Received: 6 June 2023 | Accepted: 1 October 2023

DOI: 10.1002/aps3.11563

Applications in Plant Sciences

APPLICATION ARTICLE

FloraTraiter: Automated parsing of traits from descriptive biodiversity literature

Ryan A. Folk[1] | **Robert P. Guralnick**[2,3] | **Raphael T. LaFrance**[2]

[1]Department of Biological Sciences, Mississippi State University, Mississippi State, Mississippi, USA

[2]Florida Museum of Natural History, University of Florida, Gainesville, Florida, USA

[3]Biodiversity Institute, University of Florida, Gainesville, Florida, USA

Correspondence

Ryan A. Folk, Department of Biological Sciences, Mississippi State University, 295 E. Lee Blvd., P.O. Box GY, Mississippi State, Mississippi 39762, USA.
Email: rfolk@biology.msstate.edu

Abstract

Premise: Plant trait data are essential for quantifying biodiversity and function across Earth, but these data are challenging to acquire for large studies. Diverse strategies are needed, including the liberation of heritage data locked within specialist literature such as floras and taxonomic monographs. Here we report FloraTraiter, a novel approach using rule-based natural language processing (NLP) to parse computable trait data from biodiversity literature.

Methods: FloraTraiter was implemented through collaborative work between programmers and botanical experts and customized for both online floras and scanned literature. We report a strategy spanning optical character recognition, recognition of taxa, iterative building of traits, and establishing linkages among all of these, as well as curational tools and code for turning these results into standard morphological matrices.

Results: Over 95% of treatment content was successfully parsed for traits with <1% error. Data for more than 700 taxa are reported, including a demonstration of common downstream uses.

Conclusions: We identify strategies, applications, tips, and challenges that we hope will facilitate future similar efforts to produce large open-source trait data sets for broad community reuse. Largely automated tools like FloraTraiter will be an important addition to the toolkit for assembling trait data at scale.

KEYWORDS

biodiversity literature, flora, functional trait, language model, natural language parsing

Botanists have been gathering information on plant traits, which comprise the entirety of measurable aspects of plant phenotypes, since the dawn of scientific pursuit (Figure 1A). With traits scored for many species, biologists can ask questions about phenotypic and functional differences spanning the plant tree of life, and within and across ecological communities. These questions are fundamental because traits underlie how species interact with and adapt to their surroundings and how we as humans interact with them (Freudenstein et al., 2016). Despite the long history of studying plant traits, however, this information remains mostly unavailable in a form that is usable for quantitative analysis (Hortal et al., 2015). A shortage of "computable" trait data has a very real effect on our view of global plant diversity (Pakeman and Quested, 2007; Pakeman, 2014; Sandel et al., 2015; Cornwell et al., 2019) and tends to be more pronounced in quantitative terms than other plant data and information domains such as DNA and geographic data (Sandel et al., 2015; Folk et al., 2018). This lack is distributed along biogeographic, socioeconomic, political, and other axes that impact the science performed in regions of the world where this same information is urgently needed (Meyer et al., 2015; Daru et al., 2018; Cornwell et al., 2019).

If plant traits are needed and missing, which traits should we measure and how do we best measure them? As argued by Violle et al. (2014) and Hortal et al. (2015), "the traits that are generally measured are often the most simple, rather than the most functional" (Hortal et al., 2015, p. 529),

