BarnebyLives':' an R package to create herbarium specimen labels and digital data sheets

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 were collected with the goal of describing the plant kingdoms taxonomic diversity, and documenting the worlds floristic diversity (Greve et al. (2016)). The rate of accessioning new collections to herbaria diminished

*Correspondence: rbenkendorf@chicagobotanic.org

in the 20th century as research goals in the biological sciences shifted away from describing, documenting, and understanding earths biodiversity (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)).

Recent innovations in computing, specimen digitization, data sharing, DNA sequencing, and statistics have brought about a renaissance in herbarium based studies (Greve et al. (2016), James et al. (2018), Brewer et al. (2019), Rønsted et al. (2020)). 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)). However, 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, electronic sensing (Tosa et al. (2021)).

However, we now stand at a time where we recognize the need for more specimens, but are in a difficult position where 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 in the process. 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. This scenario results in not only the delay in the deposition of many specimens, but undoubtedly the deposition of many collections 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. 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. This lack of topographic maps compounds the issues of young collectors being unable to come up with appropriate site names.

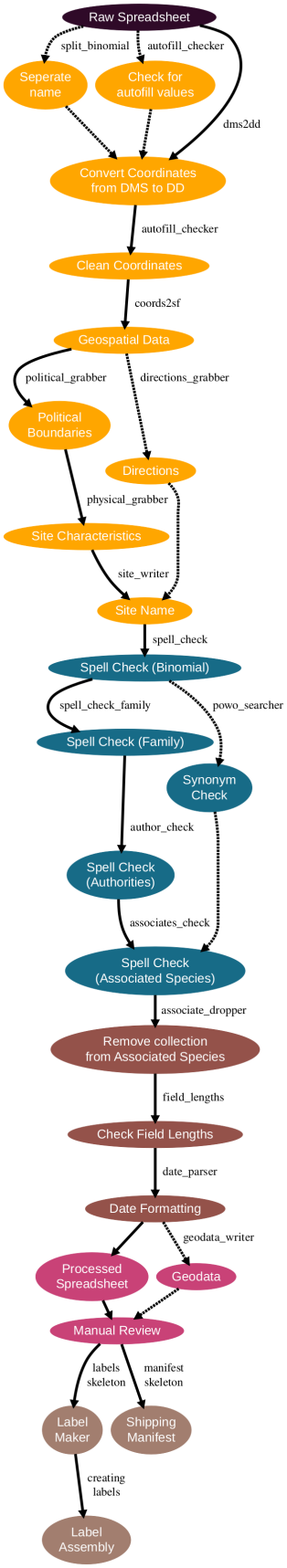

Figure 1: Recommended workflow.

Here we provide a description of the BarnebyLives R package. BarnebyLives was named for plant taxonomist extraordinaire 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)).

More evidently difficult tasks involve taxonomy and the rapid rate at which taxonomic names have changed since the publication of many Floras.

METHODS AND RESULTS

Elit turpis elementum litora volutpat eget blandit – velit quam in nullam. Nullam lobortis erat duis eleifend, torquent malesuada felis aliquet etiam sodales! Et viverra aenean bibendum cursus – nunc nascetur cursus? Morbi semper ante elementum primis orci molestie, hac, suscipit fames sem potenti proin? Suspendisse ut justo at taciti!

Adipiscing interdum sapien tortor eros cursus curabitur hendrerit dictumst; feugiat senectus! Sagittis fermentum, tortor pretium ad; non magna dui dapibus! Duis laoreet purus: cursus tristique tristique tortor auctor mauris tempus sed, integer eros.

Adipiscing dictum suspendisse; mollis diam enim pellentesque pellentesque hac justo purus tortor. Urna curabitur convallis in diam, aptent quisque scelerisque dictum ligula. Suscipit hendrerit accumsan

luctus eros. Urna ligula suscipit blandit parturient posuere magna suspendisse dictumst, nulla viverra, curabitur maecenas, hendrerit parturient ad mollis lobortis mus volutpat.

Lorem aliquet at nibh venenatis luctus cras at maecenas. Feugiat vel blandit accumsan eleifend praesent – ultricies ultricies eget luctus mi nunc in imperdiet. Et porta eu fames parturient tristique suspendisse aenean libero lobortis ullamcorper sed. Odio natoque euismod taciti posuere pretium aliquet vitae. Per tempor commodo condimentum diam donec primis purus, mus lacinia nibh potenti. Conubia ultrices congue litora proin ut ullamcorper lectus.

Consectetur volutpat odio congue vivamus nisi at? Fringilla tincidunt vitae mattis habitant commodo fringilla sociis parturient quam per lacus conubia eros varius!

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, or 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 XX.

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 XX. In total 615 species (with 557 sets of authors), with 66 infraspecies (22 authors) in 73 families were used for testing.

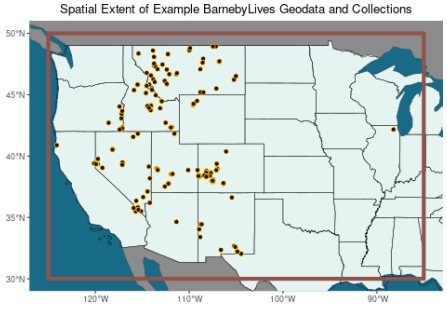

Figure 2: The spatial extent (orange), and herbarium collection sites (burgundy) tested in this manuscript.

It took roughly seven minutes (392.664s) to run all local steps of BarnebyLives, and roughly ten minutes (605.025s) to search Plants of the World Online, and 78.785s to search Google Maps and write directions to sites. Most of the local run time is attributable to the spatial (spatial: 162.376s), and taxonomic operations (228.774s), style: 1.514s. The spell check operation of the scientific name accounted for nearly all of the time (228.683s) 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

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. The number of family misspellings were XX (% percent), the number of misspelled genera were XX (% percent), the number of misspelled binomials were XX (% percent). 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 incrementation 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

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.