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. The package outputs data in a tabular format for review by the user to accept or confirm changes, before dynamically rendering labels.

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.

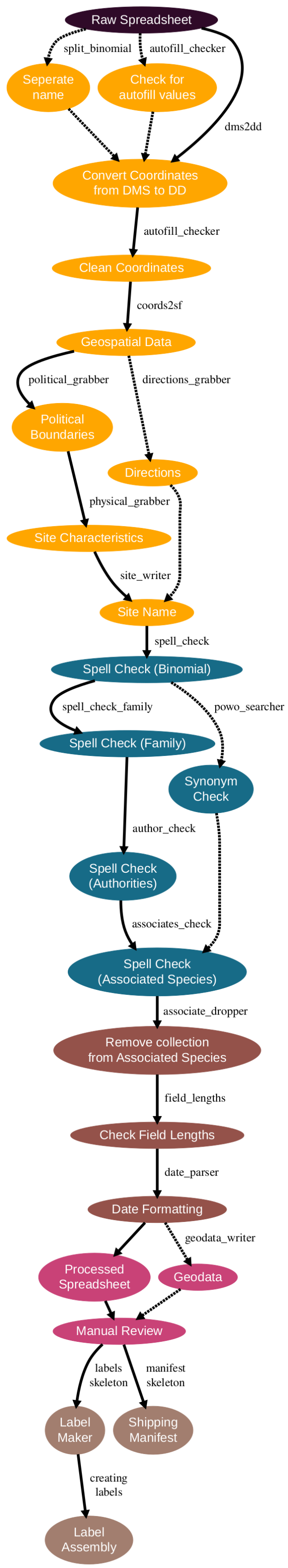
Nearly 400 million specimens are housed in herbaria around the globe (Thiers (2021)). These specimens, collected to describe the taxonomic diversity of plants and document the worlds floristic diversity, have recently found myriad new applications in several adjacent fields such as conservation biology and ecology

*Author for Correspondence: rbenkendorf@chicagobotanic.org

(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 20^{th} century as priorities in biology shifted away from describing and documenting earths biodiversity towards understanding cellular and molecular processes (Prather et al. (2004), Pyke and Ehrlich (2010), Daru et al. (2018)). This shift, among other factors, lead to a decline in the funding allocated to collections based research, the number of staff maintaining and accessioning new collections, and educating students in these practices (Funk (2014)). Fortunately, renewed interest approaches in collections generated by ‘big data approaches’ have brought herbarium collections back to the forefront of the natural sciences (Rønsted et al. (2020), Marsico et al. (2020)).

In fact 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)). The 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), Davis (2023)). Further we anticipate that collections are yet to gain their most widespread utilization as a revitalization of natural history appears underway in ecology, fostered via novel approaches such as remote and electronic sensing, meta-barcoding, and community science (Tosa et al. (2021)). While image or purely observational (rather than collections based) citizen science initializes (e.g. iNaturalist) have dovetailed to meet many needs of these studies specimens contain rich data which are not accessible via images. Namely specimens have the ability to: provide samples of DNA, secondary metabolites, or proteins, notes on the status and composition of the biotic and abiotic settings at time of collection, material for measuring (micro-)morphological attributes (Borges et al. (2020)), and seeds or pollen; eternally ensuing specimens as the ultimate data source to center most efforts around.

However, despite this renewed recognition of the utility of collections, efforts to continually grow them appear slow (Prather et al. (2004)). We conjecture this is in part because collecting and depositing specimens is a fundamentally slower process, especially for novice collectors, than simply taking photographs via well-developed apps (Daru et al. (2018), Mishler et al. (2020)). While many young botanists, capable of using dichotomous keys to reliably identify - and able to collect satisfactory - material exist, we have observed that they face difficulties navigating several aspects of data collection and preparation of labels for submission to herbaria. Apparent problems include the lack of dedicated time at a field seasons end to process specimens, a general lack of education on cartography and orienteering, natural history (e.g. geology, geomorphology), nomenclature and Latin, various computer programs (e.g. Microsoft Office suite), and increasingly - plant systematics (Nanglu et al. (2023), Woodland (2007), Barrows et al. (2016)). In the absence of suitable mentors this assuredly results in not only the delay in the deposition of many specimens, but undoubtedly in a failure for many specimens to be accessioned at all, and increasingly ever collected.