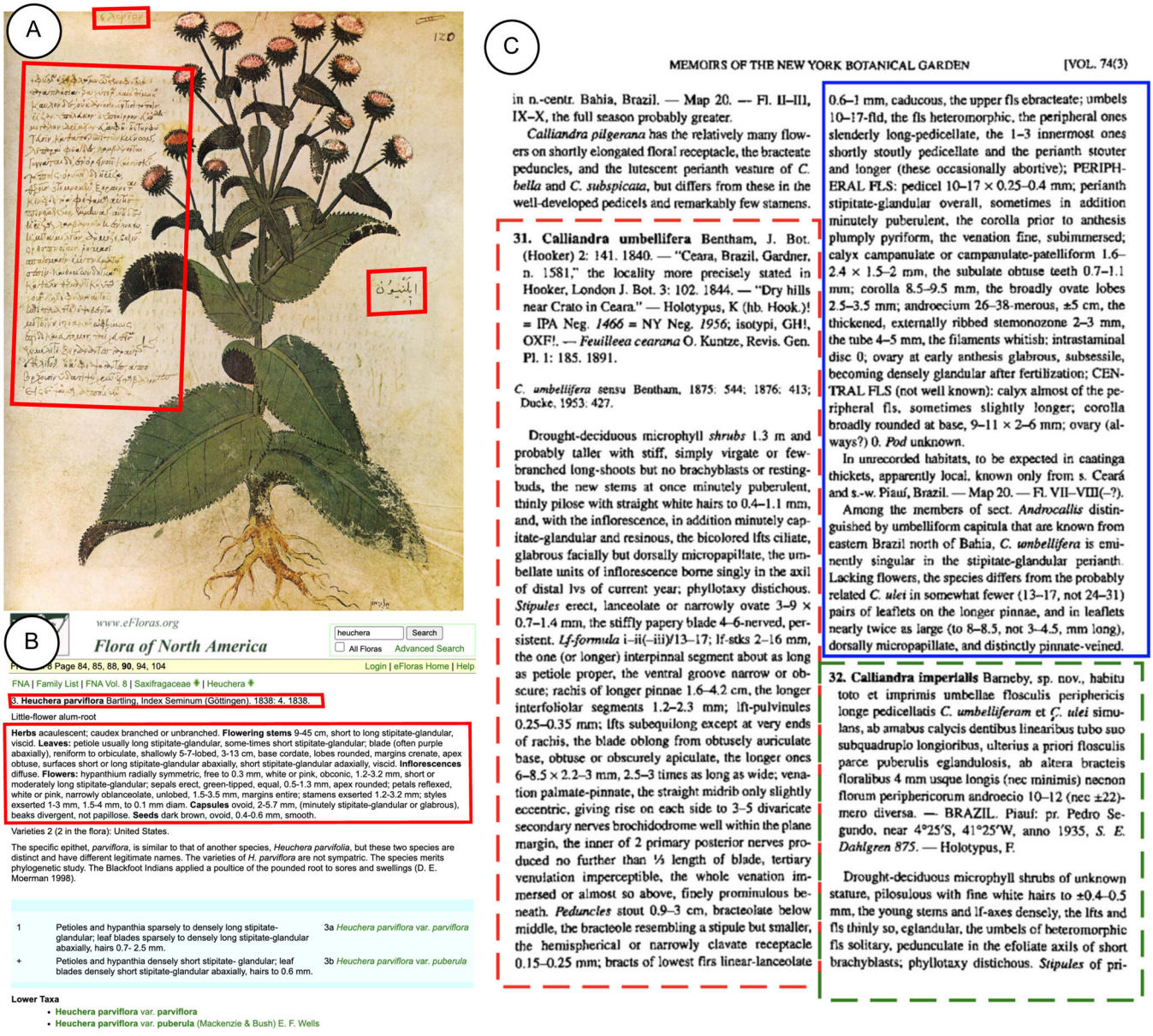


FIGURE 1 Anatomy of a piece of biodiversity literature. (A) The collection of plant trait information, while a modern challenge, has a long history. The *Vienna Dioscorides*, a manuscript from the early sixth century, represents some of the earliest direct and relatively complete survival of scientific botanical descriptions in the Western tradition and illustrates the long history of study that underlies our current understanding of plant traits. The structure of the entry for *Inula helenium* L. follows many aspects of a modern description and therefore illustrates concepts, with its taxon names marked in Greek (ἑλενίου, marked in red, upper left) and Arabic (راسِن, marked in red, right), as well as the original description (marked in red, lower left) including distribution, morphology, habitat, and medical uses, in that order. Our purpose here is to break down the latter (the *description*) to basic *traits* and link these to the former (the *taxon*). (B) Entry for *Heuchera parviflora* Bartl. from *Flora of North America*, an example of a modern online flora, where the text is already machine-readable and relevant portions of the description structure can be fetched by following links and cleaning up HTML tags. (C) A more difficult use case from a monographic work treating *Calliandra*; while also a recent treatment, this is a scanned text that had to be subjected to optical character recognition (OCR) according to the segmentation strategy reported in the Methods section to generate machine-readable text. Structural features like page breaks and page headers are not description or taxon text and therefore should be removed from OCR. This was done a priori here, with the boxes marking the start of the first full treatment on the page; the blue box continues this treatment, and the green box indicates a second treatment (this is a screenshot of one of the processing tools presented here; see Methods). All of these text boxes had to be annealed correctly before further processing. Also illustrated here are various formatting peculiarities (capitalization, italicization, font size, indentation, unusual abbreviations, multilingual material, some of which parallel [A]) that vary by work and can be variously complex or inconsistent; including or discarding these is the subject of optimizing particular sources.

whereas many ecologists might prefer the traits most strongly linked to ecosystem function be measured first. Rather than ask which traits are best measured anew, however, it is equally important to assure that we best leverage the vast information we already have. Information about many traits measured across the entire plant body for many species has long been disseminated by specialists, but these data are locked up in the inaccessible form of

biodiversity literature (Rinaldo, 2009; Thessen et al., 2012; Folk et al., 2018; Penev et al., 2019; Folk and Siniscalchi, 2021; Shirey et al., 2022). For our purposes, "biodiversity literature" refers to all forms of scientific and para-professional output that contains summary statements regarding *taxa* and their *traits*. This includes species descriptions, floras, field guides, monographic revisions, and similar works (examples in Figure 1). Such biodiversity literature is unique in that, rather than representing a directly verifiable measurement on a single organism, these resources provide measurements that represent an expert's judgment regarding a set of observations contingent on a taxon hypothesis. A literature-derived trait therefore has the downside that it is only as good as the taxonomic delimitation that underlies it. This same property is a key strength: unlike other sources of data, biodiversity literature represents the assessment of a domain expert, such as a botanist specializing in a family of plants, and is therefore likely to represent the state of the art at that time regarding taxon boundaries and attributes. An expert is able to identify and exclude diseased and underdeveloped plant organs, can interpret the sometimes-complex structural homology between plant species, and characteristically will focus on the structures most variable among closely related organisms. Biodiversity literature extraction therefore holds promise as a major source of plant trait information and a complement to other sources, but despite several previous proof-of-concept reports (Cui et al., 2016; Endara et al., 2018), the botanical community still lacks a general-purpose approach that can generate high-quality data for many species.

In this paper, we report on work that far extends pipelines originally designed for vertebrate measurements (dubbed Traiter in Guralnick et al., 2016) to extract morphological data from floristic publications. Traiter originally used regular expressions (referred to as "regex") to process text data about body length and mass from vertebrate specimen records and produce harmonized, standardized quantitative measurements. However, a regex approach has some important downsides when applied to unstructured publication text. First, it relies on text data sources that are already highly structured, such as specimen labels with free text entry for body size, embryo counts, and similar content. While there are structured reports of traits in floras or other resources, much of the data is presented in prose rather than presorted into trait categories, which often requires extensive human effort to curate the resulting parses. Second, regexes rely on patterns of characters, making it difficult to use other information not contained in the characters themselves, such as that implied by parts of speech or sentence structure. For instance, consider the description of *Comptonia peregrina* (L.) J. M. Coult. from *Flora of North America* (Bornstein, 1997), according to which the leaves are "3–15.5 × 0.3–2.9 cm, lobes alternate to nearly opposite, base truncate, cuneate to attenuate, or oblique, apex acute; surfaces abaxially pale gray-green, densely pilose to puberulent, adaxially dark green, densely pilose to glabrate, gland-dotted, especially adaxially." Correctly understanding the information in this description requires recognizing several things: (1) that every structure mentioned is a subpart of a leaf (our data set must therefore represent or at least be cognizant of hierarchical structure); and (2) that the given measurements apply to the entire leaf but all other descriptors belong to subparts (we cannot rely on word order even though the adjectives follow the nouns here, according to a typical regex strategy; this varies between publications or even between sentences). In both cases, we must use context to understand that the first reported numbers represent length and the second are width, and we must navigate qualifiers ("especially", "nearly"), positional information ("adaxially"), and multiple measurements with complex relational prepositions and conjunctions ("cuneate to attenuate, or oblique"). In short, a flexible method to extract data from text like this must be able to understand the structure of a typical sentence from a taxon description.

Here we report a new approach called FloraTraiter that shifts to using a natural language processing (NLP) approach to more flexibly extract traits. The specific form of NLP used here, rule-based parsing, leverages preexisting language models to break down biodiversity descriptions into parts of speech, with an extended vocabulary to handle technical botanical descriptions. Then, a series of newly implemented steps further process structural elements that pertain specifically to a biodiversity description, beginning with recognizing a taxon and then using an interactive process to identify partial traits and map them to taxa. We test the new approach on descriptions of the plant order Fagales as reported in online treatments on the eFloras website (http://www.efloras.org) and implement code to translate parsed descriptions into a recognizable morphological matrix format. We then score trait extractions against human observers examining the same source references to count false negatives and false positives, giving us a quantification of traits that are missed or incorrectly scored, respectively. Finally, we demonstrate some standard downstream uses for the trait extractions.

METHODS

Parsing raw data

Parsing begins with literature that has either already been entered digitally or is an image file that must be subjected to optical character recognition (OCR). OCR is the most complex task and so this will be described first and at length. Commercial solutions for PDF OCR are available, such as ABBYY FineReader (ABBYY, Milpitas, California, USA), but a custom approach was used because standard OCR packages were found unsuitable after extensive testing. Problems with standard OCR include (1) OCR text, while appropriately placed on the page, is commonly not in reading order when extracted; (2) sometimes text on the

page is missing; (3) characters are frequently substituted for other similar-looking characters (such as exchanges involving u, n, m); and, most importantly, (4) many aspects of page structure are a nuisance irrelevant to the data contained in a page; these aspects increase the challenge of parsing traits, and especially linking them back to taxa when descriptions span pages. Points 3 and 4 will be discussed further below.

The first step is to convert the PDF into images, one image per page. This function is captured in "pdf_to_images.py" (hereafter all quoted scripts are found at: https://github.com/rafelafrance/FloraTraiter unless otherwise specified; see Data Availability Statement), which is a wrapper around the pdftpcairo program, a module in poppler-utils (pre-compiled command-line programs from the Poppler library; https://poppler.freedesktop.org/) to be installed separately. Second, the PDF images are manually segmented by drawing bounding boxes around text in reading order and marking which ones denote the start of a treatment; these bounding boxes indicate where each treatment starts and ends while marking the order of the treatment text. This is done in a Python script called "slice.py". The colors indicate the order in which a page is read, with red always denoting the first box, blue the second, etc.; the dashed box outlines indicate a treatment start (Figure 1C shows an actual session using this program). Only treatment text is outlined, whereas figures, captions, headers, and similar materials are left out. This is essential because large amounts of structural information in text is a nuisance and must be discarded, and the flow of text on the page is not always easily determined programmatically. Nuisance text includes page breaks and number breaks, as well as material that does not contain formal descriptions, such as literature surveys and indices. Handling this cleanly is important because successfully identifying breaks in descriptions is essential for discovering links between traits and the taxa they belong to.

