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

[1]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022, USA
[2]Plant Biology and Conservation, Northwestern University, Evanston, Illinois 60208, USA
Reed Clark Benkendorf[1,2]*, Jeremie B. Fant[1,2]

Abstract

Premise: Depositing specimens in herbaria is a time-consuming process. Many institutions have reduced the amount of funding for herbaria and universities have reduced the amount of education dedicated to collections and curatorial tasks. However, the continual generation of herbarium specimens is essential for current and future research on evolution and ecology. To facilitate the continued growth of herbaria, BarnebyLives was developed as a tool to supplement collection notes, perform geographic and taxonomic information processing, enact spell checks and other QC steps, produce labels, and submit digital data to increase the rate of accessioning specimens.

Methods and Results: BarnebyLives uses geospatial data from the U.S. The Census Bureau provides administrative jurisdictional 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 in-house 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 nomenclature innovations. Optionally, the package writes driving directions to the sites using Google Maps. The package outputs data in tabular and spatial formats for review by the user before rendering labels using LaTeX.

Conclusions: BarnebyLives provides accurate political and physical information, reduces typos, provides users with 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 worldwide in herbaria (Thiers, 2021). However, The rate of accessioning new collections to herbaria diminished in the 20th century as priorities in biology shifted away from

*Author for Correspondence: rbenkendorf@chicagobotanic.org

describing and documenting earths biodiversity and towards understanding cellular and molecular processes underpinning life (Prather et al., 2004; Pyke and Ehrlich, 2010; Daru et al., 2018). This shift, among other factors, led to a decline in the funding allocated to collection-based research, the number of staff maintaining and accessing new collections, and educating students in these practices (Funk, 2014). Historically, specimens have been used to describe the taxonomic diversity of plants and document global floristic diversity (Greve et al., 2016; James et al., 2018; Brewer et al., 2019; Rønsted et al., 2020). However, renewed interest in herbarium collections utilizing 'big data approaches,' such as museuomics, has brought herbaria back to the forefront of the natural sciences (Rønsted et al., 2020; Marsico et al., 2020).

Innovations in specimen digitization, data sharing, computing, DNA sequencing, and statistics have perhaps 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 use of specimens and their ancillary data extends well 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). We suspect that collections are yet to realize their full potential, and as currently novel approaches, such as remote and electronic sensing and meta-barcoding, become more accessible the use of collections will increase (Tosa et al., 2021). While image-based or purely observational (rather than collection-based) citizen science approaches (e.g., iNaturalist, BudBurst) have recently dovetailed with herbarium specimens to meet many current research needs, specimens contain rich data that are not accessible via images. Only specimens have the ability to: provide samples of DNA, secondary metabolites, or proteins, material for measuring (micro-)morphological attributes (Borges et al., 2020), and seeds or pollen. These factors ensure that the specimens remain the premier botanical data source.

However, despite renewed recognition of the utility of collections, efforts to grow them appear slow (Prather et al., 2004). We conjecture that this is partly because collecting and depositing specimens is a fundamentally slower process, especially for novice collectors, relative to taking photographs via professionally developed apps on smartphones (Daru et al., 2018; Mishler et al., 2020; Manzano and Julier, 2021). While many young botanists are capable of using dichotomous keys and other resources to reliably identify and collect satisfactory material, we observed that they face difficulties navigating several aspects of data acquisition, processing, and preparation of labels for submission to herbaria. Evidently, apparent problems include the lack of dedicated time at the end of a field season to process specimens, a general lack of education on cartography and orienteering, natural history (e.g., geology, geomorphology), nomenclature, and familiarity with various computer programs (for example, Microsoft Office suite), and increasing foundational knowledge of plant systematics and phylogenetics (Woodland, 2007; Barrows et al., 2016; Nanglu et al., 2023).

