for some specimens to ever be accessioned at all. Problems which young collectors face generally include both the lack of dedicated time awarded to them at a seasons end to process specimens, and a general lack of formal education on cartography, natural history, taxonomy, and plant systematics.

The successful generation of an herbarium specimen includes many steps which are easy to take for granted. For example, while the acquisition of political information for a collection site appears simple, it is only so if the collector has the adequate resources at their disposal.

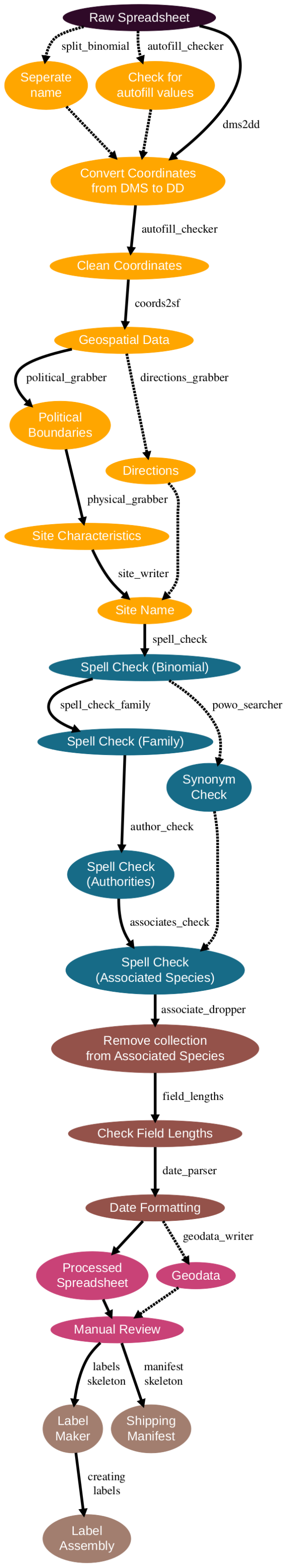


Figure 1: Recommended workflow.

Given the association of boundaries with topographically complex areas (e.g. watersheds) it often requires topographic maps, which are no longer widespread - resulting in many having difficulties interpreting them, or transcription of coordinates into a Geographic Information System (e.g. ArcMap, which is relatively expensive at 100$ year), or more likely Google Maps by individual site a time consuming process. This lack of topographic maps compounds the issues of young collectors being unable to come up with appropriate site names.

Further issues relate to the typical problems of nomenclature, to wit the pace at which taxonomic innovations are now made and which many persons of all skill levels find difficult to keep up with…

Here we provide a description of the BarnebyLives R package. BarnebyLives was named for plant taxonomist Rupert Charles Barneby (1911-2000), whom published over 6,500 pages of text, described over 750 taxa, and is notable for balancing his studies at the William & Lynda Steere Herbarium at the New York Botanical Garden with annual collection trips in Western North America from 1937-1970, and sporadically until his passing (Welsh (2001)). Select accolades of Rupert include the 1989 Asa Gray Award from the American Society of Plant Taxonomists (ASPT), the 1991 Engler Silver Medal from the International Association of Plant Taxonomists (IAPT), as well as being one of eight recipients of the International Botanical Congress's (IBC) Millennium Botany Award (1999) (Welsh (2001)).

METHODS AND RESULTS

Usage

All steps of BarnebyLives except for label generation are run from within Rstudio. Data may be read in from any common spreadsheet management system or database connection such as Excel, LibreOffice, OpenOffice, or via the cloud on Googlesheets. The latter two options are documented here and in package vignettes, detailed descriptions of the required and suggested input columns are located on the Github page (https://github.com/sagesteppe/BarnebyLives *'Input Data Column Names'*) and over 100 real-world examples are on a Google Sheets accessible from the page. BarnebyLives is atypical of R packages in that it requires a considerable amount of data to operate (Table 1). Virtually all of the on-disk memory associated with these data are for storing geo-spatial information, setting up a local instance of the program - at whichever scale a user desires (see Figure XX) is available in the package documentation. Functions which require the on-disk data require a path to the data as an argument. Manually supplying the argument allows for the users to judiciously decide a storage location suitable for there needs.

We anticipate most personal BarnebyLives instances will be less than several gigabytes, and the processing takes relatively little RAM, hence we believe installations can work on hardware as small as Chromebooks, and have the data stored entirely on thumb-drives. The final steps of Barnebylives, generating the labels require working installations of Rmarkdown, a LaTeX installation (e.g. pdflatex, lualatex, xelatex), and the open source command line tools pdfjam and pdftk. While these steps are run through bash, we have wrapped them in a R functions which bypass the need to enter the commands to a terminal. Several commands in BarnebyLives require the output from previous functions, and a workflow which satisfies these requirements is presented in Figure 1.