Once the PDF images are segmented, OCR is conducted on the text in each bounding box; boxes of text are then stitched together with markers in the text to indicate the start of treatments. This script is called "stitch.py". Finally, common OCR errors are corrected and the text is normalized with "clean_text.py". As a corollary, a limited number of typographical errors can be corrected, but in general high-quality and well-aligned scans are needed to produce good OCR content.

Parsing the raw data for HTML sources is much simpler and involves "spidering" (i.e., iteratively traversing the structure of) a base webpage for a taxon and pulling taxonomic treatments guided by HTML markup. Code for this purpose is also available at https://github.com/rafelafrance/FloraTraiter.

Controlled vocabulary

The approach described in the following section relies on prebuilt language models, but these are generally trained on text intended for a wide audience and lack vocabulary on specialist topics. Conversations between botanists and programmers on this project led to us identifying a basic botanical trait vocabulary to add to the model; early drafts were distributed to organismal specialists for comment, which led to identification of missing technical words (the final vocabulary developed is integrated throughout the GitHub repository in relevant processing steps). Fortunately, differing sources on similar plants tend to differ little in vocabulary, with the main need being to identify several commonly used synonyms and closely related terms that reflect differing authorial habits and editorial policies (pod = legume, androecium ~ stamens). The greater effort in adapting code to new projects has been in shifting taxa, as highly specialized plant families will have numerous trait terms that are not of broad application or may have restricted meanings in context.

We also built a custom vocabulary for identifying scientific names within text. This comprises a combination of four sources of known binomials and monomials assembled from these sources: the Integrated Taxonomic Information System (ITIS) (SQLite version at https://www.itis.gov/downloads/index.html), the World Flora Online (WFO) Plant List (https://wfoplantlist.org/plant-list/classifications), Plants of the World Online (http://sftp.kew.org/pub/data-repositories/WCVP/), and further miscellaneous taxa not found in the other sources (available in the "flora" subfolder, file "other_taxa.csv", in the Zenodo repository: https://doi.org/10.5281/zenodo.8336468; LaFrance, 2023).

Parsing strategy

The next major step is to parse each treatment in the resulting text. This uses a wrapper script called "extract_traits.py" to call code from the general Traiter repository (https://github.com/rafelafrance/traiter). These two repositories use a combination of rules and statistical methods to parse traits in the treatments. The basic parsing approach is a hybrid statistical and rule-based approach relying heavily on the spaCy NLP Python library (Honnibal and Montani, 2017).

spaCy is used for NLP because of its prebuilt statistical models and its flexible framework for custom-building parsing rules. First, spaCy's statistical models are used for determining the parts of speech (POS) to which each word belongs. POS is used when building rules for parsing. For instance, parsing a full taxon can be approached by looking for a binomial (species) notation followed by a proper noun or a set of proper nouns (like: John Smythe and Jane Jones) separated by conjunctions to find candidate taxon authorities. We also take advantage of customizability in spaCy; for instance, its default tokenizer (breaking text into meaningful words or word parts) is intended for general-use text such as Wikipedia but tends to fail in tokenizing the formalized and idiosyncratic text of taxonomic treatments. The addition of custom vocabularies greatly improves performance in this respect.

A full outline of the parsing pipeline (summarized and simplified in Figure 2) proceeds as follows: (1) Use a

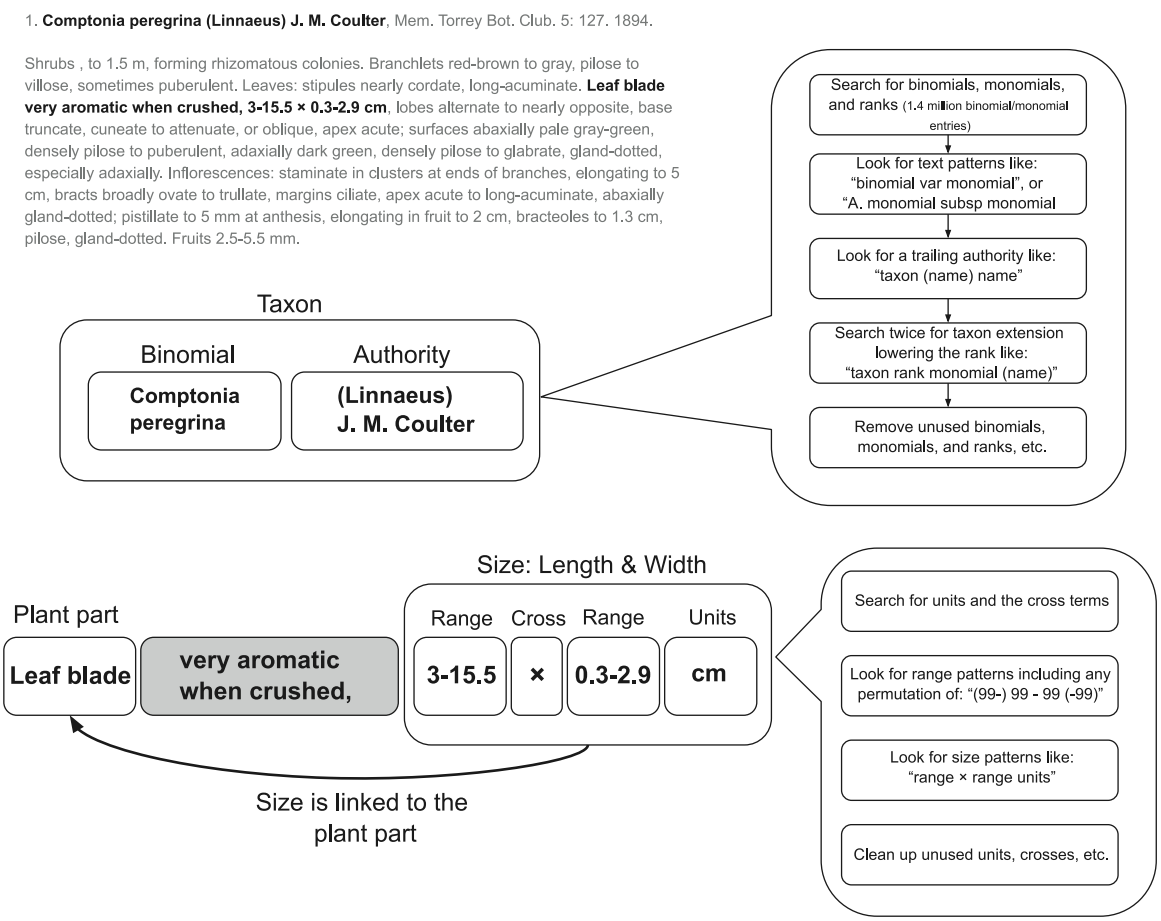


FIGURE 2 Workflow summary of FloraTraiter. Taking the example from the introduction (top left), in bold is the taxon and the first part of the leaf description of *Comptonia peregrina* in *Flora of North America*. The first step is to identify a "taxon anchor" (center), which involves finding binomials or monomials in the text and searching the vicinity for additional material representing authors and subspecific names and ranks. Once this taxon anchor has been identified, plant parts are identified and linked to the taxon (bottom; shown is a measurement example, where the process begins with a similar anchoring process around units and similar symbology) and finally attributes such as measurements are linked to them to yield "partial traits." The given example demonstrates the frequent use of context clues such as implied length and width, as well as the common need to link text material that is non-adjacent.