The generation of an herbarium specimen involves many steps that 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 adequate cartographic resources (printed topographic maps or GIS software) at their disposal. In topographically complex areas, where administrative borders are often associated with hydrological basins and the ridges defining them, collectors are liable to misinterpret their true geographic position and report details in error. Even finding appropriate site names can rarely be solved without a printed map, as many navigation-related software now consider many features that would serve as extraneous site names. Similarly, the rate at which taxonomic innovations occur, the volume of the literature, and the reluctance of some regional curators to embrace a phylogenetic approach to plant classification have made it difficult to find more recently applied names, even when these names are unanimously accepted by group experts (Hitchcock and Cronquist, 2018). Furthermore, formatting a label correctly (e.g., author abbreviations) is a time-consuming process with many opportunities to introduce errors in formatting. Anecdotally, many mail merge templates offered by herbaria still require collectors to modify many variables by hand, for example, applying italicization. Even if a collector navigates all these hurdles, the time allocated to each step is quite large.

As a result of these concerns, we have developed an R package, BarnebyLives, that aims to increase both the quality of data rendered to labels and recorded in databases and to speed up the process of producing labels. 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 (PLSS), and ownership details of public lands via the Protected-Areas Database (PAD-US) (Gap Analysis Project (GAP), 2024). Site names are suggested by finding the closest unambiguously named place feature in the Geographic Name Information System (GNIS) and the precise calculation of distance and azimuth from this feature to the collection site (Survey, 2023). Using the Global Mountain Biodiversity Assessment (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 with less 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 the International Plant Names Index (IPNI) Standard Author Abbreviation Checklist and also returned by Kew's Plants of the World Online to ensure proper abbreviations of authorities (The Royal Botanic Gardens and Herbarium, 2024; POWO, 2024). Checks to search for and flag common issues associated with spreadsheet software or data transcription, such as the auto-filling of coordinate and date columns. After a final review of the data generated by the package, it allows for the option to export spreadsheets that are suitable for mass uploading of data to multiple common herbarium databases as well as the generation of herbarium labels.

Currently, label generation functionality is provided explicitly by two programs, PLabel and Symbiota, as well as by the Microsoft Word tool Mail Merge (Perkins (2020), Gries et al. (2014)). The office suite costs money, and in our experience, it is finicky; further, its functionality ends with label creation. PLabel is a standalone program that has greatly enhanced functionality relative to a Mail Merge, allowing users to specify the layout and formatting of label components using an intuitive and local graphical user interface (GUI) functionality. However, it does not include data cleaning functionalities beyond verifying nations of collection. While some sources indicate that it can only be used in Microsoft, we expect it to be accessed on Linux and Mac using Windows 'emulators' like Wine. The increasingly popular Symbiota biodiversity data management software not only provides label generation capabilities but also provides data cleaning functionality in an attractive GUI web portal allowing for live management of collections and bypassing the need for a local installation, allowing it to be accessed on all operating systems. Symbiota offers functionality similar to the first four of our five stages of our 'Taxonomic' module and to our knowledge a check of the 'Political Boundaries' (see Figure 1). However, not all herbaria use Symbiota and many have original database systems that they maintain (for example, Harvard University Herbarium, https://kiki.huh.harvard.edu/databases/specimen_index.html; Missouri Botanical Garden https://tropicos.org/specimen/Search; and The Consortium of Pacific Northwest Herbaria https://www.pnwherbaria.org/). However, many collectors prefer to generate their own labels, especially as they are likely to send different sets of collections to different institutions. Accordingly, the functionality of Symbiota should exist in an ecosystem with alternative systems. In scenarios where users want to keep rendering labels in either of the three existing alternatives, they can easily export data in the appropriate formats after utilizing BLs data cleaning utilities.

BarnebyLives was named by plant taxonomist Rupert Charles Barneby (1911-2000), who published over 6,500 pages of text, described over 750 taxa, and is notable for balancing his studies at the William and Lynda Steere Herbarium at the New York Botanical Garden with annual collection trips in Western North America from 1937-1970 and sporadically until he passed 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 germanely, 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

[Figure 1 about here.]

BarnebyLives was iteratively developed based on data submitted by approximately 20 seasonal field botany teams over two years. Essentially, continual updates were made as the developers became aware of the idiosyncrasies of collection notes and data entry.

