Applications in Plant Sciences

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

--Manuscript Draft--

Manuscript Number:	APPS-D-25-00042
Full Title:	BarnebyLives: an R package to create herbarium specimen labels and clean spreadsheets
Article Type:	Software Note
Keywords:	herbarium; software; QC; automation
Corresponding Author:	Reed Clark Benkendorf, MSc Northwestern University Glencoe, IL UNITED STATES OF AMERICA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Northwestern University
Corresponding Author's Secondary Institution:	
First Author:	Reed Clark Benkendorf, MSc
First Author Secondary Information:	
Order of Authors:	Reed Clark Benkendorf, MSc
	Jeremie B. Fant
Order of Authors Secondary Information:	
Abstract:	Premise: Accessioning herbarium specimens is labor-intensive, yet remains vital for research in ecology, evolution, and conservation. As institutional support for herbaria declines, efficient tools are needed to streamline this process. BarnebyLives was developed to assist collectors by supplementing collection notes, verifying taxonomic data, conducting quality checks, generating labels, and submitting digital records. Methods and Results: It integrates geospatial data from U.S. government sources to provide jurisdictional and site information, and checks taxonomic names using in-house spell checkers, IPNI author standards, and Kew's Plants of the World Online. Optional features include generating Google Maps driving directions. The tool outputs data in tabular and spatial formats for review before producing LaTeX-based labels and shipping manifests. Conclusions: BarnebyLives improves data accuracy, ensures up-to-date taxonomy, and significantly reduces the time and effort required to accession herbarium specimens.
Suggested Reviewers:	Michelle (Shelly) Gaynor, Ph.D postdoc, University of Michigan mlgaynor@umich.edu Has developed several R packages, and published a couple in APPS software notes, member (former?) of iDigBio team.
	Ben Legler University of Wyoming blegler@uwyo.edu Database administrator and creator of two large herbarium consortia websites, and a bona fide field botanist.
	Jason Alexander, Ph.D Jepson Herbarium: University of California Berkeley University and Jepson Herbaria I don't actually know Jason, word from the mutual friends is he know his stuff though.
	Charles Davis, Ph.D

	Professor & Curator, Harvard University Herbaria cdavis@oeb.harvard.edu Well you asked for four... Charles is a profuse user of herbarium specimens for modelling global patterns, testing ecoloical hypothesis at regional scales, and for systematics.
Opposed Reviewers:	
Funding Information:	

Applications in Plant Sciences Author Agreement Form

Corresponding Author's Name: Reed Benkendorf
Date: 4/23/25

Respond to all the statements below either by typing your initials or checking the appropriate box. After you have completed this form, save it to your desktop and upload it with your manuscript submission in Editorial Manager (http://www.edmgr.com/apps).

1. All authors know of and concur with the submission of this manuscript to *APPS*. (Single authors please also initial.)
Initials: RCB

2. All authors of this research paper have directly contributed (https://casrai.org/credit/)
AND
All authors of this paper have read and approved the final version submitted.
Initials: RCB

3. The contents of this manuscript have not been copyrighted or published previously and are not now under consideration for publication elsewhere. The contents of this manuscript will not be copyrighted, submitted, or published elsewhere while acceptance by *APPS* is under consideration.
Initials: RCB

4. *APPS* operates under a Creative Commons Attribution license, under the CC BY, CC BY-NC, and CC BY-NC-ND licenses. More information is available here: https://bsapubs.onlinelibrary.wiley.com/hub/journal/21680450/homepage/open_access_license_and_copyright.

You will be asked to select the appropriate license on completion of the licensing agreement after article acceptance. Each of these licenses ensures wide availability of the article and that the article can be included in any scientific archive. No permission for reuse is required from the author or the Botanical Society of America.
Initials: RCB

5. Authors are responsible for recognizing and disclosing any duality of interest that could be perceived to bias their work, acknowledging all financial support and any other personal connections.

☑ **No**, there is no duality of interest that I should disclose, having read the above statement.

☐ **Yes**, having read the above statement, there is potential duality of interest. This has been fully detailed in my cover letter.

6. Have the results/data/figures in this manuscript been published or are they under consideration for publication elsewhere?

☑ **No**, the results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from any of the authors) by another publisher.

☐ **Yes**, some portion of the results/data/figures in this manuscript has been published or is under consideration for publication elsewhere.