customized tokenizer to break treatment text into tokens. (2) Allow a spaCy model to identify what POS each token belongs to, as well as other attributes like its lemma (the "normalized" version of a token that reduces it to its root meaning; e.g., "good" would be the lemma of "better"). (3) Parse traits using rules and phrases. Most traits have a predefined vocabulary of words and phrases (noted above) that anchor rules for further parsing. These terms are obtained from organismal experts and other authoritative sources; as described above, the approach used for this study consisted of iterative improvements of drafts that were submitted to expert botanists for comment. The initial product of parsing is termed a "partial trait." (4) Use the anchor phrases in rules that build up full traits from partial traits. This involves finding patterns of words and symbols around previously identified anchors to build up the traits themselves. Sometimes these phrases are the final trait, but more often this step is repeated to build up increasingly larger traits; in such cases, the order of steps affects the outcome and is controlled in repeated iterations to build traits into their final form. (5) Clean up any leftover phrases and partially applied rules so the words/tokens can be used in other traits. For instance, in Figure 2, information about leaf scent (grayed out) splices a statement about leaf length and width, but this discarded information can be used to populate a trait about general leaf properties. (6) Link traits after they are parsed. All built-up traits are linked to the taxon in the treatment title, and other traits like plant or flower parts get linked based on proximity and other trait-specific considerations (i.e., rules about whether links across sentences are allowed). Rules are also applied based on the permissible structure of linkages considering the nature of the trait. For example, a plant part may have several colors but typically only has one size, barring sexual dimorphism; multiple entries are forbidden in certain contexts.

In general, the overall process follows this outline, but special strategies were also used. Some traits can be ambiguous because their meaning depends on context; for instance, "green" can be a last name, a color, or an administrative unit, so context is used extensively in parsing to distinguish among related meanings.

Curation

An important component of achieving high-quality extractions was a series of structured conversations between botanists and programmers as the project developed. This was facilitated by the preparation of interactive visuals that enable non-specialists to understand how the NLP method reads meaning into and pulls information out of text. An example of marked-up HTML output may be seen in the output folder of the Zenodo repository (https://doi.org/10.5281/zenodo.8336468; LaFrance, 2023). Each color corresponds to a trait, with trait labels corresponding to CSV outputs (see below). Common colors between highlighted sections of the treatment and extracted trait data speed up the process of comparing the two, as do different options for color-coding traits. To assist with accessibility, a mouseover action also provides the full trait label. Special data models for reported traits apply to complex numerical data as reported in standard botanical descriptions. For instance, "(1–)2–4(–5) cm" is separately parsed into measurements labeled "min", "low", "high", and "max", as well as a field representing measurement units. HTML outputs were distributed and marked up by participating botanists for correction in the form of prose commentary or marked-up CSV outputs in several iterations.

Post-processing

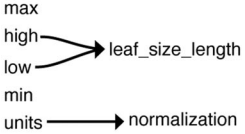
The output of FloraTraiter, even after extensive curation, looks very different from a standard morphological matrix (Figure 3A). Because NLP aims to capture as much text as possible, (1) numerous fields are returned because species descriptions discuss numerous structures. Parts of the internal data structure also lead to this, as "flattened" CSV outputs use separate fields for measurement ranges and multiple values. Additionally, (2) while basic botanical vocabularies are integrated, the philosophy employed here was conservative in the matter of trait synonymy and, generally, measurements and descriptions were kept separate that could conceivably be different (e.g., seed width and height are separate even if only one is mentioned, fruiting and flowering pedicel measurements are not combined, stamen and androecium length were not considered identical). Finally, (3) integrating a large taxonomic and literature scope leads to large, sparse matrices because different groups have some commonalities but numerous differences that reflect specialized morphology. Because a typical downstream user will likely have a biologist's training, we felt it was important to simulate a data structure closer to a morphological matrix.

The approach implemented is a generalization of the morphological matrix harmonization reported in

A Raw outputs

A	B	AUL	AUM	AUN	AUO	AUP
family	taxon	leaf_size.1.length_high	leaf_size.1.length_low	leaf_size.1.length_max	leaf_size.1.length_min	leaf_size.1.length_units
Betulaceae	Alnus					
Betulaceae	Alnus glutinosa	9	3			cm
Betulaceae	Alnus incana					
Betulaceae	Alnus incana subsp. rugosa	11	4			cm
Betulaceae	Alnus incana subsp. tenuifolia	10	4			cm
Betulaceae	Alnus maritima	9	4.5			cm
Betulaceae	Alnus oblongifolia	9	5			cm
Betulaceae	Alnus rhombifolia	9	4			cm
Betulaceae	Alnus rubra	16	6			cm
Betulaceae	Alnus serrulata	14	5			cm
Betulaceae	Alnus viridis	11	3			cm
Betulaceae	Alnus viridis subsp. crispa	6	3.5	10		cm
Betulaceae	Alnus viridis subsp. fruticosa	8	5	10		cm
Betulaceae	Alnus viridis subsp. sinuata	10	4			cm

B Column-smashed outputs



A	B	C	D	E
taxon	leaf_size_length	leaf_size_width	leaf_shape	leaf_apex_shape
Alnus	nan	nan	['ovate']	[]
Alnus glutinosa	6	5.5	['obovate']	['obcordate', 'retuse']
Alnus incana	nan	nan	['ovate']	['acute', 'acuminate', 'obtuse']
Alnus incana subsp. rugosa	7.5	5.5	['ovate']	['acute', 'acuminate', 'obtuse']
Alnus incana subsp. tenuifolia	7	5.25	['ovate']	['acute', 'obtuse']
Alnus maritima	6.75	3.5	['elliptic']	['acute']
Alnus oblongifolia	7	4.5	['valvate']	['acuminate']
Alnus rhombifolia	6.5	3.5	['valvate']	['acute', 'orbicular', 'obtuse']
Alnus rubra	11	7	['ovate']	['acute', 'obtuse']
Alnus serrulata	9.5	5.75	['elliptic']	['orbicular', 'obtuse']
Alnus viridis	7	5.5	['ovate']	['acute', 'orbicular']
Alnus viridis subsp. crispa	4.75	4	['ovate']	['acute', 'obtuse']
Alnus viridis subsp. fruticosa	6.5	4.5	['ovate']	['acute', 'acuminate']
Alnus viridis subsp. sinuata	7	5.5	['ovate']	['acuminate']

FIGURE 3 Examples of trait extractions from FloraTraiter. (A) Raw FloraTraiter output, demonstrating the very long field format due to the detailed method of parsing traits. (B) A result of "column-smashing," where quantitative data were reported as the midpoint excluding extreme measurements, and qualitative data were reported as a list.

Folk et al. (2019), deposited on GitHub at https://github.com/ryanafolk/fagales_traits/. First, a controlled vocabulary for fields is specified by a simple spreadsheet format. This is easily edited by a non-specialist to map field definitions that can be combined. Fields are specified by three separate CSVs to represent (1) categorical data that should be represented as a list, termed "concatenate_terms.csv"; (2) count data that should be summarized in ways appropriate for cardinal numbers such as the mode, termed "range_terms_count.csv"; and (3) quantitative measurements that can be summarized by the mean or other methods appropriate for continuous data, termed "range_terms_quantitative.csv". Additional controlled vocabularies specify (4) equivalency among sex terms (e.g., male and staminate flowers are the same) in "sexes.csv", (5) terms to be excluded in "discard_terms.csv", and (6) measurement unit data in "unit_columns.csv". Second, a synonymy among data in the fields is established in "synonyms.csv". This data file represents some of the most "opinionated" decisions as it includes judgements about term usage: terms that certainly mean the same thing, such as "flabellate = flabelliform"; terms that are likely to mean the same thing, such as "orbicular = round"; fine but unneeded distinctions like "pentagonal = polygonal"; and removal of subjective qualifiers ("sub-", "usually", "quite"). All of these behaviors are completely customizable.

