Received: 26 August 2024 | Accepted: 18 November 2024

DOI: 10.1002/aps3.11631

Applications in Plant Sciences

SOFTWARE NOTE

mvh: An R tool to assemble and organize virtual herbaria from openly available specimen images

Thais Vasconcelos[1,2] | **James D. Boyko**[1,3]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[2]University of Michigan Herbarium, University of Michigan, Ann Arbor, Michigan 48108, USA

[3]Michigan Institute of Data Science, University of Michigan, Ann Arbor, Michigan 48109, USA

Correspondence

Thais Vasconcelos, Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA.
Email: tvasc@umich.edu

Abstract

Premise: Recent advances in imaging herbarium specimens have enhanced their use in biodiversity studies. However, user-friendly tools that facilitate the assembly of customized sets of herbarium specimen images on personal devices are still lacking.

Methods and Results: Here we present the R package mvh ("my virtual herbarium"), which includes functions designed to search and download metadata and openly available images associated with herbarium specimens based on taxon or geography. We tested the functionalities of mvh by searching metadata associated with five sets of 10 vascular plant species and five sets of 10 terrestrial coordinates. The download function had a success rate of 99%, downloading 291 out of the 293 images found in the search. Possible reasons for download failure are discussed.

Conclusions: As long as an internet connection is available, mvh simplifies the assembly and organization of virtual herbaria, thereby facilitating the investigation of novel empirical questions as well as trends in digitization efforts.

KEYWORDS

biodiversity, GBIF, natural history collections, specimen digitization

Herbarium collections, originally established to preserve dried plant specimens for pharmaceutical and taxonomic purposes, have evolved into irreplaceable resources of data that can be used to answer a broad range of scientific questions (Funk, 2003; Lavoie, 2013; Davis, 2023). In the past decade, these collections have been used to address such diverse topics as historical changes in plant distribution (Feeley, 2012), extinction risk assessment (Nic Lughadha et al., 2019), tracking changes in phenological patterns associated with climate change (Park et al., 2019), and phylogenomics (Maurin et al., 2021), demonstrating that their applications are difficult to predict but continue to grow in significance. Many of the novel applications of herbarium specimens in biodiversity research have been made possible due to the increased efforts associated with specimen digitization and imaging. As herbarium collections enter the digital era, digitization projects have become a priority for many institutions, both to safeguard the specimens, ensuring that this data source is preserved should anything happen to the physical collections, and to allow botanists easy access to specimens deposited across many institutions (Nelson et al., 2012; Sweeney et al., 2018).

Image data, in particular, has become increasingly important in recent years as the accessibility of computer vision tools has expanded (Weinstein, 2018; Lürig et al., 2021; Hussein et al., 2022). These new tools enable botanists to use large datasets of herbarium specimen images to automate tasks including the extraction of label information (Weaver and Smith, 2023; Weaver et al., 2023), the scoring of phenological stage (Davis et al., 2020), and the measurement of phenotypic traits such as color (Boyko, 2024) and leaf size (Weaver et al., 2020). Of course, the size of the virtual collection and quality of the associated metadata are critical for the successful application of computer vision tools, and therefore they need to be supported by user-friendly applications that allow botanists to easily assemble broad datasets of specimen images. Currently, assembling virtual collections of herbarium images can be slow, unintuitive, and often requires programming knowledge or manual data assembly, which is difficult to track and reproduce.

Here we present the package mvh ("my virtual herbarium"), which includes a set of R functions (R Core Team, 2024) designed to facilitate building datasets of herbarium specimen images and their associated metadata in personal devices, based on taxon or geography. The package interfaces with the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) application programming interface (API) through rgbif (Chamberlain and Boettiger, 2017) to compile a list of specimens with images that are openly available for download. It then accesses the URLs linked to these images and downloads them into a user-determined directory, naming the files according to taxon name and GBIF ID, a unique identifier for each record. The R package mvh is available through GitHub at https://github.com/tncvasconcelos/mvh (see Data Availability Statement).