The generation of an herbarium specimen includes many steps which are easy to take for granted (Forman and Bridson (1989)). For example while acquiring appropriate political information for a collection site appears simple, young collectors rarely have the adequate resources (printed topographic maps, or GIS software) at their disposal. In topographically complex areas, where borders are often associated with hydrologic basins and the ridges defining them, collectors are liable to misinterpret their position. Finding appropriate sites names is another problem which can rarely be solved without a printed map, as many software maps now consider many features which would serve as site names extraneous in the era of GPS. The rate at which taxonomic innovations are occurring has left many Floras difficult for young users to interpret and has made it difficult for them to find more recently applied names. Upon finding a name they may find it frustrating to hear that while published, the proposal has been accepted by few practitioners and they have unwittingly offended certain curators. Formatting a label correctly (e.g. abbreviations), if successful upon even setting up a mail merge, is a time consuming process and likely to introduce several errors in formatting. Even if a collector navigates all of these hurdles successfully, the time allocated to each step is quite large. Further each step of interfacing with different resources increases the opportunity for transcription errors.

Here we provide a description of the BarnebyLives R package. BarnebyLives aims to increase both the data quality of labels, and to speed up the process of producing them. It rapidly provides political and administrative boundary information for a collection site using data from the U.S. Census Bureau (Walker (2024)), the Public Land Survey System, and ownership details of public lands via the Protected-Areas Database (PAD-US) from (Gap Analysis Project (GAP) (2024)). Site names are suggested via finding the closest unambiguously named place feature via the Geographic Name Information System (GNIS), and by precise calculation of the distance and azimuth from these localities to the collection site (Survey (2023)). Using the GMBA Mountain Inventory V. 2, a standardized named mountain data set with global coverage, which we have supplemented with over XXXX valleys allows for a relevant descriptor of the general region without any ambiguity (Snethlage et al. (2022)). Spell checks on all scientific names (including associated species) are performed using a copy of the World Checklist of Vascular Plants, and the collected species may be searched via Kew's Plant of the World Online for relevant synonyms (Govaerts et al. (2021), POWO (2024)). Author abbreviations are verified using IPNI's Standard Author Abbreviation Checklist and also returned by Kew's Plants of the World Online to ensure proper abbreviation of authorities (The Royal Botanic Gardens and Herbarium (2024), POWO (2024)). Checks are performed to search for common issues associated with spreadsheets, or transcription, such as the auto-filling of coordinate and date columns. After final review of the data generated by the package, it allows for the option to export spreadsheets which are usable for mass upload of data to multiple common herbarium databases, as well as the generation of herbarium labels.

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 in 2000 (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)). Most importantly, Rupert was remembered as being generous with his time to assist younger botanists with the more arcane aspects of field botany and taxonomy (Holmgren and Holmgren (1988)).

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, or free of charge alternatives such as: 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 spatial data, setting up a local instance of the program - at whichever scale a user desires (see Figure XX) is fully documented in the package documentation. Functions which require the on-disk data require a path to the data as an argument.

Manually supplying the path argument allows for users to judiciously decide a storage location suitable for their needs.

We anticipate most personal BarnebyLives instances will be less than several gigabytes (ours covering all of the conterminous Western U.S. is XX GiB), and the processing takes relatively little RAM, hence we believe installations can work on hardware as limited as Chromebooks, while having 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 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.