Reading in the field-controlled vocabularies leads to what is referred to here as "column-smashing" (Figure 3B), where fields judged combinable are summarized according to their data type. At this stage, an additional spreadsheet representing taxa extracted by hand can be included, e.g., for data taken from direct measurements or small works such as single species descriptions that do not lend themselves to automated methods. Two types of output are produced: one with all of the data, and one with summarized single data points per cell. By default, the summaries are means for quantitative data, midpoints for count data, and random selections for categorical data. Finally, output can be filtered by missing data proportions, which functions primarily to exclude special descriptive material not shared among species. As well as outputting a raw harmonization of the data, the last function provided by the codes is a distance matrix that can be used to easily perform trait ordinations such as multidimensional scaling (MDS). The distance metric as described by Folk et al. (2019) is a hybrid metric comparable to Gower's distance, partitioning categorical and quantitative data separately but with the ability to specify arbitrary weights to each.

Ground-truthing

To capture Type I (false positive) and Type II (false negative) errors, which measure incorrectly captured and missed trait variation, respectively, we manually scored the extracted trait data using interactive HTML outputs. The two measures were calculated slightly differently: Type I error was calculated with the denominator as the complete count of partial traits because these are reported exhaustively by FloraTraiter and could be individually checked. Type II error was instead calculated with the denominator defined as the count of full traits (i.e., counting whether content was captured for each basic trait topic, while not scoring how verbal details were parsed out) because it was not possible to manually duplicate the tokenization process. As a manageable use case involving diverse plants with many specialized structures, we focused on all available species-level treatments for Betulaceae in *Flora of North America* and *Flora of China* as represented on eFloras.org, comprising 125 species (Figure 4).

Worked example

Fagales treatments as a whole were used, again sourcing from *Flora of North America* and *Flora of China*, to demonstrate one typical downstream use focused on quantifying trait spaces. Using the MDS procedure noted above, which is well-adapted to sparse matrices, we quantified trait spaces across Fagales using all morphological features populated with at least 5% of species. This analysis also demonstrates the reconciliation of automated and user-coded features, which are demonstrated in the empirical GitHub repository (https://github.com/ryanafolk/fagales_traits) and were scored from additional online sources including the *Flora of Australia* (https://profiles.ala.org.au/opus/foa) and *Flora Malesiana* (https://portal.cybertaxonomy.org/flora-malesiana/). This repository demonstrates the original extraction and conversion to controlled fields.

RESULTS

Parsing success

FloraTraiter was implemented on the complete Fagales treatments on eFloras.org as represented in *Flora of North America* (212 taxa successfully extracted) and *Flora of China* (518 taxa extracted). On average, 55 traits per species were captured. The most densely populated traits included leaf shape (83.0%), leaf size (79.6%), leaf margin (61.0%), and leaf surface (54.5%).

Ground-truthing

Type I (false positive) error rates were less than 1% (0.98%; Figure 5A and 5C break this down by source and genus). Many errors tend to be removed when filtering on missing data because they often involve uncommon or very detailed traits. This was because Type I errors were generally of the linkage type; a measurement was captured and parsed correctly but associated with a wrong, spatially adjacent trait. Most often, such association issues happen for less common traits such as vein numbers (accounting for almost all errors; this is reflected in the higher error in *Flora of China*; Figure 5A) that were unparsed as such; during the

A

	Family	Taxon	Text
+	Betulaceae	Corylus cornuta	Shrubs or trees , open-spreading, 4--8(--15) m. Bark light to dark brown, smooth. Branches ascending; twigs glabrous to sparsely pubescent, sometimes with glandular hairs. Winter buds containing inflorescences ovoid, 3--5 × 3--4 mm, acute. Leaves: petiole glabrous to moderately pubescent, with or without glandular hairs. Leaf blade nearly orbiculate to narrowly ovate or ovate-oblong, often nearly angular and slightly lobulate near apex, 4--10 × 3.5--12 cm, thin to leathery, base narrowly cordate to narrowly rounded, margins coarsely and often irregularly doubly serrate, apex obtuse to acute or acuminate; surfaces abaxially glabrous to moderately pubescent, usually pubescent on major veins and in vein axils. Inflorescences: staminate catkins lateral along branchlets on short shoots, usually in clusters of 2--3, 4--6 × 0.5--0.8 cm; peduncles 0.5--10 mm. Nuts in clusters of 2--6, completely concealed; bracts bristly, connate at summit, lengthened into extended tubular beak.
+	Betulaceae	Corylus cornuta subsp. californica	Shrubs or trees , open-spreading, to 8(--15) m; trunks usually several. Bark dark brown to blackish. Branches ascending; twigs sparsely to moderately pubescent, bearing glandular hairs. Winter buds containing inflorescences broadly ovoid, 3--5 × 3--5 mm, apex acute. Leaves: petiole pubescent, often bearing well-developed glandular hairs. Leaf blade nearly orbiculate or broadly elliptic, 4--7 × 3.5--7 cm, leathery, base nearly cordate, margins coarsely doubly serrate, apex obtuse to acute, abaxially moderately pubescent, villous to tomentose on major veins and in vein axils. Inflorescences: staminate catkins usually in clusters of 2--3, 4--6 × 0.5--0.8 cm; peduncles mostly 5--10 mm. Nuts in clusters of 2--4; involucral tubular beak less than 2 times length of nuts, hispid. 2 n = 22.
+	Betulaceae	Corylus cornuta subsp. cornuta	Shrubs , open spreading, to (4--)6 m. Bark light brown, smooth. Branches ascending; twigs glabrous to sparsely pubescent, without glandular hairs. Winter buds containing inflorescences ovoid, 3--5 × 3--4 mm, apex acute. Leaves: petiole glabrous to moderately pubescent, without glandular hairs. Leaf blade ovate to obovate or narrowly elliptic, often nearly angular and lobulate near apex, 5--12 × 3.5--9 cm, base narrowly cordate to narrowly rounded, margins coarsely and often irregularly doubly serrate, apex usually distinctly acuminate; surfaces abaxially glabrous to moderately pubescent, usually pubescent on major veins and in vein axils. Inflorescences: staminate catkins usually in clusters of 2--3, 4.5--6 × 0.5--0.8 cm; peduncles mostly 0.5--2 mm. Nuts in clusters of 2--6; involucral tubular beak long, narrow, 2--3(--4) times length of nuts, densely bristly. 2 n = 22, 28.
+	Betulaceae	Ostrya	Trees , 9--18 m; trunks usually 1, branching mostly deliquescent, trunk and branches terete. Bark of trunk and branches brownish gray to light brown, thin, smooth, breaking and shredding into shaggy vertical strips and scales; lenticels generally inconspicuous. Wood nearly white to light brown, very hard and heavy, texture fine. Branches, branchlets, and twigs conspicuously 2-ranked; young twigs differentiated into long and short shoots. Winter buds sessile, ovoid, somewhat laterally compressed, apex acute; scales many, imbricate, longitudinally striate. Leaves on long and short shoots, 2-ranked. Leaf blade narrowly ovate to ovate, elliptic, or obovate with 10 or more pairs of lateral veins, 2.5--13 × 1.5--6 cm, thin, margins doubly serrate to serrulate; surfaces abaxially glabrous to tomentose. Inflorescences: staminate catkins terminal on branches, mostly in small, racemose clusters, formed previous growing season and exposed