METHODS AND RESULTS

Functionalities

The main pipeline of mvh includes two functions: *search_specimen_metadata* and *download_specimen_images* (Table 1, Figure 1A). The first function, *search_specimen_metadata*, interacts with rgbif's *occ_search* to search for records of preserved specimens with images in the GBIF dataset. This search can be either taxon-based using the argument *taxon_name*, geography-based using the arguments *coordinates* and *buffer_distance*, or both. Note that, because the function *search_specimen_metadata* is essentially a wrapper of *rgbif::occ_search*, it will take any argument that *rgbif::occ_search* takes; therefore, more experienced R users may want to use this functionality to customize their search in other ways. Any scientific name included in GBIF's taxonomic backbone can be used in the taxon-based search, returning a list of specimens under the currently accepted name of the taxon name used in the search. For instance, a search using "Myrcia fallax" for the argument *taxon_name* will return specimens listed as *Myrcia splendens* (Sw.) DC., the currently accepted name for *Myrcia fallax* (A. Rich.) DC. Similarly, a search using "Myrcia splendens" for the argument *taxon_name* will also return specimens listed as *Myrcia fallax* and all other synonyms of *Myrcia splendens*. Geography-based searches will take any latitude and longitude passed through the *coordinates* argument and create a square around that point of edge length in degrees determined by the *buffer_distance* argument. For instance, if the user inputs "coordinates = c(42.28, −83.74)" and "buffer_distance=1", it will create a square of edge 1 degree with centroid in latitude 42.28°N and longitude 83.74°W. As a result, *search_specimen_metadata* will return a data.frame object containing all metadata associated with the specimens found under the taxon and/or geography criteria, including a column with the URL for the specimen image (column "media_url") and the license regulating data usage. For geography-based searches, it is particularly important to note that the function *search_specimen_metadata* is not able to distinguish between correctly and incorrectly georeferenced specimens. Therefore, further filtering of the metadata using pipelines such as CoordinateCleaner (Zizka et al., 2019) or rWCVP (Brown et al., 2023) before downloading may be useful to remove specimens that are not correctly georeferenced.

The second function in the pipeline, *download_specimen_images*, will take as the main argument the resulting data.frame object from *search_specimen_metadata* to download in .jpg format all images linked to the URL in the "media_url" column (Figure 1B). The naming convention for each image uses a combination of taxon name and GBIF ID, a unique number allowing the user to connect the specimen to the metadata table (given that many GBIF observations lack collector name and number). Other arguments in the

TABLE 1 Main functionalities of the R package mvh.

Function name[a]	Short description	Main arguments	Output
search_specimen_metadata	Interacts with rgbif to search metadata associated with preserved specimens with images available.	*taxon_name* (for taxon-based searches), and *coordinates* and *buffer_distance* (for geography-based searches)	A data.frame including all metadata associated with specimens returned by the search.
download_specimen_images	Takes the output of *search_specimen_metadata* to download images of herbarium specimens through their URLs.	*metadata* (the output from *search_specimen_metadata*)	A data.frame with summarized metadata and the status of the download.
plot_specimens_by_institution	Takes the output of *search_specimen_metadata* to plot a barplot of the institutions where most specimens are deposited.	*metadata* (the output from *search_specimen_metadata*)	A barplot sorted in decreasing order of number of specimens per institution.
plot_specimens_by_country	Takes the output of *search_specimen_metadata* to plot a barplot of the countries where most specimens were collected.	*metadata* (the output from *search_specimen_metadata*)	A barplot sorted in decreasing order of number of specimens per country.

[a]Note that both *search_specimen_metadata* and *download_specimen_images* require an internet connection to run.