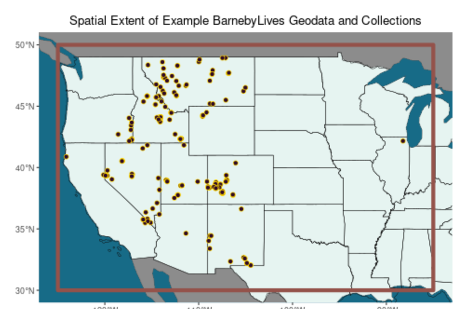


Figure 2: The spatial extent (orange), and herbarium collection sites (burgundy) tested in this manuscript.

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).
The total computer run time for processing these 819 specimens was 16 minutes.

Results

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 none of the 73 families were misspelled, BarnebyLives made 24 suggestions on naming, identified 16 typos, identified 2 instances where an incorrect family was entered, and 0 instances of an outdated circumscription applied. At the level of family BarnebyLives flagged 6 records where the author follows an alternative taxonomy, and flagged 0 records in error.

In the 292 genera analysed BarnebyLives 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 (20 unique genera). BL flagged 5 records where the author follows an alternative taxonomy (3 genera total), and flagged 0 records in error.

Of the 819 species analysed (615 distinct species) BarnebyLives flagged 55 records. BL 28 detected instances of misspelled epithets (28 unique species). In 15 of these instances the user supplied an outdated name (15 unique species). BL flagged 2 records where the author follows an alternative taxonomy (2 unique species), and flagged 9 records in error. The final record was an egregious error where the order of the specific epithet and the genus name.

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.

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. Sofia Garcia is acknowledged for creating the

'Valleys' data set which place naming in the package relies on. 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 manuscript 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

Barrows, C. W., M. L. Murphy-Mariscal, and R. R. Hernandez. 2016. At a crossroads: The nature of natural history in the twenty-first century. *BioScience* 66: 592–599.

Borges, L. M., V. C. Reis, and R. Izbicki. 2020. Schrödinger's phenotypes: Herbarium specimens show two-dimensional images are both good and (not so) bad sources of morphological data. *Methods in Ecology and Evolution* 11: 1296–1308.

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.

Davis, C. C. 2023. The herbarium of the future. *Trends in Ecology & Evolution* 38: 412–423.

Forman, L., and D. Bridson. 1989. The herbarium handbook. Royal Botanic Gardens Kew.

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

Gap Analysis Project (GAP), U. S. G. S. (USGS). 2024. Protected areas database of the united states (PAD-US) 4.0.

Govaerts, R., E. Nic Lughadha, N. Black, R. Turner, and A. Paton. 2021. The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Scientific data* 8: 215.

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.

Holmgren, N., and P. Holmgren. 1988. Intermountain flora v. 7. The New York Botanical Garden Press, New York.

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.

Nanglu, K., D. de Carle, T. M. Cullen, E. B. Anderson, S. Arif, R. A. Castañeda, L. M. Chang, et al. 2023. The nature of science: The fundamental role of natural history in ecology, evolution, conservation, and education. *Ecology and Evolution* 13: e10621.

POWO. 2024. Geographic names information system (GNIS) - USGS national map downloadable data collection: U.s. Geological survey.

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.

Snethlage, M. A., J. Geschke, A. Ranipeta, W. Jetz, N. G. Yoccoz, C. Körner, E. M. Spehn, et al. 2022. A hierarchical inventory of the world's mountains for global comparative mountain science. *Scientific data* 9: 149.

Survey, U. S. G. 2023. Geographic names information system (GNIS) - USGS national map downloadable data collection: U.s. Geological survey.

The Royal Botanic Gardens, H. U. H. &. L., Kew, and A. N. Herbarium. 2024. International plant names index.

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.

Walker, K. 2024. Tigris: Load census TIGER/line shapefiles.

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

Woodland, D. W. 2007. Are botanists becoming the dinosaurs of biology in the 21st century? *South African Journal of Botany* 73: 343–346.

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.