B

	Family	Taxon	Text
−	Betulaceae	Corylus cornuta	Shrubs or trees , open-spreading, 4--8(--15) m. Bark light to dark brown, smooth. Branches ascending; twigs glabrous to sparsely pubescent, sometimes with glandular hairs. Winter buds containing inflorescences ovoid, 3--5 × 3--4 mm, acute. Leaves: petiole glabrous to moderately pubescent, with or without glandular hairs. Leaf blade nearly orbiculate to narrowly ovate or ovate-oblong, often nearly angular and slightly lobulate near apex, 4--10 × 3.5--12 cm, thin to leathery, base narrowly cordate to narrowly rounded, margins coarsely and often irregularly doubly serrate, apex obtuse to acute or acuminate; surfaces abaxially glabrous to moderately pubescent, usually pubescent on major veins and in vein axils. Inflorescences: staminate catkins lateral along branchlets on short shoots, usually in clusters of 2--3, 4--6 × 0.5--0.8 cm; peduncles 0.5--10 mm. Nuts in clusters of 2--6, completely concealed; bracts bristly, connate at summit, lengthened into extended tubular beak.
		Trait Label	Trait Data
		bark_color	color: brown
		bark_size	length_low: 4.0, length_high: 8.0, length_max: 15.0, length_units: m
		bark_surface	surface: smooth
		beak_shape	shape: tubular
		branch_count_staminate	low: 2, high: 3, count_group: cluster, sex: staminate
		habit	habit: ascending
		inflorescence_shape	shape: ovoid shape: acute
		inflorescence_size	length_low: 3.0, length_high: 5.0, length_units: mm, width_low: 3.0, width_high: 4.0, width_units: mm

FIGURE 4 Demonstration of the HTML outputs. Two screenshots are shown for *Corylus* from *Flora of North America*: (A) shows the unexpanded view (the "+" symbol on the HTML will expand the view and show the atomized traits as shown in [B]), with every highlighted portion corresponding to trait content that was extracted; (B) shows an expanded view of one species with the actual scores for each trait. This view allows organismal experts to evaluate the operation of FloraTraiter's natural language processing approach. Example output with an accessible color palette is available in the output folder of the Zenodo repository (https://doi.org/10.5281/zenodo.8336468).

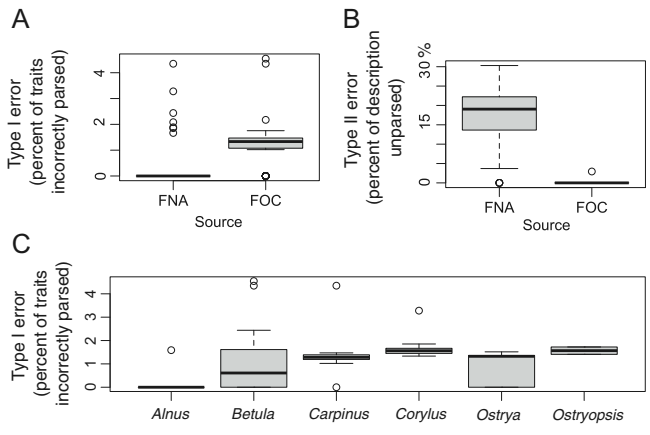


FIGURE 5 False positives (Type I error) and false negatives (Type II error). Type I error is shown vs. the two flora sources (A) and genus (C). Type II error is shown vs. flora source (B). Much of the structuring of errors appears to reflect differing editorial practices. Type II error was higher in the *Flora of North America* (FNA) treatment; the missing portions largely comprised prolix discussions of fine features. Type I error was higher in *Flora of China* (FOC), reflecting extensive discussion of leaf vein characters that were not parsed in detail but could easily be discarded.

iterative process summarized in Figure 2, unused tokens were then likely reused and misassigned. This example is illustrative of the most straightforward way to address Type I error: maximizing the text content that is parsed and explicitly assigned to traits. Rare traits are also where we see the most Type II errors, which was about 4.7% (i.e., the average treatment captured 95.3% of traits). Some of these errors involved phenological traits (e.g., flowering before leaf-out), scent and taste descriptors, and other aspects that we decided not to capture because they are not basic morphology; however, these more frequently comprised detailed treatments of taxon-specific topics like extensive prose on bark structure, winter buds, and elaborations of fruit and seeds that did not contain straightforward scorings. This last point is likely responsible for the higher Type II error in *Flora of North America* (Figure 5B); there is a greater proportion of less structured, informal discussion of traits in this source. Connected to this, invariant "nuisance" traits were sometimes captured because the original description mentioned structures without providing relevant details. These were not counted

as errors because they are uninteresting rather than incorrect and can easily be removed. For instance, lateral veins were captured as present with no detail because they were mentioned in relation to pubescence or other features. The capture was also sometimes incomplete because minor details were not included; however, this was generally not recorded as Type II error because partial traits were not counted. Examples of this include color change over time or extensive prose describing hair or ridge distributions. Excluding these from consideration is reasonable because these are unlikely to be useful in a morphological analysis without further standardization.

Worked example

The MDS analysis (Figure 6A; each dot represents a taxon) demonstrated differences in morphological variability among families, with Fagaceae (the most species-rich family) showing a widespread and Betulaceae showing the smallest morphospace as quantified by between-species morphological distances. The within-family results show differing levels of morphospace occupancy per family, with Fagaceae showing the strongest variability; it alone occupies more than half of MDS axis 1 (Figure 6A), which largely captures variation in vegetative and reproductive size (below) and therefore reflects morphological diversification in this family. Between-family comparisons likely underestimate morphospace divergence due to lack of homology; Casuarinaceae in particular possesses numerous fields not shared with close relatives relating to its specialized morphology. We note in this connection that this challenge is not unique to automated analysis; manual extractions possessed similar patterns of unshared fields between divergent taxa.

Following Folk et al. (2019), "loadings" in the MDS analysis were evaluated by univariate R^2, which quantifies variance explained by each ordinated axis. In the first axis, the most important traits were leaf width ($R^2 = 0.85$), nut width (0.74), leaf length (0.72), and nut length (0.54); hence, overall, axis 1 captures size. The most important traits in the second axis were leaf duration (0.81), branch shape (0.28; this trait mostly captured information related to position such as "ascending"), peduncle shape (0.23), and inflorescence shape (0.20); hence, overall, axis 2 represents shape and other categorical descriptors. Finally, Figure 6B demonstrates an individual trait, showing tree height (the typical quantification of body size in plant ecology) against genera with sufficient subspecific sampling.

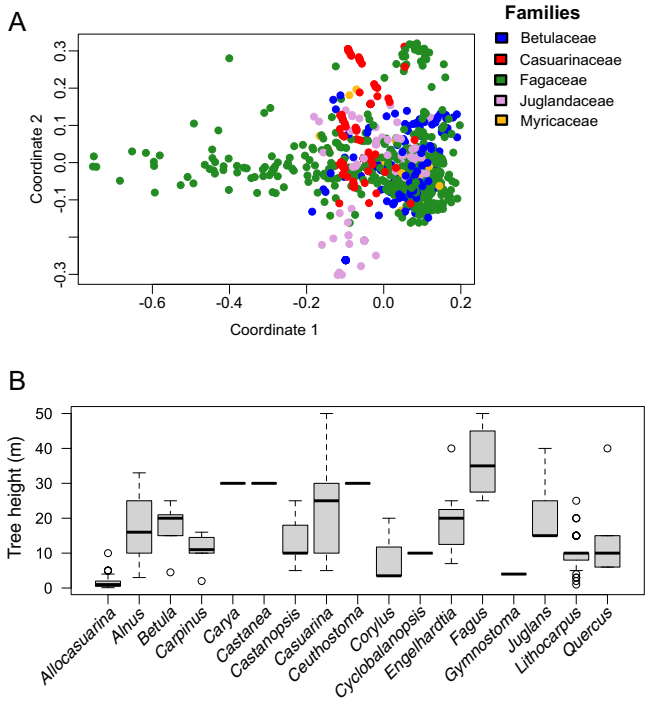


FIGURE 6 Worked example in Fagales. In this proof-of-concept, sampling primarily reflects taxa available in standard online floras; note Nothofagaceae and Ticodendraceae are missing. (A) Multidimensional scaling demonstrating variability of overall morphospaces within and between families. Each dot represents one taxon (mostly species; subspecific and higher-level taxa are also plotted). (B) Tree height plotted by genus.