A
```
metadata <- search_specimen_metadata(
                taxon_name = "Vaccinium",
                coordinates = c(42.28, -83.74),
                limit=8)
download_specimen_images(metadata,
                dir_name="Vaccinium_in_AnnArbor_example/specimens",
                result_file_name="Vaccinium_in_AnnArbor_example/result_download")
```

B

C
```
Download completed! Don't forget to acknowledge the collections of OUHC & BRIT
if you use the specimens in your research.
```

FIGURE 1 Example script and images downloaded with the mvh pipeline. (A) Example script of an mvh pipeline to search and download up to eight specimens ("limit=8") of the blueberry genus *Vaccinium* (Ericaceae) from the Ann Arbor (Michigan, USA) area ("coordinates = c(42.28, −83.74)"). (B) Specimen images downloaded using the pipeline. (C) Message reminding the user to acknowledge the collections where specimens are deposited if they are used in publications.

download_specimen_images function allow the user to customize and organize the download of the images. The argument *dir_name* will determine the name of the folder created in the working directory at the beginning of the downloading process where the images are going to be saved. The *result_file_name* argument allows the user to rename the results table to be written in the working directory with the download status. This results table (Figure 2) is a spreadsheet created in the designated working directory containing summarized metadata about the specimens returned by the search. As the function attempts to download each of the images returned by the search, it will write in the last two columns of the spreadsheet whether the download was successful or not (column "status") and, if the download failed, the message associated with the error for further investigation (column "error_message") (e.g., so the user knows if the download failed due to an issue of broken URL or internet instability). The spreadsheet will also return the license regulating the use of the image (if available), the holder of the rights to the image, and the image size in bytes.

scientificName	gbifID	institutionCode	eventDate	country	original_filesize	megapixels	status	error_message
Vaccinium corymbosum L.	4442292449	OUHC	2023-04	United States of America	547868	15.36	succeeded	NA
Vaccinium corymbosum L.	4454196306	BRIT	5/18/23	United States of America	5201155	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454199306	BRIT	5/18/23	United States of America	5294834	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454197307	BRIT	5/18/23	United States of America	5331727	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454198308	BRIT	5/18/23	United States of America	6233814	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454195309	BRIT	5/18/23	United States of America	5343762	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454194307	BRIT	5/18/23	United States of America	4929296	22.1184	succeeded	NA
Vaccinium corymbosum L.	4454196307	BRIT	5/18/23	United States of America	4943905	22.1184	succeeded	NA

FIGURE 2 Summarized example of a results table from a search and download of up to eight specimens ("limit=8") of the blueberry genus *Vaccinium* (Ericaceae) from the Ann Arbor (Michigan, USA) area ("coordinates = c(42.28, −83.74)"). The columns "license" and "rightHolder" were omitted due to limited space.

The *resize* argument will specify whether the images should be resized after the download. The default of this argument is NULL, meaning that if nothing is changed, the images will keep their original quality and size after download. However, if the user chooses to use this argument, it will take any value from 1 to 100 to determine how much the size and quality of the original file should be reduced. For instance, a resize=75 will reduce the original image to 75% of its original size. Depending on the size of the herbarium images and the user's storage capacity, the *resize* feature can be essential when building large collections. Herbarium specimen images of high enough quality to allow the study of relatively small morphological structures (e.g., trichomes) are large in terms of file size, and building a virtual herbarium of 100 specimen images of original quality may require over 1 GB of storage space in the personal device. While the *resize* feature can be beneficial, it should be used cautiously. Reducing image size may negatively impact the accuracy of automated methods of trait scoring such as LeafMachine (Weaver et al., 2020) as it is often resolution dependent. The user will need to evaluate the use of the argument on a case-by-case basis. With that in mind, we added an additional argument called *max_megapixels* to this function. This argument reduces the quality of images to the specified maximum megapixels set by the user but does not alter images that are already below that resolution. It will then reduce the file size of very large images, while leaving lower-resolution images unchanged.

Finally, the argument *timeout_limit* will determine how long mvh should spend trying to download an image before the connection fails. The default for the *timeout_limit* is 300 seconds, meaning that each download will try to connect to the network for 5 minutes before crashing. Adjusting this argument can be helpful if the user has an unstable internet connection and requires a longer buffer time to complete a given download. At the end of the download, the function will also print on the console a message reminding the user to acknowledge the collections where the images come from if they are used in publications (Figure 1C).