6a. If you select Yes, please identify results/data/figures taken from other published/pending manuscripts in the textbox below and explain why this does not constitute dual publication. (Note: The existence of pending or previously published articles that use or have used any of the same results presented in the submitted manuscript does not generally prejudice review and acceptance.)

7. All *APPS* papers are published as Open Access articles. Article Processing Charges (APCs) are as listed here: https://bsapubs.onlinelibrary.wiley.com/hub/journal/21680450/homepage/article_publication_charges:

Authors will be invoiced and payment must be received before publication. This policy applies to all articles accepted, except for invited articles for which the Editorial Board has agreed to waive fees at the time of article submission. Reduced APCs are available for authors who are members of the Botanical Society of America.

☐ **No,** No authors of this manuscript are members of the BSA.

☑ **Yes,** I confirm that at least one author is a BSA member.

9. *APPS* requires that genetic information be submitted to an appropriate data bank or repository. See the Author Guidelines for more information. I have read and understand this policy.
Initials: RCB

Contact the Editorial Office at apps@botany.org for more information.

[last revised 28 October 2020]

Brianna Gross
Editor in Chief
Applications in Plant Sciences

April 24th, 2025

Dear Dr. Gross,

I am pleased to submit a software note titled *"BarnebyLives: an R package to create herbarium specimen labels and clean spreadsheets"* for consideration in Applications in Plant Sciences..

This package was developed to assist field botanists and collectors—particularly those engaged in large-scale federal efforts such as germplasm collection (e.g., Seeds of Success) or ecological monitoring programs (e.g., National Wetland Condition Assessment, or Assess, Inventory and Monitor)—in generating herbarium specimen labels and ensuring data quality in a reproducible, standardized way. While several tools exist for working with already accessioned herbarium data, few are tailored to the needs of collectors preparing specimens for deposit. Fewer still are designed with the realities of tight timelines and varying herbarium requirements in mind.

The software has already been used to prepare data for collections from a variety of teams deposited at approximately 15 herbaria, and the output from it has been received quiet warmly. Feedback from curators has helped shape its features and scope, and we hope that making it widely available will help empower field botanists to contribute high-quality collections with greater efficiency and confidence.

I think that APPS is a wonderful place for this paper to be published, and APPS would be happy to have this paper. Several editors of APPS have been unrelenting in their support of herbaria and that herbaria maintain their role as the center of academic and field botany. We believe that on occasion herbaria are not seen as resources for the future, but rather catalogues of the past, and would like to make our small contribution to ensure that does not become true.

We affirm that this submission is original, has not been published elsewhere, and is not under consideration by another journal. There are no conflicts of interest to disclose.

Thank you for considering this submission. I look forward to your response.

Sincerely,
Reed

Running headline: Benkendorf & Fant. - BarnebyLives

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

Authors and Affiliations: Reed Clark Benkendorf[1,2*], Jeremie B. Fant[1,2]

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

[2] Chicago Botanic Garden, Glencoe, Illinois 60022, USA

*Corresponding author reedbenkendorf2021@u.northwestern.edu

Reed Benkendorf *https://orcid.org/0000-0003-3110-6687*

Jeremie Fant *https://orcid.org/0000-0001-9276-1111*

Manuscript received ___; revision accepted ___.

Word count: 4098

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.

ACKNOWLEDGMENTS

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. Several prominent associated collectors of specimens used in this study are thanked: Dani Yashinovitz, Hannah Lovell, Dakota Becerra, 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). The package includes three real use-case vignettes (tutorials) available on a Github Pages site (https://sagesteppe.github.io/BarnebyLives/). The first vignette *"Preparing to use BarnebyLives!"* shows how to set up an instance for a certain geographic area (domain). The next two vignettes *"BarnebyLives! Running pipeline"* showcases the usage of the package for processing data entered on a spreadsheet, and *"Printing herbarium labels and exporting a digital copy of data"* how to export data in both digital and analog formats. *"Custom label templates"* shows how to customize labels in LaTeX, and *"Rendering a shipping manifest"* details how to produce a shipping manifest for gifting or transferring material to an herbarium. All data used in this manuscript are available at: https://github.com/sagesteppe/Barneby_Lives_dev/manuscript.

