A Natural Language Processing Pipeline to Extract Phenotypic Data from Formal Taxonomic Descriptions with a Focus on Flagellate Plants

Lorena Endara, J. Gordon Burleigh
Department of Biology
University of Florida
Gainesville, FL, USA

Laurel Cooper, Pankaj Jaiswal
Department of Botany and Plant Pathology
Oregon State University
Corvallis, OR, USA

Marie-Angélique Laporte
Bioversity International
Montpellier, France

Hong Cui,
School of Information
University of Arizona
Tucson, AZ, USA

Abstract— Assembling large-scale phenotypic datasets for evolutionary and biodiversity studies of plants can be extremely difficult and time consuming. New semi-automated Natural Language Processing (NLP) pipelines can extract phenotypic data from taxonomic descriptions, and their performance can be enhanced by incorporating information from ontologies, like the Plant Ontology (PO) and the Plant Trait Ontology (TO). These ontologies are powerful tools for comparing phenotypes across taxa for large-scale evolutionary and ecological analyses, but they are largely focused on terms associated with flowering plants. We describe a bottom-up approach to identify terms from flagellate plants (including bryophytes, lycophytes, ferns, and gymnosperms) that can be added to existing plant ontologies. We first parsed a large corpus of electronic taxonomic descriptions using the Explorer of Taxon Concepts tool (http://taxonconceptexplorer.org/) and identified flagellate plant specific terms that were missing from the existing ontologies. We extracted new structure and trait terms, and we are currently incorporating the missing structure terms to the PO and modifying the definitions of existing terms to expand their coverage to flagellate plants. We will incorporate trait terms to the TO in the near future.

Keywords—Natural Language Processing; Plant Ontology; Plant Trait Ontology; taxonomic descriptions; flagellate plants; phenotypic traits; matrices; phylogeny

I. Introduction

Assembling phenotypic datasets is a major bottleneck for many studies in evolutionary biology and biodiversity science [1]. New computer-mediated methods facilitate and expedite the assembly of plant trait datasets from digital images and the natural history literature [1–3]. For example, Natural Language Processing (NLP) approaches can be used to extract phenotypic data from formal taxonomic descriptions [4, 5]. The phenotypic characters can be organized quickly and inexpensively into character x taxon matrices that can be used for tasks such as phylogenetic inference, ancestral state reconstruction, or key building.

Ontologies, structured vocabularies of standardized terms and the logical relationships between those terms [6], can enhance NLP approaches for assembling phenotypic datasets by increasing the precision of the data extracted and consequently the number of usable characters. For example, into parsing analyses, ontologies can establish complex relationships among plant parts. For example, 'apicula' *is_part* of 'apex', and 'apex' *is part* of 'leaf'. In this example, this representation of knowledge enables the system to extract the qualifiers of the apicula (e.g., vestigial/prominent, length of the apicula), relate them to the leaf, and distinguish this information from apicula present in other structures (e.g., petals)

There has been much recent work to develop ontologies and controlled vocabularies for botanical terms, such as the Plant Ontology (PO) and the Plant Trait Ontology (TO) [6–10]. However, these efforts have largely focused on terms associated with flowering plants. There is a need to enrich plant ontologies with terms from 'flagellate plants', land plants including bryophytes, lycophytes, ferns, and gymnosperms that mostly have flagellated sperm and lack flowers.

Many of the terms that are used to describe plant structures and traits in flagellate plants have not been formalized in controlled vocabularies and ontologies. Additionally, other terms included in the ontologies have definitions that do not encompass the usage found in descriptions of flagellate plants. The lack of terms in existing plant ontologies for flagellate plants limits the effectiveness of NLP approaches to generate comparative phenotypic datasets.

In this study, we demonstrate a bottom-up approach to extract structures and traits that can be used to enrich the available ontologies. We used a semi-automatic Natural Language Processing (NLP) pipeline to extract terms from plant taxonomic descriptions. We then evaluated whether these terms were represented in the PO and TO, and if they should be added to the ontologies, or if the definitions of existing terms should be expanded to accommodate all the uses of the term. This bottom-up approach can identify candidate terms for plant ontologies and capture the diversity of semantic usage of the terms. By considering this variation in the use of a term, we can develop ontologies with broader phylogenetic coverage and thus improve the efficiency of assembling character matrices across plants.