Herbarium Collections

The package was finalized using the primary authors collections from 2023. The testing of the package within this manuscript was performed using a subset of their collections from 2018-2022, *all* of which are un-accessioned. Only collections which had identifications to the level of species or lower, and transcribed collection dates and coordinates were used. This results in a data set of 819 records for testing, from 204 sites located across Western North America (Figure 2). In total 615 species (with 557 sets of authors), with 66 infraspecies (22 authors) in 73 families were used for testing.

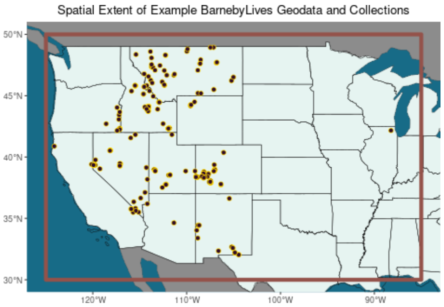

Spatial Extent of Example BarnebyLives Geodata and Collections
50°N
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W

Figure 2: The spatial extent (orange), and herbarium collection sites (burgundy) tested in this manuscript.

BarnebyLives took roughly three minutes (190.246s) to run all local steps, and roughly twelve minutes (703.167s) to search Plants of the World Online, and 73.73s to search Google Maps and write directions to sites. Most of the local run time is attributable to the spatial (spatial: 174.69s), and taxonomic operations (14.132s), style: 1.424s. The spell check operation of the scientific name accounted for nearly all of the time (14.092s) spent performing local taxonomic operations. The generation of labels consumed around seven minutes (424.042s) for the rendering, 50.54s to combine individual labels four per single sheet of landscape orientated paper, and 2.97s to combine the 205 sheets to a single Portable Document Format (PDF).

Results

```
## character(0)
```

Even on data which had been manually cleaned and error-checked by a human several times BarnebyLives was able to reduce transcription errors, identify typos, make nomenclature suggestions, and reformat text elements for downstream use. While no families were misspelled, BL made 24 suggestions on naming, 16 manually entered typos were found, it identified 2 instances where an incorrect family was entered, and 0 instances of an outdated circumscription applied. BL flagged 6 records where the author follows an alternative taxonomy, and flagged 0 records in error.

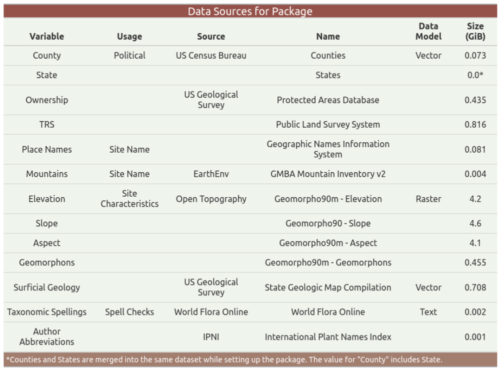
BL identified 57 discrepancies at the level of genus between user submitted and processed data. In 36 of these instances the user supplied an outdated name instance of an outdated circumscription applied (20 genera total). BL flagged 5 records where the author follows an alternative taxonomy (3 genera total), and flagged 0 record in error.

BL flagged 75 species

The number of author abbreviations which were not in the appropriate format were XX (% percent), in nearly all cases the presence or absence of a period were the issue. Plants of the World Online was able to identify XX new names for the submitted taxa, XX of which the author adopted. 5 records were appropriately flagged for issues with auto fill increment of the longitude value, and 3 records were also auto-flagged for increases in latitude values (% of records).

CONCLUSIONS

BarnebyLives is a tool which is able to rapidly acquire relevant geographic, and taxonomic data. It is also capable of performing specialized spell checks, and assorted curatorial tasks to produce both digital and analog data. The package relies on no licensed Software, such as the Microsoft suite, and is suitable for install on all major operating systems (Windows, Mac, Linux), with a small amount of use of the command line, which may be called from the Rstudio rather than a 'traditional' terminal.