Abstract

Premise: Accessioning herbarium specimens is labor-intensive, yet remains vital for research in ecology, evolution, and conservation. As institutional support for herbaria declines, efficient tools are needed to streamline this process. BarnebyLives was developed to assist collectors by supplementing collection notes, verifying taxonomic data, conducting quality checks, generating labels, and submitting digital records.

Methods and Results: It integrates geospatial data from U.S. government sources to provide jurisdictional and site information, and checks taxonomic names using in-house spell checkers, IPNI author standards, and Kew's Plants of the World Online. Optional features include generating Google Maps driving directions. The tool outputs data in tabular and spatial formats for review before producing LaTeX-based labels and shipping manifests.

Conclusions: BarnebyLives improves data accuracy, ensures up-to-date taxonomy, and significantly reduces the time and effort required to accession herbarium specimens.

KEYWORDS: herbarium; software; QC; automation

Introduction

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 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 and grearly expanded their roles in science (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 electronic and remote 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 will ensure that the specimens remain the premier botanical data source into perpetuity.

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 commercially developed apps on smartphones (Daru et al., 2018; Mishler et al., 2020; Manzano and Julier, 2021). While many novice botanists are capable of using dichotomous keys and other resources to reliably identify and collect satisfactory material, we observe that they face difficulties navigating several aspects of data acquisition, processing, and preparation of labels for submission to herbaria. Some of the 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, novice 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 administrative details in error. Even finding appropriate site names can rarely be resolved without a printed map, as many navigation-related software now consider most features that would serve as site names extraneous. 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 scientific names, even when these names are unanimously accepted by taxonomic specialists in the group and other regional curators (Hitchcock and Cronquist, 2018). Furthermore, formatting a label correctly (e.g., author abbreviations, italicization, etc.) is a time-consuming process with many opportunities to introduce errors in formatting which reduce the apparent credibility of a collector. 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 successfully navigates all these hurdles, the time allocated to each step is quite large, and may discourage them from further collecting.

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 generation of labels. BarnebyLives 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 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 resolved 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, flagged or 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, to our knowledge label generation functionality is provided explicitly by two programs, PLabel and Symbiota, and by the Microsoft Word tool Mail Merge (Gries et al., 2014; Perkins, 2020). The office suite costs money, and in our experience, 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, beyond verifying the nations of collection it does not include data cleaning functionalities. While some sources indicate that it can only be used on Microsoft, we expect it to be usable 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, and most importantly 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 for 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

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. Several commands in BarnebyLives require output from previous functions, and a workflow that satisfies these requirements is presented in Figure 1.

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 Pages (https://sagesteppe.github.io/BarnebyLives/) and around 100 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 3-arc second (~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. Given that the attributes which the package collects data on are tailored to the Western U.S. region, we do not expect local installs to exceed the size of ours. 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 shell terminal outside of RStudio. Unfortunately, we have not found Windows alternatives to pdfam and pdftek, so we are unable to offer the final label-generating functionality on that operating system, but suspect Ubuntu subsystem for Windows may allow for integration of these tools.

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 write distance and azimuth to collection site from the nearest official named place from the GNIS database. *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.

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).

Please note that the function *physical_grabber* is the one portion of the package where a decoupling may exist between the collection site, and the resolution of the spatial data. While we expect the mismatch to be negligible for all effective purposes relating to: elevation, major geology type, and in general aspect, estimates of slope at this resolution may be biased - generally to lower angles. For these reasons collectors must always make notes on the truly local environment which taxa are found in, and consider that the notes from BL reflect the greater landscape which a microfeature may be present in. While this mis-match will seldom effect landscape ecologists, it may have implications for other data users.

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. Two are optional; however, they are run locally and so quickly that there is no reason to skip them. *date_parser* parses an input date into various formats for notating collection and determination dates on labels.

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.

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.

label_writer write from a flatfile or spreadsheet to small 4x4 inch herbarium labels (users can modify these dimensions as they see fit). *format_database_import* will write out a spreadsheet of cleaned data in a variety of formats, currently: Jepson, Symbiota, and Consortium of Pacific Northwest herbaria are supported.

Herbarium Collections

{Figure 2}

The testing of the package within this manuscript was performed using a subset of the authors collections from 2018-2022, while most development was performed on their 2023 and 2024 collections. Only collections which had identifications to the level of species or lower, and transcribed collection dates and coordinates were used for most functionality. In total 980 records were used for testing various functions, these records were 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.