II. Natural Language Processing Pipeline and Identification of Candidate Terms to be Included in Ontologies.

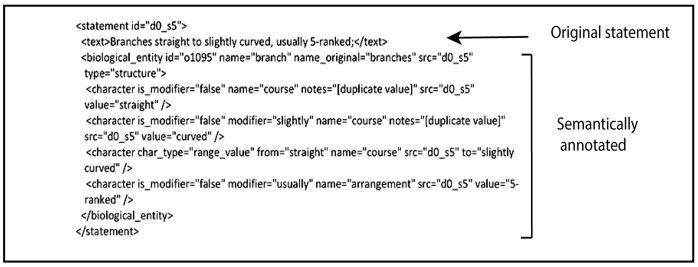
A. Extraction of terms from taxonomic descriptions

We gathered 3978 taxonomic descriptions of flagellate plant taxa from electronic versions of seven floras and monographic treatments (Table 1). These descriptions were written in the telegraphic syntax (i.e., abbreviated English language; Fig. 1). Only the text in the body of descriptions was used, and we removed the parenthetical remarks and extended descriptions, which often violate the rules of telegraphic syntax.

We input the formatted descriptions into the Explorer of Taxon Concepts pipeline (ETC) [4], an online application that uses an unsupervised machine learning model to analyze the formulaic sentences used in descriptions. These sentences consist of a structure followed by a string of qualifiers separated by commas (i.e., Structure (noun), qualifier1, qualifier2, …qualifier n;). The ETC pipeline is composed of five tools. However, to extract terms from the descriptions, we used only the 'Text Capture Tool', which transforms the input text into XML format, identifies sentences and the terms within them (i.e., parsing), and semantically annotates the components of each sentence (Fig.1). This step of the analysis is facilitated by built-in reference glossaries specific for each group of organisms. To parse the flagellate plant dataset, we used the "Plant Glossary" [8].

During the initial phases of the parsing analysis, the Text Capture Tool recognizes terms based on the reference glossaries and places them into discrete, predefined categories. It also presents the user with unrecognized terms, along with the corresponding context sentences, to facilitate the evaluation of terms (Fig. 2). The context sentences enable the user to see all the ways in which a term has been used throughout the descriptions, and the user can manually categorize any terms which were not automatically assigned a category. Using the context sentences, we categorized terms that were unrecognized by the system, and we also verified the categorizations performed by the software.

Figure 1. Example of sentence of a taxonomic description written in telegraphic syntax that has been semantically annotated by ETC.



We downloaded all the categorized terms extracted by the system for each of the seven datasets (Table 1) using the 'File Download' function of the Review step of the Text Capture tool (Fig. 2). The files downloaded in this step were in comma separated values (csv) format and contained the terms extracted by the ETC and their corresponding categories. For example, the term *blue* would be associated with the *coloration category*, whereas *leaf* would be with assigned to the *structure category*.

TABLE 1: SUMMARY OF THE TAXONOMIC DESCRIPTIONS PARSED USING NLP PIPELINE AND STRUCTURAL TERMS EXTRACTED FOR EACH DATASET.

Datasets (sources)	Number of descriptions	Number of 'structure' terms
Cycads [11]	312	170
Ferns of Australia [12]	463	334
Ferns of China [13]	422	159
Ferns of Mexico [14]	950	51
Ferns of North America [15]	649	23
Gymnosperms, exc. Cycads [16]	646	101
Moss Flora of China- Vol. I, II [17]	536	174
Total Number:	3978	1012 (575-unique terms)

Although we extracted terms describing both structures and traits, we are first focusing on evaluating and adding structure terms to the Plant Ontology only. We extracted 1012 plant structure terms from across flagellate plants (Table 1), 575 of which were unique. Because structure terms are defined differently in ETC and the PO, our first effort was to distinguish structure terms that can be added to the PO. The nature of the difference is that structure terms extracted by ETC include external and internal anatomical entities, as well

as terms that refer to parts, spaces, lines, scars, constrictions, and derived products. In contrast, PO structures are defined more strictly as parts of a plant (i.e. anatomical structure). We evaluated the terms extracted (575 unique terms) and separated terms that refer to anatomical structures (494) from non-specific nouns like aperture, border, or center (81 terms).