DISCUSSION

A lack of trait data is a major obstacle in plant science because traits underlie how organisms interact with their environment and other organisms. However, researchers are met with the enormous challenge of collecting many traits from many taxa. The effort needed, for example, at the familial or ordinal level for many speciose clades is immense, especially if starting from scratch. For many of the most basic traits, the problem is not that the data do not exist but that researchers cannot easily assemble them. With resources like the Biodiversity Heritage Library (https://www.biodiversitylibrary.org/) and the Botanicus project (http://www.botanicus.org/), large troves of trait data are available but disseminated in such a way that they are currently available only to specific subfields and applications. As we have argued, unlocking the potential of biodiversity literature will be a key component of overcoming the lack of trait data, when used in a complementary role with other approaches such as generating new in-situ field-based measurements. However, this heritage data is vast in scope and can be dissimilar in content, which highlights the need for automated and semi-automated approaches, especially as these data are not limited to a single application or purpose, but can be re-used by others in new applications as befits their value. Such FAIR data re-use approaches (Wilkinson et al., 2016) have not been particularly well-facilitated by traditional scientific literature dissemination modalities.

Here, we developed and demonstrated the application of FloraTraiter, a fully open-source NLP approach to extracting computable trait data from taxon descriptions in floras, monographs, and similar biodiversity literature. FloraTraiter is close to full automation, but contains human input at

several points that adds customizability as well as quality control. Some human roles are specific to particular projects, such as overcoming challenges in digitizing treatments and in controlling and interpreting output structures. Most centrally, however, expert botanist perspectives are built into the NLP model; this is particularly important for specialized contents where applying previously developed approaches "out of the box" may result in limited parsing (Endara et al., 2018). We identify a number of challenges related to the structure (or lack thereof) of the data that will inform future efforts (see also Appendix S1), but also demonstrate a level of accuracy already sufficient for re-use in diverse applications and in a form that will facilitate further curation.

FloraTraiter fits into recent efforts unlocking similar literature data in other organisms, such as lepidopterans (Shirey et al., 2022) and vertebrates (Guralnick et al., 2016). The present effort differs from Shirey et al. (2022) in that much of that work relied on human effort to assemble and parse text blocks, with automation for only the simplest continuous traits (e.g., wing length). Unlike Guralnick et al. (2016), FloraTraiter interrogates much more complicated written text blocks and assembles a much more complicated set of both continuous and categorical traits. Our work not only demonstrates the power of semi-automation to scale up trait assembly, but also illustrates the critical importance of a more explicit model for collaboration between botanists and programmers to structure improving language processing models. Similar efforts have previously been recorded in plants (e.g., Endara et al., 2018 report an NLP approach), but here we have moved beyond proof-of-concept in a large-scale study using a data set comprising hundreds of taxa ready for reuse.

Challenges encountered

Challenges and recommendations in the effort overall are summarized in Appendix S1. In this section, we will focus on empirical properties of the output. Type I error rates are low (<1%) and similar to those reported in previous NLP efforts (Endara et al., 2018). The specific Type I errors found were attributed to the completeness of parsing, with information involving unparsed traits most prone to incorrect trait linkage. This suggests a straightforward strategy in addressing Type I error focused on maximizing the vocabulary that would enable tokenization in the first steps of FloraTraiter. Type II error was also fairly low (<5%); this is more difficult to control as it was found to be highly dependent on authorial style among sources, as standard taxonomic styling lends itself to NLP but more prolix approaches to the discussion of traits tend not to be parsed. However, excluded trait information tends to be highly detailed and may not be specifically of interest for many projects. In Betulaceae, this often involved relative flower and leaf phenologies (not included by decision) and extensive bark descriptions (which could be challenging to score even for a human observer).

As noted by Endara et al. (2018), matrices created by automated trait extraction tend to be sparse, with basic leaf and other traits well filled-in, but other traits less so. This is a feature rather than a bug; it reflects the higher thoroughness of an automated approach as opposed to the more selective strategy a human observer would likely take. Trait presence is primarily a function of both morphological specialization and editorial differences among sources. At a large taxonomic scope, fields must be invoked to cover all extracted traits, but within subtaxa taxonomists generally will focus only on those traits that differ between species. Some traits are also difficult to homologize because features are not shared across taxa in a straightforward way (for instance, when comparing details of berries to samaras, or when comparing Casuarinaceae overall to other families). In the Betulaceae application used for error analysis, there were fewer problems because Betulaceae trees are fairly similar; however, their diverse inflorescence structure requires enforcing controlled fields and synonymies when comparing individuals across genera.

A further challenge concerns the nature of expert literature rather than methodological approaches. Many taxa and geographic regions are understudied due to a shortage of taxonomists (Bebber et al., 2014), so many taxonomic works will be out of date to varying degrees. For relatively similar species, applying synonymy tools such as Plants of the World Online (http://sftp.kew.org/pub/data-repositories/WCVP/) can identify and often reduce the scope of the problem, but the changes in plant taxonomy have been more extreme at the levels of genus, family, and above, rendering some familiar taxa unrecognizable (e.g., Scrophulariaceae; Olmstead et al., 2001). Higher-level taxon treatments also differ in their scope and selection of morphological traits from the more basic unit of species. While FloraTraiter reports data at all ranks identified in its inputs, higher-level trait reports are certainly suspect in older treatments that antedate molecular phylogenetic work, and trait reports for such taxa may best be generated from aggregating species- or subspecific-level data. For instance, genus-level length measurements could be generated from the range of variation in all species successfully captured and reconciled to currently recognized generic boundaries. In summary, beyond the more difficult problem of promoting the basic taxonomic work that would best address outdated taxonomic treatments (Stuessy, 1993), the careful selection of recent authoritative sources and a focus on data from species and varieties or subspecies represents a compromise that best summarizes current expert knowledge.

CONCLUSIONS

Among the extracted traits that are broadly scored across species are body size, measurements and shape descriptors of leaves and flowering and fruiting structures, scorings of plant sex, and other attributes that have broad use potential.

For instance, these traits capture aspects of plant and leaf size that have recently seen global-scale studies in a phylogenetic (Testo and Sundue, 2018) or spatial framework (Baird et al., 2021). Aside from investigations of targeted questions, traits could be used to capture regional or site-level morphospaces, e.g., via calculation of convex hulls (Cornwell et al., 2006) or other methods; ordinations of morphospaces could also be used in a phylogenetic context to investigate trait evolution (Folk et al., 2019). These examples underline the value of making biodiversity literature more available and computable (Thessen et al., 2012; Folk and Siniscalchi, 2021) to broad audiences beyond the traditional readership of such literature, including scientists who will be able to identify important, unusual, and creative applications. Most centrally, FloraTraiter demonstrates the major promise of combining expert perspectives with guided automated approaches, as well as the promise of tools designed to facilitate curation for non-specialists. Future challenges will include data curation and reconciliation challenges, and assuring that outputs are broadly usable for the long term and conform to community definitions. We also foresee that large language models and model-based parsing may rapidly improve and provide an alternative to the rule-based parsing shown here; still, those models are likely to require the same set of lexicons that are specific to botany and therefore to benefit from collaboration with domain experts. Here, we show a pathway to making data and methods available that can serve as an important first step toward broadly available plant trait data for the community.