BarnebyLives took roughly four minutes (227.481sec) to run all local steps, and roughly ten minutes (595.294sec) to search Plants of the World Online for preferred synonyms, and a minute 64.869sec to search Google Maps and write directions to sites.

Most of the local run time is attributable to the spatial (209.089sec), and taxonomic operations (17.932sec), while formatting data for labels took 0.46sec. The spell check of the scientific name accounted for nearly all of the time (17.688sec) spent performing local taxonomic operations. The generation of labels consumed around nine minutes (523.5sec) for the rendering, and an additional 61.08sec to combine the 182 sheets to a single Portable Document Format (PDF). The total label generation run time for processing these 728 collections was 15 minutes. In total the 728 collections, which underwent all processing steps, took 25 minutes to process.

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 6 instances where the user entered an unequivocally incorrect family (or taxonomic entity), identified 5 records where families were autofilled, and 1 instance where an outdated circumscription was applied. At the level of family BarnebyLives flagged 6 records where the author follows an alternative taxonomy, and flagged 7 records in error, it appears most of these errors are due to issues in the backbone used by the earlier spell check function.

In the 326 genera analysed BarnebyLives identified 74 discrepancies at the level of genus between user submitted and processed data. In 42 of these instances the user supplied an outdated name (21 unique genera) flagged 4 records where the author follows an alternative taxonomy (2 genera total), and flagged 2 record in error.

Of 728 distinct species analysed BarnebyLives flagged 62 records, and detected 33 instances of misspelled epithets (33 unique species). In 15 of these instances the user supplied an outdated name (15 unique species). It also flagged 2 records where the author follows an alternative taxonomy (2 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. All flags were correct, and in several instances more errors were found in the rows following the flagged values.

{Figure 3}

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 assume that responsibility 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.

LaTeX, a software system used for typesetting, allows users to focus on the content rather than the style of the documents rendered from it. However, using its default settings, it can produce aesthetically pleasing results (Figure 4). Additionally LaTeX offers users a wide variety of ways which they can modify labels which are under-explored in the package. Very good documentation of LaTeX capabilities is offered in multiple areas; for instance, via the Overleaf project. While the templates in the package are quite simple, LaTeX also offers the ability to use custom fonts, to alter font weights and colors, alter line spacing, to include images (e.g. dot maps) and customize labels beyond what the default templates support.

Thematically, BarnebyLives is set up to cover Western North America. However, the package supports the use of a ‘domain’ being drawn over any of the conterminous United States. Several of the attributes which it collects and displays on labels, relate to topics which more senior curators are interested in, i.e. the administrative information on Township Section and Range (or ‘TRS’), but are considered less value in other geographic regions.

Further several of the abiotic variables which it acquires information on: slope, aspect, and geology have long been considered prominent drivers of plant distributions in semi-arid and montane systems and warranted on a label in these types of systems, whereas curators in other regions may find this information superfluous. Finally, it is plausible people in other geographic areas are less interested in displaying which land management agency has jurisdiction over a collection; however in the west we believe this is useful information which may help a collector interested in revisiting a site to determine if they will require permits for access or to make new collections.

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 VisualStudio (VS) Code, now offer 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.

{Figure 4}

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.

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. Schrodinger'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.