Data Sources for Package

Variable	Usage	Source	Name	Data Model	Size (GiB)
County	Political	US Census Bureau	Counties	Vector	0.073
State			States		0.0*
Ownership		US Geological Survey	Protected Areas Database		0.435
TRS			Public Land Survey System		0.816
Place Names	Site Name		Geographic Names Information System		0.081
Mountains	Site Name	EarthEnv	GMBA Mountain Inventory v2		0.004
Elevation	Site Characteristics	Open Topography	Geomorpho90m - Elevation	Raster	4.2
Slope			Geomorpho90 - Slope		4.6
Aspect			Geomorpho90m - Aspect		4.1
Geomorphons			Geomorpho90m - Geomorphons		0.455
Surficial Geology		US Geological Survey	State Geologic Map Compilation	Vector	0.708
Taxonomic Spellings	Spell Checks	World Flora Online	World Flora Online	Text	0.002
Author Abbreviations		IPNI	International Plant Names Index		0.001

*Counties and States are merged into the same dataset while setting up the package. The value for "County" includes State.

Figure 3: Sources of Data required for operations

AUTHOR CONTRIBUTIONS

The project was conceptualized by R.C.B. The program was written by R.C.B. Data collection and analysis were performed by R.C.B. R.C.B. wrote the manuscript with input from all other authors. All authors approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The Bureau of Land Management are graciously acknowledged as providers of funding to R.C.B for the majority of his specimen collection activities. Two anonymous peer reviewers who increased the quality of this manuscript are thanked. Several prominent associated collectors of specimens used in this study are thanked: Dani Yashinovitz, Dakota Becerra, Hannah Lovell, Caitlin Miller & Hubert Szczygiel.

DATA AVAILABILITY STATEMENT

The BarnebyLives R package is open source, the development version is available on GitHub (https://github.com/sagesteppe/BarnebyLives), and the stable version is available on CRAN. The package includes three real use-case vignettes (tutorials) on usage. One vignette "setting_up_files" explores setting up a instance for a certain geographic area. Another vignette "running_pipeline" showcases the usage of the package for processing data entered on a spreadsheet. A final vignette "creating_labels" shows the usage of an R, and Bash script launched from RStudio to produce print-ready labels. All data used in this mansucript are available at: https://github.com/sagesteppe/Barneby_Lives_dev/manuscript

ORCID

Reed Benkendorf https://orcid.org/0000-0003-3110-6687
Jeremie Fant https://orcid.org/0000-0001-9276-1111

REFERENCES

Brewer, G. E., J. J. Clarkson, O. Maurin, A. R. Zuntini, V. Barber, S. Bellot, N. Biggs, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in plant science* 10: 1102.

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. Whitfeld, T. G. Seidler, et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217: 939–955.

Funk, V. A. 2014. The erosion of collections-based science: Alarming trend or coincidence. *The Plant Press* 17: 1–13.

Greve, M., A. M. Lykke, C. W. Fagg, R. E. Gereau, G. P. Lewis, R. Marchant, A. R. Marshall, et al. 2016. Realising the potential of herbarium records for conservation biology. *South African Journal of Botany* 105: 317–323.

James, S. A., P. S. Soltis, L. Belbin, A. D. Chapman, G. Nelson, D. L. Paul, and M. Collins. 2018. Herbarium data: Global biodiversity and societal botanical needs for novel research. *Applications in plant sciences* 6: e1024.

Marsico, T. D., E. R. Krimmel, J. R. Carter, E. L. Gillespie, P. D. Lowe, R. McCauley, A. B. Morris, et al. 2020. Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. *American journal of botany* 107: 1577–1587.

Mishler, B. D., R. Guralnick, P. S. Soltis, S. A. Smith, D. E. Soltis, N. Barve, J. M. Allen, and S. W. Laffan. 2020. Spatial phylogenetics of the north american flora. *Journal of Systematics and Evolution* 58: 393–405.

Prather, L. A., O. Alvarez-Fuentes, M. H. Mayfield, and C. J. Ferguson. 2004. The decline of plant collecting in the united states: A threat to the infrastructure of biodiversity studies. *Systematic Botany* 29: 15–28.

Pyke, G. H., and P. R. Ehrlich. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological reviews* 85: 247–266.

Rønsted, N., O. M. Grace, and M. A. Carine. 2020. Integrative and translational uses of herbarium collections across time, space, and species. *Frontiers in Plant Science* 11: 1319.

Thiers, B. M. 2021. The world's herbaria 2021: A summary report based on data from index herbarium.

Tosa, M. I., E. H. Dziedzic, C. L. Appel, J. Urbina, A. Massey, J. Ruprecht, C. E. Eriksson, et al. 2021. The rapid rise of next-generation natural history. *Frontiers in Ecology and Evolution* 9: 698131.

Welsh, S. L. 2001. Rupert c. Barneby (1911-2000). *Taxon*.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. A table of all time trials for each function.