Usage

All steps of BarnebyLives, except for label generation are run within the freely available Rstudio. Data may be read from any common spreadsheet management system or database connection such as Excel, or free alternatives such as LibreOffice, OpenOffice, or via the cloud on Google Sheets. The latter two options are documented here and in package vignettes, detailed descriptions of the required and suggested input columns are located on a Github page (https://github.com/sagesteppe/BarnebyLives 'Input Data Column Names') and 96 real-world examples are on a Google Sheets accessible from the page. BarnebyLives is atypical for R packages in that it requires a considerable amount of data to operate (Table 1). Virtually all on-disk memory associated with the package are used to store spatial data. The amount of spatial data varies according to the domain that the user decides to support (Figure 3). Functions that require on-disk data require a path to data as an argument. Manually supplying the path argument allows users to determine an appropriate storage location suitable for their needs.

We anticipate that for a typical user, BarnebyLives will require less than a couple gigabytes of memory (ours covering all of the conterminous Western U.S. at 90m resolution is ~16 GiB), while the processing requires 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, requires working installations of R Markdown, a LaTeX installation (e.g. pdfTeX, LuaTeX, XeLaTeX), and the open source command line tools pdfjam and pdftk. While these steps are run through a shell scripting language such as bash, we have wrapped them in R functions that bypass the need to enter the commands directly into a terminal. Unfortunately, we did not find Windows alternatives to pdfam and pdftek, so we are unable to offer the final label-generating functionality on that operating system. BarnebyLives was iteratively developed based on the data submitted by approximately 20 seasonal field botanists. Essentially, additions to the code have been continually made, as developers were exposed to more idiosyncrasies of collection notes and data entry. Several commands in BarnebyLives require output from previous functions, and a workflow that satisfies these requirements is presented in Figure 1.

Functionality

BarnebyLives can be thought of as consisting of five main modules (Figure 1): spatial, taxonomic, formatting, manual review, and data exporting.

The spatial module has five required functions and two optional functions.

autofill_checker searches for patterns in the input latitude and longitude data associated with autofilling from various spreadsheet programs and will emit a warning if they are encountered. *coords2sf* creates a spatially explicit simple feature (sf) geometry dataset for the input data. political_grabber determines many levels of administrative ownership, including land management and public land survey system sections. *physical_grabber* provides various geographic data, such as elevation, landform position, and aspect using 90m resolution spatial data. site_writer will write directions from an officially named placename to the collection site. dms2dd is an optional function used to convert from coordinates denoted in the degrees minutes and second format (for example, 42°08'39.9"N 87°47'08.3"W) to decimal degree format (for example 42.14439, -87.78569). *directions_grabber* is an optional function that writes driving directions from a reasonably sized town to the closest drivable area to the site using the Google Maps API, which will require a valid Google account that is free per month for most personal and smaller academic usages.

The taxonomic module has four required functions and one optional function. *spell_check* will perform a spell check on the entered scientific name based on a local copy of Kew Plants of the World database filtered to the local continents or a user-specified backbone. *spell_check_family* performs a spell check on the family entered for each scientific name. author_check ensures that the authors are entered in a valid format, for example, the correct standard abbreviations are used. *associates_check* performs a spell check on all associated species using the local taxonomic database. *powo_searcher* can be used in tandem with the functions spell_check_family and author-check, but we use it in lieu of them to search the current Plants of the World Online to determine relevant synonyms and alternative higher taxonomy for the focal species. No API key or registration is required to use powo_searcher.

The formatting module has three functions. The first two we will detail are technically optional; however, they are run locally and so quickly that there is no reason to skip them. *associate_dropper* silently removes the collected species from the list of associated species; however, it searches for the species to be removed using the scientific name entered initially by the user rather than returned via spell checks. *field_lengths* will emit messages for any fields that we suspect will create an 'overflow' on the physical label and should be truncated for clarity. *date_parser* is mandatory and parses an input date into various formats for notating collection and determination dates on labels.