AUTHOR CONTRIBUTIONS

R.A.F. and R.P.G. acquired the funding and conceived the research; R.T.L. and R.A.F. designed the code and analyzed the data with input from R.P.G. R.A.F. and R.T.L. wrote the manuscript; all authors edited the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

We first thank the collective efforts of taxonomists over hundreds of years; the automated approach described here focuses on recent works, but this still substantial legacy depends on the expertise of botanical experts. For input on our work, we thank Leonardo Borges (Universidade Federal de São Carlos), Yago Barros (Universidade Federal de São Carlos), Carolina Siniscalchi (Mississippi State University), and Joshua Doby (University of Florida). This work was supported by the National Science Foundation (NSF DEB-1916632).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available at https://doi.org/10.5281/zenodo.8336468 and https://doi.org/10.5281/zenodo.8349315.

DATA AVAILABILITY STATEMENT

The codebase for FloraTraiter is disseminated on GitHub at https://github.com/rafelafrance/FloraTraiter, with a publication release of this repository available at https://doi.org/10.5281/zenodo.8336468 (LaFrance, 2023). A worked example processing Fagales content from eFloras.org is deposited at https://github.com/ryanafolk/fagales_traits, with a publication release of this repository available at https://doi.org/10.5281/zenodo.8349315 (Folk, 2023).

ORCID

Ryan A. Folk http://orcid.org/0000-0002-5333-9273
Robert P. Guralnick http://orcid.org/0000-0001-6682-1504

REFERENCES

Baird, A. S., S. H. Taylor, J. Pasquet-Kok, C. Vuong, Y. Zhang, T. Watcharamongkol, C. Scoffoni, et al. 2021. Developmental and biophysical determinants of grass leaf size worldwide. *Nature* 592: 242–247.

Bebber, D. P., J. R. I. Wood, C. Barker, and R. W. Scotland. 2014. Author inflation masks global capacity for species discovery in flowering plants. *New Phytologist* 201: 700–706.

Bornstein, A. J. 1997. *Comptonia peregrina*. *In* Flora of North America Editorial Committee [eds.], Flora of North America North of Mexico, vol. 3. Magnoliophyta: Magnoliidae and Hamamelidae, 356–357. Oxford University Press, New York, New York, USA.

Cornwell, W. K., L. D. W. Schwilk, and D. D. Ackerly. 2006. A trait-based test for habitat filtering: convex hull volume. *Ecology* 87: 1465–1471.

Cornwell, W. K., W. D. Pearse, R. L. Dalrymple, and A. E. Zanne. 2019. What we (don't) know about global plant diversity. *Ecography* 42: 1819–1831.

Cui, H., D. Xu, S. S. Chong, M. Ramirez, T. Rodenhausen, J. A. Macklin, B. Ludäscher, et al. 2016. Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. *BMC Bioinformatics* 17: 471.

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. S. Whitfeld, T. G. Seidler, et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217: 939–955.

Endara, L., H. Cui, and J. G. Burleigh. 2018. Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Applications in Plant Sciences* 6: e1035.

Folk, R. A. 2023. ryanafolk/fagales_traits: New release (v1.0). Available at Zenodo repository: https://doi.org/10.5281/zenodo.8349315 [posted 15 September 2023; accessed 23 December 2023].

Folk, R. A., and C. M. Siniscalchi. 2021. Biodiversity at the global scale: The synthesis continues. *American Journal of Botany* 108: 912–924.

Folk, R. A., M. Sun, P. S. Soltis, S. A. Smith, D. E. Soltis, and R. P. Guralnick. 2018. Challenges of comprehensive taxon sampling in comparative biology: Wrestling with rosids. *American Journal of Botany* 105: 433–445.

Folk, R. A., R. L. Stubbs, M. E. Mort, N. Cellinese, J. M. Allen, P. S. Soltis, D. E. Soltis, and R. P. Guralnick. 2019. Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proceedings of the National Academy of Sciences, USA* 116: 10874–10882.

Freudenstein, J. V., M. B. Broe, R. A. Folk, and B. T. Sinn. 2016. Biodiversity and the species concept—Lineages are not enough. *Systematic Biology* 66: 644–656.

Guralnick, R. P., P. F. Zermoglio, J. Wieczorek, R. LaFrance, D. Bloom, and L. Russell. 2016. The importance of digitized biocollections as a source of trait data and a new VertNet resource. *Database* 2016: baw158. https://doi.org/10.1093/database/baw158

Honnibal, M., and I. Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and

incremental parsing. Website https://spacy.io/ [accessed 15 December 2023].

Hortal, J., F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 46: 523–549.

LaFrance, R. 2023. rafelafrance/FloraTraiter: Color blind friendly CSS (v1.2.0). Available at Zenodo repository: https://doi.org/10.5281/zenodo.8336468 [posted 12 September 2023; accessed 23 December 2023].

Meyer, C., H. Kreft, R. Guralnick, and W. Jetz. 2015. Global priorities for an effective information basis of biodiversity distributions. *Nature Communications* 6: 8221.

Olmstead, R. G., C. W. Depamphilis, A. D. Wolfe, N. D. Young, W. J. Elisons, and P. A. Reeves. 2001. Disintegration of the Scrophulariaceae. *American Journal of Botany* 88: 348–361.

Pakeman, R. J. 2014. Functional trait metrics are sensitive to the completeness of the species' trait data? *Methods in Ecology and Evolution* 5: 9–15.

Pakeman, R. J., and H. M. Quested. 2007. Sampling plant functional traits: What proportion of the species need to be measured? *Applied Vegetation Science* 10: 91–96.

Penev, L., M. Dimitrova, V. Senderov, G. Zhelezov, T. Georgiev, P. Stoev, and K. Simov. 2019. OpenBiodiv: A knowledge graph for literature-extracted linked open data in biodiversity science. *Publications* 7: 38.

Rinaldo, C. 2009. The Biodiversity Heritage Library: Exposing the taxonomic literature. *Journal of Agricultural & Food Information* 10: 259–265.

Sandel, B., A. G. Gutiérrez, P. B. Reich, F. Schrodt, J. Dickie, and J. Kattge. 2015. Estimating the missing species bias in plant trait measurements. *Journal of Vegetation Science* 26: 828–838.

Shirey, V., E. Larsen, A. Doherty, C. A. Kim, F. T. Al-Sulaiman, J. D. Hinolan, M. G. A. Itliong, et al. 2022. LepTraits 1.0 A globally comprehensive dataset of butterfly traits. *Scientific Data* 9: 382.

Stuessy, T. F. 1993. The role of creative monography in the biodiversity crisis. *Taxon* 42: 313–321.

Testo, W. L., and M. A. Sundue. 2018. Are rates of species diversification and body size evolution coupled in the ferns? *American Journal of Botany* 105: 525–535.

Thessen, A. E., H. Cui, and D. Mozzherin. 2012. Applications of natural language processing in biodiversity science. *Advances in Bioinformatics* 2012: 391574.

Violle, C., P. B. Reich, S. W. Pacala, B. J. Enquist, and J. Kattge. 2014. The emergence and promise of functional biogeography. *Proceedings of the National Academy of Sciences, USA* 111: 13690–13696.

Wilkinson, M. D., M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.

SUPPORTING INFORMATION

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

Appendix S1. Guide to generalizing natural language processing and overcoming challenges.

How to cite this article: Folk, R. A., R. P. Guralnick, and R. T. LaFrance. 2024. FloraTraiter: Automated parsing of traits from descriptive biodiversity literature. *Applications in Plant Sciences* 12(1):e11563. https://doi.org/10.1002/aps3.11563