The package mvh also includes two plotting functions to allow the user to visualize two important components of the metadata associated with the search. The first is in which countries the collections in their personal herbaria were collected (*plot_specimens_by_country*) and the second is in which institutions they are currently deposited (*plot_specimens_by_institution*), which follows the herbarium acronyms used by GBIF's Global Registry of Scientific Collections (https://scientific-collections.gbif.org/) (Table 1). Both functions will take as argument the data.frame object resulting from the *search_specimen_metadata* function to plot a barplot showing the number of specimens in each category in decreasing order (Figure 3).

Study case

We tested the functionality of the mvh package by running the main pipeline in two study cases, one based on taxon and one on geography. In the first, we took five random samples of 10 species each from the list of accepted species of vascular plants from the Plants of the World Online (POWO) database (https://powo.science.kew.org/). In the second, we used as coordinate input the latitude and longitude of 10 global national parks. Because the intention was merely to test the success rate of the downloads, in both cases we set the *limit* argument of *search_specimen_metadata* to "limit=5" for speed, meaning the search will be limited to five specimens. The *resize* argument of *download_specimen_images* was set to "resize=5", allowing the downloaded images to be resized to 5% of their original quality. The pipeline was run at an internet speed of 350 Mbps using a MacBook Pro macOS Ventura 13.1 (Apple, Cupertino, California, USA). The full script for the search, including our seed number for the random sample of species, latitude and longitude of the 10 global national parks, and code to summarize the results are available in Appendices S1–S3 (see Supporting Information with this article).

The pipeline found 242 specimen images in the taxon-based search, successfully downloading 240 of them in 22.32 minutes, resulting in a success rate of 99% and an average of 5.58 seconds per specimen. In the geography-based search, the pipeline found 51 specimen images and successfully downloaded 51 of them in 6.42 minutes, for a 100% success rate and an average download speed of 7.55 seconds per image. The original size of the files at the time of download ranged from 4.34 kB to 80.01 MB, with a

A

```
metadata <- search_specimen_metadata(taxon_name = "Myrcia splendens",limit=100)
plot_specimens_by_institution(metadata)
plot_specimens_by_country(metadata)
```