The manual review process technically only has one function that is optional and may be executed during the spatial process (after coords2sf), but the importance of manual review is important enough to warrant explicit mention. *geodata_writer* will write out a spatial copy of the data set to any geospatial format supported by the sf package, but defaults to writing out 'kmls' which are readily used with Google Earth, and can also be opened in several other free geographic information system (GIS) softwares such as QGIS. Notably, many of the flags that BarnebyLives generates will be placed into columns with obviously flagged names and can be manually reviewed by the analyst, and many of these issues can be resolved by simply addressing the relevant issues in the original data input spreadsheet.

The data exporting module contains three functions that interact with LaTeX templates and require slightly more advanced R user interactivity, such as setting up mapping functions using the tidyverses purrr package. *labels_skeleton* is an R 'script' which will require a few modification steps to tailor to each institution, these R scripts will put data into a user specified template, and serve as the interface to LaTeX. "

Herbarium Collections

[Figure 2 about here.]

The package was released into beta testing 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. Only collections which had identifications to the level of species or lower, and transcribed collection dates and coordinates were used. Resulting in a data set of 978 records for testing, from 234 sites located across Western North America (Figure 2). In total this data set had 728 species (with 558 distinct sets of authors), with 83 infraspecies (22 authorships) in 74 families which were used for testing.

BarnebyLives took roughly four minutes (222.886sec) to run all local steps, and roughly ten minutes (584.05sec) to search Plants of the World Online, and a minute 63.945sec to search Google Maps and write directions to sites. BarnebyLives took roughly four minutes (222.886sec) to run all local steps, and roughly ten minutes (584.05sec) to search Plants of the World Online, and 63.945sec to search Google Maps and write directions to sites.

Most of the local run time is attributable to the spatial (205.254sec), and taxonomic operations (17.253sec), while formatting data for labels took 0.379sec. The spell check of the scientific name accounted for nearly all of the time (17.006sec) spent performing local taxonomic operations. The generation of labels consumed around eight minutes (509.931sec) for the rendering, and an additional 58.60sec to combine the 0 **sheets** to a single Portable Document Format (PDF). The total computer run time for processing these 978 specimens was 15 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 74 families were misspelled, BarnebyLives made 25 suggestions on naming, identified 15 typos, identified 2 instances where an incorrect family was entered, and 0 instances where an outdated circumscription was applied. At the level of family BarnebyLives flagged 6 records where the author follows an alternative taxonomy, and flagged 2 records in error.

In the 326 genera analysed BarnebyLives identified 75 discrepancies at the level of genus between user submitted and processed data. In 43 of these instances the user supplied an outdated name (20 unique genera) flagged 5 records where the author follows an alternative taxonomy (2 genera total), and flagged 1 records in error.

Of the 978 records analysed (728 distinct species) BarnebyLives flagged 61 records, and detected 29 instances of misspelled epithets (29 unique species). Of the 978 species analysed (728 distinct species) BarnebyLives flagged 61 records, and detected 29 instances of misspelled epithets (29 unique species). In 18 of these instances the user supplied an outdated name (18 unique species). It also flagged 4 records where the author follows an alternative taxonomy (4 unique species), and flagged 8 records in error. The final record was an egregious error where the order of the specific epithet and the genus name.

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). Vegetation data were entered for 904 records, and after removal of duplicate site information BL flagged 72 records. BL was able to correct the spelling in 55 instances, and in 12 or instances the user entered data in a format beyond the capabilities of BL to fix. BL returned an incorrect spell check match for 3 records, two of which had the same misspelled species, in 1 instance BL had an internal errors where it returned an incorrect match due to abbreviations.

[Figure 3 about here.]

DISCUSSION

While numerous tools have been developed for cleaning existing herbarium and museum records, few tools help to ensure that the data entered are accurate (Patten et al. (2024)). We argue that the original collectors are the most qualified individuals to perform quality control checks and that BarnebyLives allows them to do so in a relatively fast and streamlined format. By utilizing both R and LaTeX and having publicly available source code on Github, this program allows users immediate familiarity with the system for troubleshooting issues and implementing upgrades and modifications in project branches.