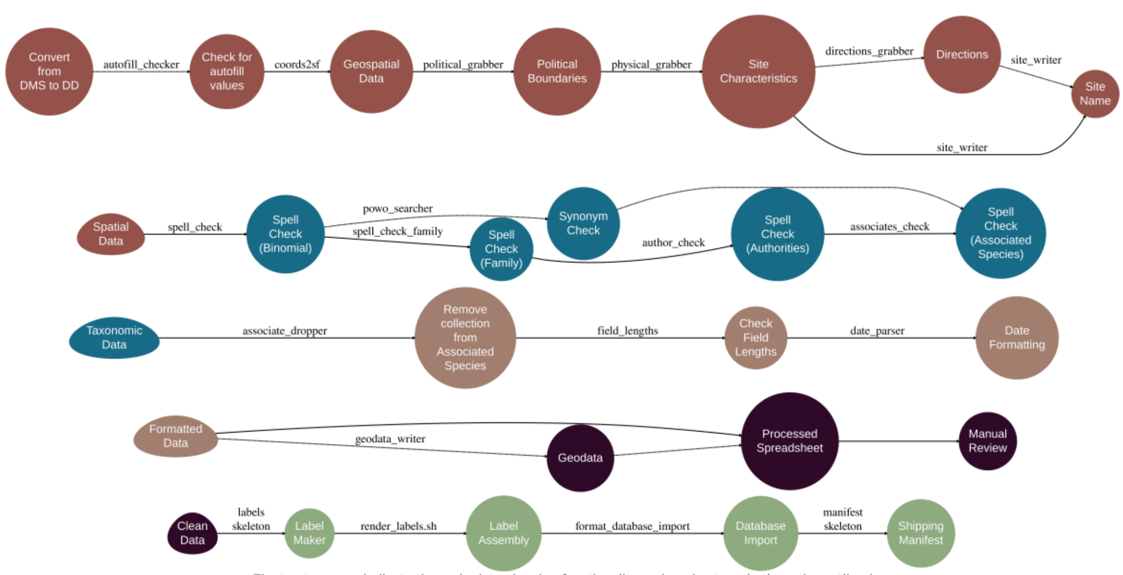
Tosa, M. I., E. H. Dziedzic, C. L. Appel, J. Urbina, A. Massey, J. Ruprecht, C. E. Eriksson, et al. 2021. The rapid rise of next-generation natural history. *Frontiers in Ecology and Evolution* 9: 698131.

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.

Figures



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.

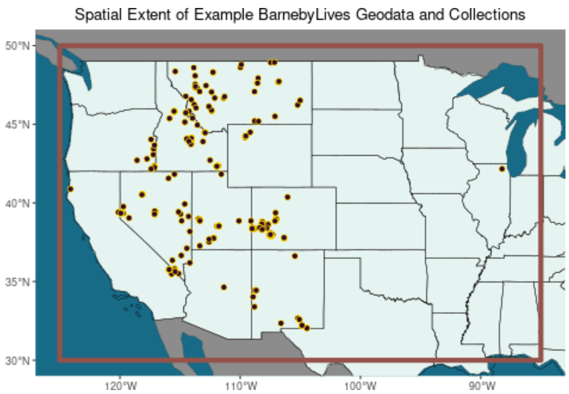


Figure 2 The spatial extent-or domain- (orange), and herbarium collection sites (burgundy) tested in this manuscript.

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

Assess, Inventory, and Monitor

ASTERACEAE

Tetraneuris ivesiana Greene

U.S.A., Colorado, Montrose Co., Uncompahgre Plateau, BLM Uncompahgre FO 48N 11W 35. 0.4mi at 138° from Cottonwood crk. 38.36884 -108.05796 (NAD83 +/- 5m).

Sandstone soils above cliff face. At 7,930 ft (2,417 m), on a slope, 15° slo. 257° asp.; geology: Sedimentary, clastic.

Veg.: *Amelanchier alnifolia* var. *utahensis, Quercus gambelii, Cercocarpus montanus, Symphoricarpos rotundifolius, Artemisia tridentata* var. *wyomingensis, Petradoria pumila, Gutierrezia sarothrae, Bouteloua gracilis.* Ass.: *Petradoria pumila, Eremogone congesta, Heterotheca villosa.*

Reed Clark Benkendorf 2759, Hannah Lovell; 28 Jul, 2022. Fide: *Flora of Colorado*, det.: R.C. Benkendorf, 31 Dec, 2022.

Figure 4. Example label.

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

Reed Clark Benkendorf[1,2]*, 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: Accessioning herbarium specimens is labor-intensive, yet remains vital for research in ecology, evolution, and conservation. As institutional support for herbaria declines, efficient tools are needed to streamline this process. BarnebyLives was developed to assist collectors by supplementing collection notes, verifying taxonomic data, conducting quality checks, generating labels, and submitting digital records.

Methods and Results: It integrates geospatial data from U.S. government sources to provide jurisdictional and site information, and checks taxonomic names using in-house spell checkers, IPNI author standards, and Kew's Plants of the World Online. Optional features include generating Google Maps driving directions. The tool outputs data in tabular and spatial formats for review before producing LaTeX-based labels and shipping manifests.

Conclusions: BarnebyLives improves data accuracy, ensures up-to-date taxonomy, and significantly reduces the time and effort required to accession herbarium specimens.