B

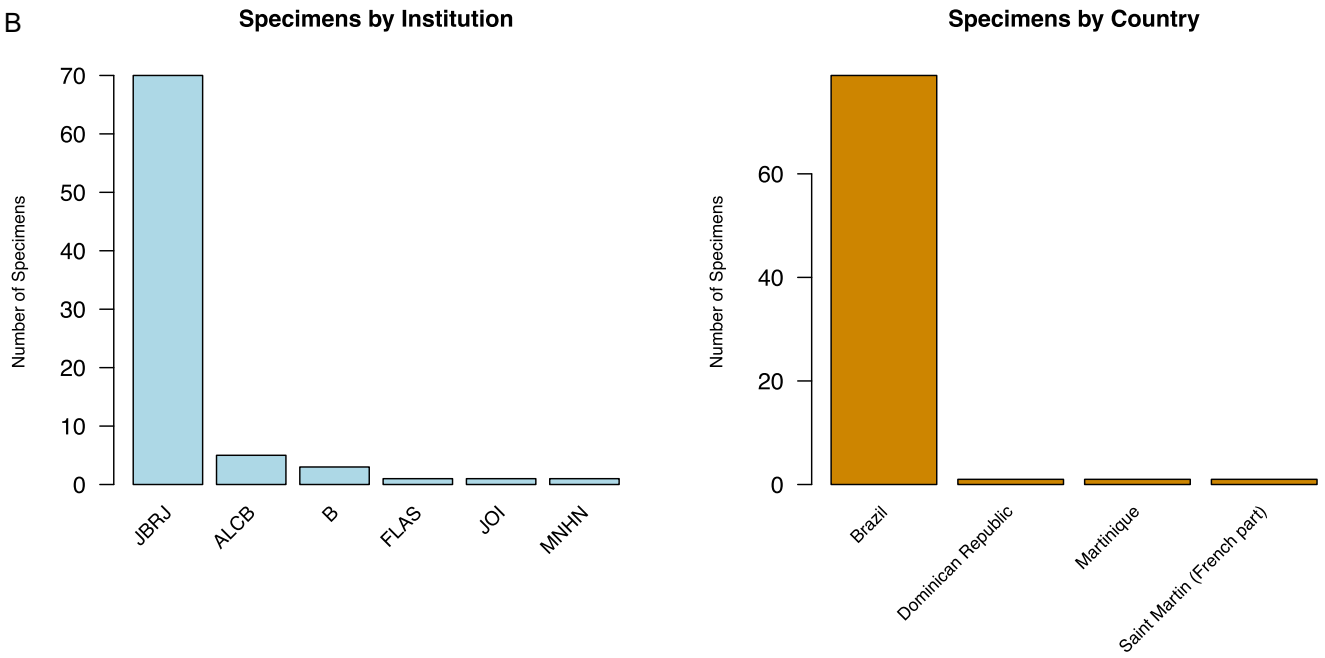

FIGURE 3 Example script and plots generated by the mvh pipeline. (A) Example script of an mvh pipeline to search up to 100 specimens ("limit=100") of the widespread species *Myrcia splendens* (Myrtaceae) and (B) the resulting barplots of the number of specimens per institution and country. Herbarium acronyms follow GBIF's Global Registry of Scientific Collections (https://scientific-collections.gbif.org/).

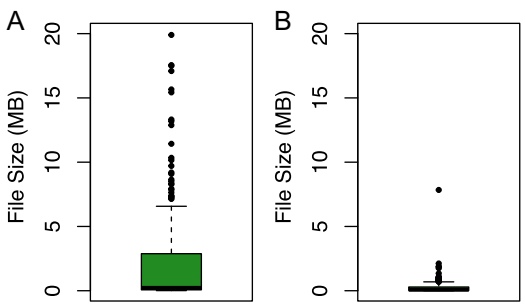

FIGURE 4 Box plots illustrating the file sizes (A) before ("resize=100") and (B) after ("resize=5") resizing.

median of 226.4 kB (Figure 4A), requiring 1.26 GB of space to complete the download at the original quality. After resizing images to 5% of their original quality, however, the 291 downloaded images occupied a total of 109.2 MB of space, with images ranging from 1.97 kB to 7.83 MB, with a median of 78.80 kB (Figure 4B).

Comparison with other software and additional notes

To our knowledge, there is currently no other user-friendly pipeline to download and organize images of herbarium specimens available as an R package, although it is possible to perform an automated download of specimen images in R through a combination of functions in the R package rgbif and some programming. An advantage of using GBIF as a base for our URL search is that it is not only the largest data aggregator available for biodiversity data, but it also allows orthographic variants linked as synonyms in the GBIF taxonomic backbone to be matched to the search term and returned through the function *search_specimen_metadata* (as is the case for *rgbif::occ_search*). It is important to note that our pipeline only generates a digital object identifier (DOI) for a proper dataset citation, as recommended by the GBIF guidelines, if the username, password, and email linked to the user's GBIF account are added as arguments in the *search_specimen_metadata* function. We advise users to use this option if images or metadata are used in publications, and to cite the generated DOI accordingly. Properly acknowledging the collections where the specimen data comes from is crucial, as these are the entities responsible for data availability.

Limitations

As we worked on the development of this package and in improving the success rate of the downloads, we observed that, as long as the metadata has a valid URL in the identifier

slot of the media file on GBIF, the download is very likely to be successful. However, two main types of errors in failed downloads were recurrent: (1) error of URL not existing or not linked with an image, which will appear as "cannot open URL [url]" in the "status" column of the results table; and (2) error of time limit for download reached, which will appear as "download from [url] failed" in the "status" column of the results table. The first error is difficult to address because it can be caused by many different issues, including human error during specimen digitization (URL in the wrong format or in the wrong data slot) and media data that used to exist but have since been excluded by the collections after being exported to GBIF. Different collections export their metadata in different formats, and we extensively tested the functionalities of mvh so that it is able to capture this variation for most of the large collections. The second error can usually be dealt with by increasing the *time.limit* arguments of the *download_specimen_images* function. In fact, we observed that the percentage of failed downloads in very large queries decreased from 2% (98% success rate) to 1% (99% success rate) once a *time.limit* of 300 seconds was set, which is why we set this as the function's default.