Accessioning often relies on the use of the Microsoft Office suite of programs and may utilize other costly software such as ArcPro or Adobe Acrobat. While BarnebyLives does not have its own graphic user interface, the functionality of commonly used Interactive Development Environments (IDE's), such as Rstudio and Visual-Studio (VS) Code, now offers functionality to readily view and filter datasets using familiar spreadsheet-like formats, making them more accessible to many users.

While other software often cost money, these are also free, and we recommend that users install an open-source PDF viewer such as Okular to review their rendered documents.

While numerous tools have been developed for cleaning existing herbarium and museum records, few help ensure that the entered data are accurate (Patten et al. (2024)). We argue that the original collectors are the most qualified individuals to perform quality control checks, and BarnebyLives allows them to do so in a relatively fast and streamlined format. By utilizing both R and LaTeX and having publicly available source code on Github, this program allows most users immediate familiarity with the system for troubleshooting issues and implementing upgrades and modifications to project branches.

Accessioning often relies on the use of the Office Suite of programs and may utilize other costly software such as ArcPro or Adobe Acrobat. While BarnebyLives does not have its own graphic user interface, the functionality of commonly used Interactive Development Environments (IDE's), such as Rstudio and Visual-Studio (VS) Code, now offers functionality to readily view and filter datasets using familiar spreadsheet-like formats.

LaTeX offers well-documented and detailed functionality to customize labels for individuals and institutions. Anecdotally, using its default settings, it can produce more aesthetically pleasing results than typical word processors. Very good documentation of LaTeX capabilities is offered in multiple areas; for instance, via the Overleaf) project.

CONCLUSIONS

BarnebyLives is an R package that can be used to rapidly acquire relevant geographic and taxonomic data. It can also perform 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 Office suite, and is suitable for install on all major operating systems (Windows, Mac, Linux), however currently label generation support is only offered on Linux and Mac, with a small amount of use of the command line, which may be called from the Rstudio rather than a 'traditional' terminal.

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. & J.B.F wrote the manuscript, and both authors approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The Bureau of Land Management is gratefully acknowledged as a provider of funding to R. C. B. for most of his specimen collection activities. We thank the two anonymous peer reviewers who have increased the quality of this manuscript. 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 and, Caitlin Miller & Hubert Szczygiel. We thank Rosalind Rowe for providing useful feedback during the later developmental stages of the program.

DATA AVAILABILITY STATEMENT

The BarnebyLives R package is open source, the development version is available on GitHub (https://github.com/sagesteppe/BarnebyLives). The package includes three real use-case vignettes (tutorials) available on a Github page site (https://sagesteppe.github.io/BarnebyLives/). 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.

The 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 Clark 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. Schr
"odinger'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.

Gries, C., M. E. E. Gilbert, and N. M. Franz. 2014. Symbiota–a virtual platform for creating voucher-based biodiversity information communities. *Biodiversity data journal.*

Hitchcock, C. L., and A. Cronquist. 2018. Flora of the pacific northwest: An illustrated manual. University of Washington Press.

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.

Manzano, S., and A. C. Julier. 2021. How FAIR are plant sciences in the twenty-first century? The pressing need for reproducibility in plant ecology and evolution. *Proceedings of the Royal Society B* 288: 20202597.

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.

Patten, N. N., M. L. Gaynor, D. E. Soltis, and P. S. Soltis. 2024. Geographic and taxonomic occurrence r-based scrubbing (gatoRs): An r package and workflow for processing biodiversity data. *Applications in Plant Sciences* 12: e11575.

Perkins, K. 2020. Plabel.