INTRODUCTION

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 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

*Author for Correspondence: rbenkendorf@chicagobotanic.org

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 and grearly expanded their roles in science (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 electronic and remote 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 will ensure that the specimens remain the premier botanical data source into perpetuity.

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 commercially developed apps on smartphones (Daru et al., 2018; Mishler et al., 2020; Manzano and Julier, 2021). While many novice botanists are capable of using dichotomous keys and other resources to reliably identify and collect satisfactory material, we observe that they face difficulties navigating several aspects of data acquisition, processing, and preparation of labels for submission to herbaria. Some of the 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, novice 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 administrative details in error. Even finding appropriate site names can rarely be resolved without a printed map, as many navigation-related software now consider most features that would serve as site names extraneous. 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 scientific names, even when these names are unanimously accepted by taxonomic specialists in the group and other regional curators (Hitchcock and Cronquist, 2018). Furthermore, formatting a label correctly (e.g., author abbreviations, italicization, etc.) is a time-consuming process with many opportunities to introduce errors in formatting which reduce the apparent credibility of a collector. 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 successfully navigates all these hurdles, the time allocated to each step is quite large, and may discourage them from further collecting.

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 generation of labels. BarnebyLives 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 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 resolved 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, flagged or 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, to our knowledge label generation functionality is provided explicitly by two programs, PLabel and Symbiota, and by the Microsoft Word tool Mail Merge (Gries et al., 2014; Perkins, 2020). The office suite costs money, and in our experience, 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, beyond verifying the nations of collection it does not include data cleaning functionalities. While some sources indicate that it can only be used on Microsoft, we expect it to be usable 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, and most importantly 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 for 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. Several commands in BarnebyLives require output from previous functions, and a workflow that satisfies these requirements is presented in Figure 1.

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 Pages (https://sagesteppe.github.io/BarnebyLives/) and around 100 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 3-arc second (~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. Given that the attributes which the package collects data on are tailored to the Western U.S. region, we do not expect local installs to exceed the size of ours. 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 shell terminal outside of RStudio. Unfortunately, we have not found Windows alternatives to pdfam and pdftek, so we are unable to offer the final label-generating functionality on that operating system, but suspect Ubuntu subsystem for Windows may allow for integration of these tools.

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 write distance and azimuth to collection site from the nearest official named place from the GNIS database. *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.

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).

Please note that the function *physical_grabber* is the one portion of the package where a decoupling may exist between the collection site, and the resolution of the spatial data. While we expect the mismatch to be negligible for all effective purposes relating to: elevation, major geology type, and in general aspect, estimates of slope at this resolution may be biased - generally to lower angles. For these reasons collectors must always make notes on the truly local environment which taxa are found in, and consider that the notes from BL reflect the greater landscape which a microfeature may be present in. While this mis-match will seldom effect landscape ecologists, it may have implications for other data users.

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. Two are optional; however, they are run locally and so quickly that there is no reason to skip them. *date_parser* parses an input date into various formats for notating collection and determination dates on labels. *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.

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. *label_writer* write from a flatfile or spreadsheet to small 4x4 inch herbarium labels (users can modify these dimensions as they see fit). *format_database_import* will write out a spreadsheet of cleaned data in a variety of formats, currently: Jepson, Symbiota, and Consortium of Pacific Northwest herbaria are supported.

Herbarium Collections

[Figure 2 about here.]

The testing of the package within this manuscript was performed using a subset of the authors collections from 2018-2022, while most development was performed on their 2023 and 2024 collections. Only collections which had identifications to the level of species or lower, and transcribed collection dates and coordinates were used for most functionality. In total 980 records were used for testing various functions, these records were 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.

BarnebyLives took roughly four minutes (227.481sec) to run all local steps, and roughly ten minutes (595.294sec) to search Plants of the World Online for preferred synonyms, and a minute 64.869sec to search Google Maps and write directions to sites.

Most of the local run time is attributable to the spatial (209.089sec), and taxonomic operations (17.932sec), while formatting data for labels took 0.46sec. The spell check of the scientific name accounted for nearly all of the time (17.688sec) spent performing local taxonomic operations. The generation of labels consumed around nine minutes (523.5sec) for the rendering, and an additional 61.08sec to combine the 182 sheets to a single Portable Document Format (PDF). The total label generation run time for processing these 728 collections was 15 minutes. In total the 728 collections, which underwent all processing steps, took 25 minutes to process.