Another observed source of download failure is authentications such as CAPTCHAs that some collection websites require to access images through the URL available on GBIF. In those cases, we observed that after the conditions for accessing the image have been accepted once, all subsequent attempts to download images using our R pipeline proceed normally when the same computer is used. Therefore, we advise users to manually access the URLs printed in the "error_message" column of the results table and rerun the pipeline after verifying if this is the reason for failed downloads. An additional limitation is that, because the function *download_specimen_images* works by accessing the URL of the specimen image as provided by the original institutions, it will occasionally download images that are not of herborized material (e.g., living plants) if their URL address is indistinguishable from that of the equivalent herbarium specimen. This happens occasionally in large queries, and we currently have found no way to avoid this. However, we anticipate that this will not be a big issue with most projects. Finally, as the current version of mvh is essentially a wrapper for rgbif, it comes as no surprise that it can only find and download images available on GBIF. While we do not consider this a major limitation, as most digitized herbarium specimens eventually make their way to GBIF, future versions of this package may allow users to search smaller data aggregators as well. We note that this also means that most queries will only return a subset of the potential universe of specimens available for that taxon and/or geographical area.

CONCLUSIONS

The R package mvh provides a flexible and user-friendly pipeline to download images of herbarium specimens deposited across institutions. It can accelerate and facilitate the use of this information for taxonomy, floristics, phenotypic analyses, and biodiversity studies in general.

Currently, most herbarium collections are not completely digitized, so virtual herbaria are not perfect representations of the whole range of physical specimens deposited in these institutions. However, we do think that tools like this package will become increasingly useful as more collections are being digitized and imaged, and novel uses of herbarium specimen images are unlocked.

AUTHOR CONTRIBUTIONS

T.V. and J.D.B. conceptualized and developed the software. T.V. wrote the manuscript with input from J.D.B. Both authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank Will Weaver (University of Michigan) for conversations that improved this manuscript and the mvh pipeline. We also thank the following herbaria for making their specimen images openly available for non-commercial use: A, B, BAYLU, BBM, BR, BRI, BRIT, BRLU, BRY, CHR, CM, COI, COLO, CR, DBG, DES, E, F, GH, GJO, HIFP, JBRJ, K, KYO, LBV, MEL, MICH, MNHN, MO, MW, NBF, NCU, NEBC, NEON, NGCPR, NHMUK, NMNZ, NSW, NY, P, PRC, RBGE, RSA, TASM, TRH, UCSB, UCR, US, W, WS, WU, Z (all acronyms following GBIF's Global Registry of Scientific Collections [(https://scientific-collections.gbif.org/]).

DATA AVAILABILITY STATEMENT

The R scripts and data necessary to reproduce the results of this work are available in the Supporting Information. The R package mvh is available through GitHub at https://github.com/tncvasconcelos/mvh.

ORCID

Thais Vasconcelos http://orcid.org/0000-0001-9991-7924

REFERENCES

Boyko, J. 2024. SegColR: Deep learning for automated segmentation and color extraction. bioRxiv 2024-07 [Preprint]. Available at https://doi.org/10.1101/2024.07.28.605475 [posted 29 July 2024; accessed 31 December 2024].

Brown, M. J., B. E. Walker, N. Black, R. H. Govaerts, I. Ondo, R. Turner, and E. Nic Lughadha. 2023. rWCVP: A companion R package for the World Checklist of Vascular Plants. *New Phytologist* 240(4): 1355–1365.