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.*

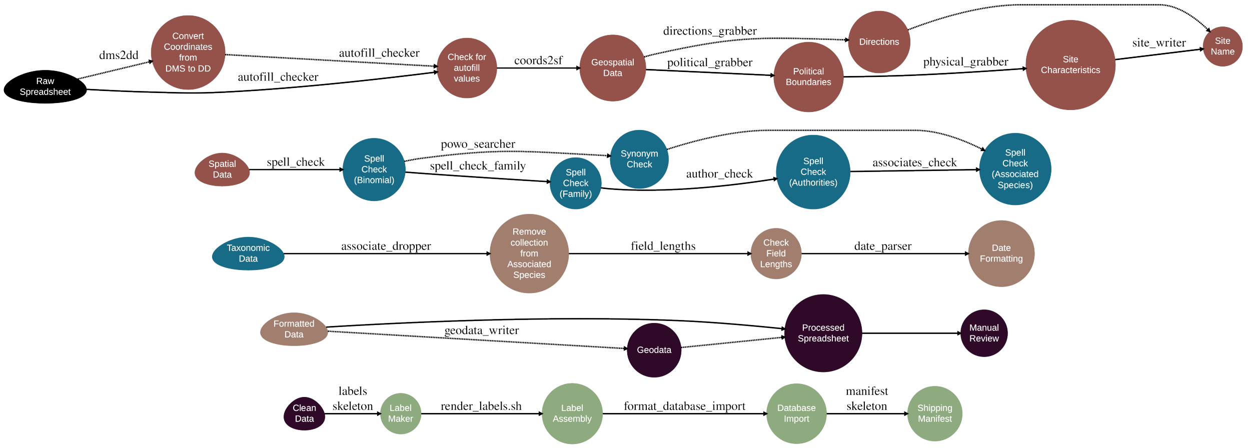
Woodland, D. W. 2007. Are botanists becoming the dinosaurs of biology in the 21st century? *South African Journal of Botany* 73: 343–346.

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.

List of Figures



Sample workflow for using BarnebyLives on data which has been manually entered into a spreadsheet. The top two rows indicate the main data cleaning functionality and are best run in the order outlined above although taxonomic steps may be ran before spatial steps. The third row can be interspersed with the above two, includes creation of labels, which allows for detection of formatting or other issues which were not captured by the pipeline or in earlier manual review. Further support is offered to export data in a format which allows mass upload at the receiving institution, and to create a shipping manifest and transfer notice.

Figure 1: Recommended workflow

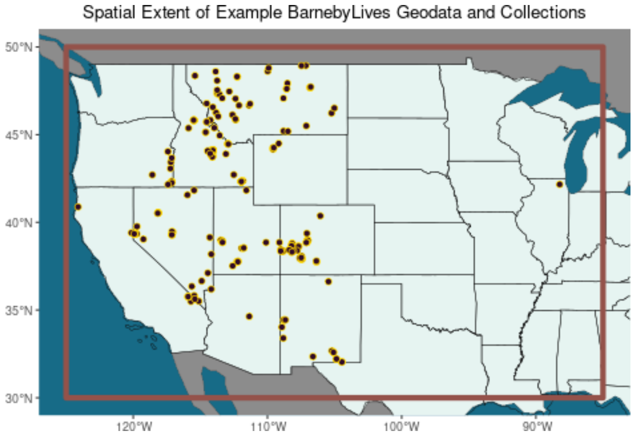
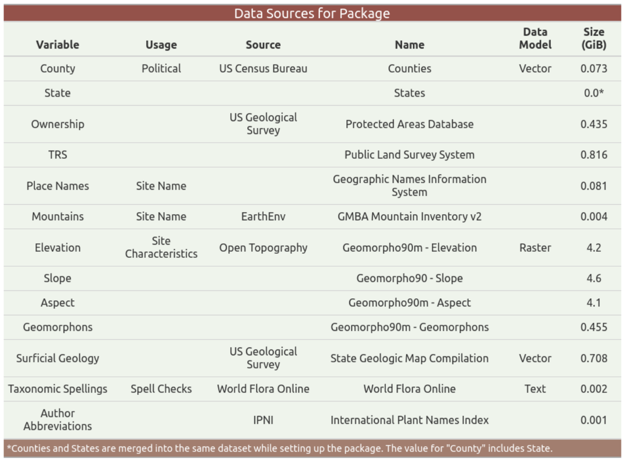

Spatial Extent of Example BarnebyLives Geodata and Collections
50°N
45°N
40°N
35°N
30°N
120°W
110°W
100°W
90°W

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



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: Data Sources