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 6 instances where the user entered an unequivocally incorrect family (or taxonomic entity), identified 5 records where families were autofilled, and 1 instance where an outdated circumscription was applied. At the level of family BarnebyLives flagged 6 records where the author follows an alternative taxonomy, and flagged 7 records in error, it appears most of these errors are due to issues in the backbone used by the earlier spell check function.

In the 326 genera analysed BarnebyLives identified 74 discrepancies at the level of genus between user submitted and processed data. In 42 of these instances the user supplied an outdated name (21 unique genera) flagged 4 records where the author follows an alternative taxonomy (2 genera total), and flagged 2 record in error.

Of 728 distinct species analysed BarnebyLives flagged 62 records, and detected 33 instances of misspelled epithets (33 unique species). In 15 of these instances the user supplied an outdated name (15 unique species). It also flagged 2 records where the author follows an alternative taxonomy (2 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. All flags were correct, and in several instances more errors were found in the rows following the flagged values.

[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 assume that responsibility 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.

LaTeX, a software system used for typesetting, allows users to focus on the content rather than the style of the documents rendered from it. However, using its default settings, it can produce aesthetically pleasing results (Figure 4). Additionally LaTeX offers users a wide variety of ways which they can modify labels which are under-explored in the package. Very good documentation of LaTeX capabilities is offered in multiple areas; for instance, via the Overleaf project. While the templates in the package are quite simple, LaTeX also offers the ability to use custom fonts, to alter font weights and colors, alter line spacing, to include images (e.g. dot maps) and customize labels beyond what the default templates support.

Thematically, BarnebyLives is set up to cover Western North America. However, the package supports the use of a 'domain' being drawn over any of the conterminous United States. Several of the attributes which it collects and displays on labels, relate to topics which more senior curators are interested in, i.e. the administrative information on Township Section and Range (or 'TRS'), but are considered less value in other geographic regions.

Further several of the abiotic variables which it acquires information on: slope, aspect, and geology have long been considered prominent drivers of plant distributions in semi-arid and montane systems and warranted on a label in these types of systems, whereas curators in other regions may find this information superfluous. Finally, it is plausible people in other geographic areas are less interested in displaying which land management agency has jurisdiction over a collection; however in the west we believe this is useful information which may help a collector interested in revisiting a site to determine if they will require permits for access or to make new collections.

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 VisualStudio (VS) Code, now offer 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.

[Figure 4 about here.]

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.

ACKNOWLEDGMENTS

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. Several prominent associated collectors of specimens used in this study are thanked: Dani Yashinovitz, Hannah Lovell, Dakota Becerra, 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). The package includes three real use-case vignettes (tutorials) available on a Github Pages site (https://sagesteppe.github.io/BarnebyLives/). The first vignette *"Preparing to use BarnebyLives!"* shows how to set up an instance for a certain geographic area (domain). The next two vignettes *"BarnebyLives! Running pipeline"* showcases the usage of the package for processing data entered on a spreadsheet, and *"Printing herbarium labels and exporting a digital copy of data"* how to export data in both digital and analog formats. *"Custom label templates"* shows how to customize labels in LaTeX, and *"Rendering a shipping manifest"* details how to produce a shipping manifest for gifting or transferring material to an herbarium. 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. Schrodinger'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.