Chamberlain, S. A., and C. Boettiger. 2017. R Python, and Ruby clients for GBIF species occurrence data. *PeerJ Preprints* 5: e3304v1.

Davis, C. C. 2023. The herbarium of the future. *Trends in Ecology & Evolution* 38(5): 412–423.

Davis, C. C., J. Champ, D. S. Park, I. Breckheimer, G. M. Lyra, J. Xie, A. Joly, et al. 2020. A new method for counting reproductive structures in digitized herbarium specimens using mask R-CNN. *Frontiers in Plant Science* 11: e1129.

Feeley, K. J. 2012. Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology* 18(4): 1335–1341.

Funk, V. A. 2003. The importance of herbaria. *Plant Science Bulletin* 49(3): 94–95.

Hussein, B. R., O. A. Malik, W. H. Ong, and J. W. F. Slik. 2022. Applications of computer vision and machine learning techniques for digitized herbarium specimens: A systematic literature review. *Ecological Informatics* 69: e101641.

Lavoie, C. 2013. Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics* 15(1): 68–76.

Lürig, M. D., S. Donoughe, E. I. Svensson, A. Porto, and M. Tsuboi. 2021. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in Ecology and Evolution* 9: e642774.

Maurin, O., A. Anest, S. Bellot, E. Biffin, G. Brewer, T. Charles-Dominique, R. S. Cowan, et al. 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *American Journal of Botany* 108(7): 1087–1111.

Nelson, G., D. Paul, G. Riccardi, and A. R. Mast. 2012. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* 209: 19–45.

Nic Lughadha, E., B. E. Walker, C. Canteiro, H. Chadburn, A. P. Davis, S. Hargreaves, E. J. Lucas, et al. 2019. The use and misuse of herbarium specimens in evaluating plant extinction risks. *Philosophical Transactions of the Royal Society B, Biological Sciences* 374(1763): e20170402.

Park, D. S., I. Breckheimer, A. C. Williams, E. Law, A. M. Ellison, and C. C. Davis. 2019. Herbarium specimens reveal substantial and unexpected variation in phenological sensitivity across the eastern United States. *Philosophical Transactions of the Royal Society B, Biological Sciences* 374(1763): e20170394.

R Core Team. 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website https://www.R-project.org/ [accessed 31 December 2024].

Sweeney, P. W., B. Starly, P. J. Morris, Y. Xu, A. Jones, S. Radhakrishnan, C. J. Grassa, and C. C. Davis. 2018. Large-scale digitization of herbarium specimens: Development and usage of an automated, high-throughput conveyor system. *Taxon* 67(1): 165–178.

Weaver, W. N., J. Ng, and R. G. Laport. 2020. LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens. *Applications in Plant Sciences* 8(6): e11367.

Weaver, W. N., and S. A. Smith. 2023. From leaves to labels: Building modular machine learning networks for rapid herbarium specimen analysis with LeafMachine2. *Applications in Plant Sciences* 11(5): e11548.

Weaver, W. N., B. R. Ruhfel, K. J. Lough, and S. A. Smith. 2023. Herbarium specimen label transcription reimagined with large language models: Capabilities, productivity, and risks. *American Journal of Botany* 110(12): e16256.

Weinstein, B. G. 2018. A computer vision for animal ecology. *Journal of Animal Ecology* 87(3): 533–545.

Zizka, A., D. Silvestro, T. Andermann, J. Azevedo, C. Duarte Ritter, D. Edler, H. Farooq, et al. 2019. Coordinate Cleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10(5): 744–751.

SUPPORTING INFORMATION

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

Appendix S1. Latitude and longitude of the 10 global national parks used in the study.

Appendix S2. List of accepted species of vascular plants from the Plants of the World Online (POWO) database used in the study.

Appendix S3. R code to reproduce the study.

How to cite this article: Vasconcelos, T., and J. D. Boyko. 2025. mvh: An R tool to assemble and organize virtual herbaria from openly available specimen images. *Applications in Plant Sciences* 13(2): e11631. https://doi.org/10.1002/aps3.11631