List of Figures

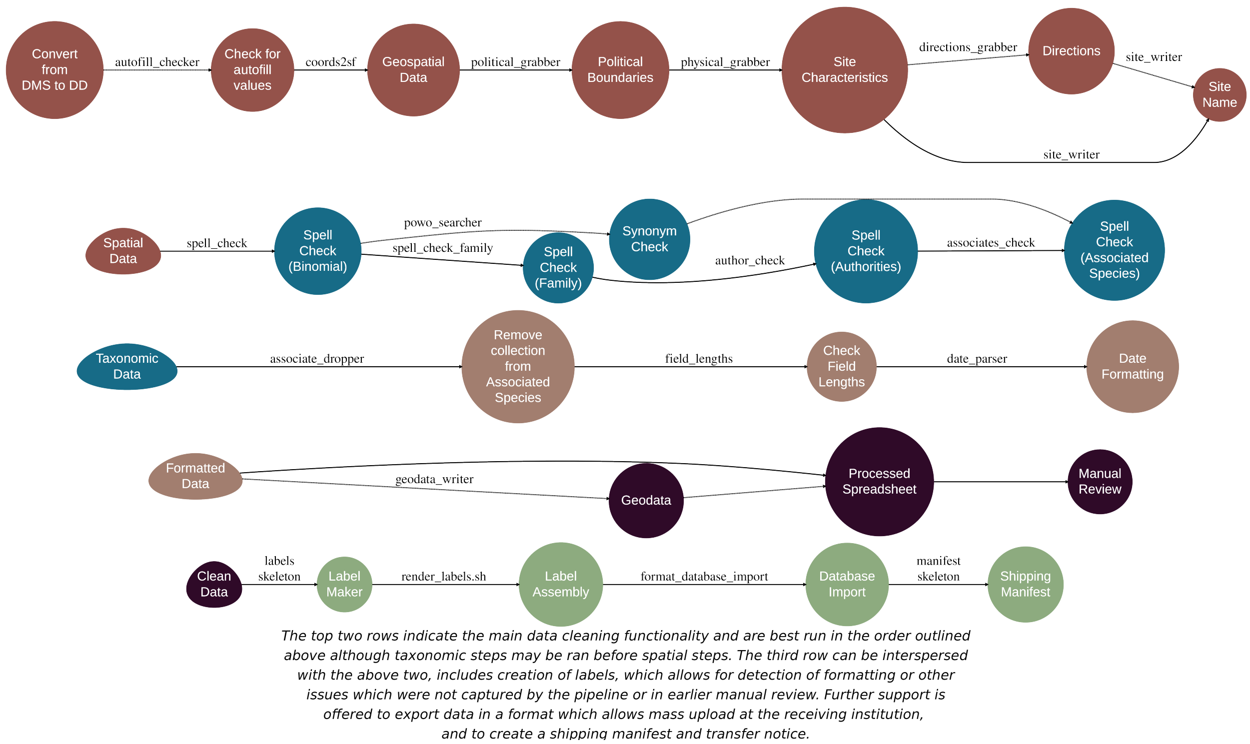


Figure 1: Recommended workflow

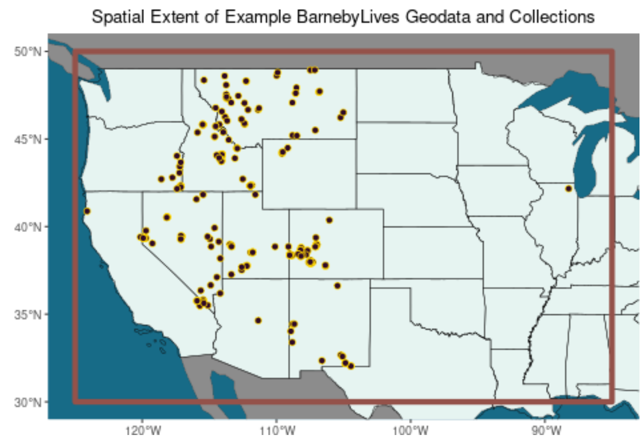


Figure 2: The spatial extent-or domain- (orange), and herbarium collection sites (burgundy) tested in this manuscript.

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

Assess, Inventory, and Monitor

ASTERACEAE

Tetraneuris ivesiana Greene

U.S.A., Colorado, Montrose Co., Uncompahgre Plateau, BLM Uncompahgre FO 48N 11W 35. 0.4mi at 138° from Cottonwood crk. 38.36884 -108.05796 (NAD83 +/- 5m).

Sandstone soils above cliff face. At 7,930 ft (2,417 m), on a slope, 15° slo. 257° asp.; geology: Sedimentary, clastic.

Veg.: *Amelanchier alnifolia* var. *utahensis, Quercus gambelii, Cercocarpus montanus, Symphoricarpos rotundifolius, Artemisia tridentata* var. *wyomingensis, Petradoria pumila, Gutierrezia sarothrae, Bouteloua gracilis.* Ass.: *Petradoria pumila, Eremogone congesta, Heterotheca villosa.*

Reed Clark Benkendorf 2759, Hannah Lovell; 28 Jul, 2022. Fide: *Flora of Colorado*, det.: R.C. Benkendorf, 31 Dec, 2022.

Figure 4: A label generated from a default template