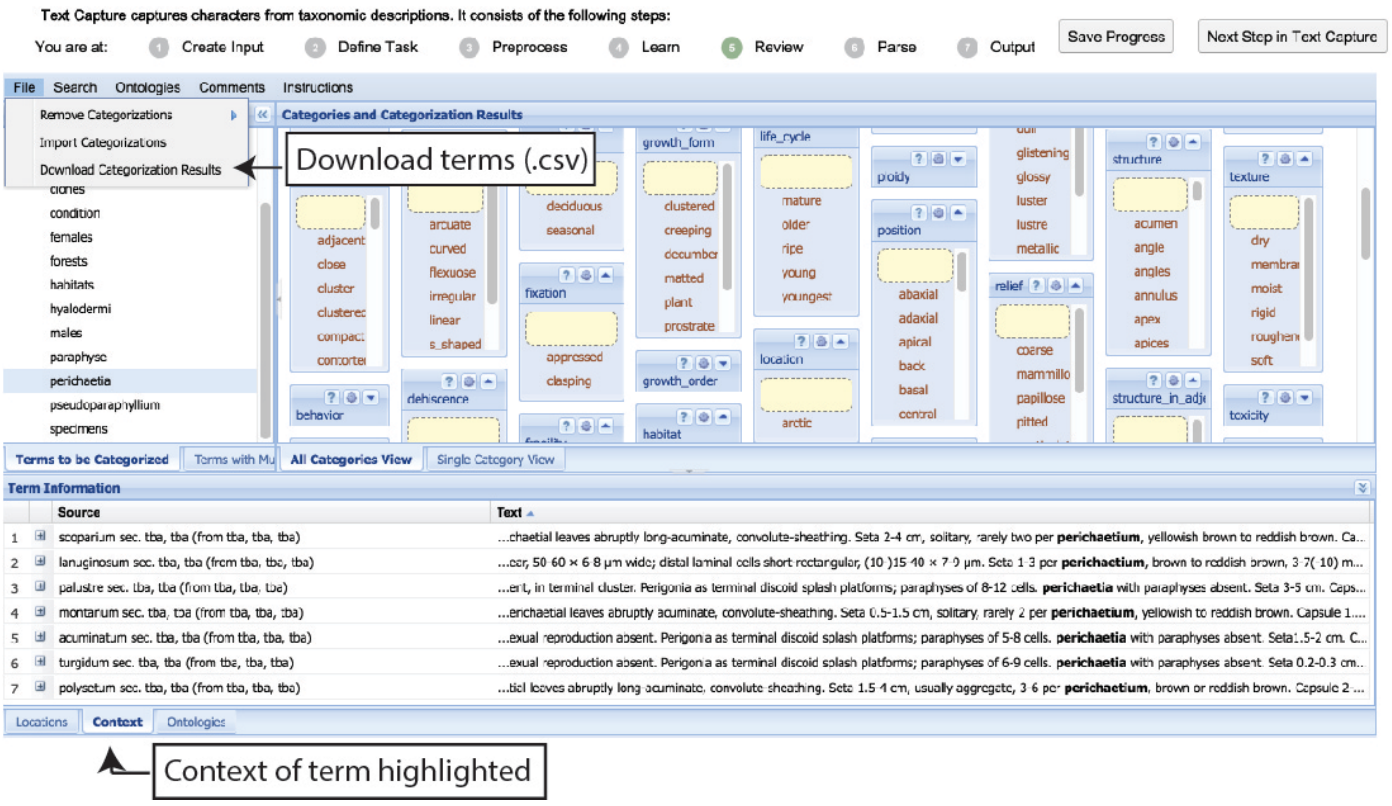


Figure 2. Interface of the review phase of ETC Text Capture tool. Users can download the terms identified by the system and evaluate terms with their context sentences.

B. Mapping the terms to exisitng ontologies

We mapped the extracted candidate terms to the existing Plant and Trait Ontologies based on string similarity and ontology design patterns using an in-house script (https://github.com/Planteome/common-files-for-ref-ontologies/tree/master/scripts). A total of 222 structure terms were mapped to the Plant Ontology using this method, but they still required a curator to review. For many of the extracted candidate terms that were not mapped automatically to PO terms, we used the context sentences (Fig. 2), and manually matched the term to an ontology term based on the human-readable definition. The many terms that were not mapped (272 terms) are good candidates terms that can be submitted to the existing ontology, either as new terms or as synonyms of existing terms. The context sentences can be helpful for building the definitions.

C. Adding terms to the ontologies- using the GitHub issue tracker

Once we identified a term for addition to the PO, we opened an issue on the Plant Ontology GitHub repository (https://github.com/Planteome/plant-ontology/issues).

The proposed definitions for the terms were determined by flagellate plant experts, working with the ontology curators. For example, we recently added the term gametophore coma to the PO:

Issue tracker: https://github.com/Planteome/plant-ontology/issues/682

```
gametophore coma (PO:0028005): A
collective plant organ structure
(PO:0025007) which is a cluster of
gametophore branches (PO:0030021) or non-
vascular leaves (PO:0025075) at the top
of the gametophore axis (PO:0030020),
forming a tuft.
```

III. Discussion and future directions

Our bottom-up approach of using ETC to parse flagellate plant descriptions has produced a wealth of candidate terms for inclusion in existing plant ontologies. These efforts have the potential to greatly enhance the phylogenetic breadth of terminology in plant ontologies. We have parsed descriptions from all the genera and species of conifers and cycads, most of the genera and some species of ferns, and some of the gnetales. Although we have parsed descriptions of the Moss Flora of China (Table 1), our sampling of the diversity of bryophytes (i.e., mosses, liverworts and hornworts) and lycophytes is still low. We are focusing our efforts to gather descriptions of the main lineages of bryophytes and lycophytes. Other future efforts will include adding the additional new terms to the Plant Ontology and extending this effort to incorporate terms to the TO. From the corpus of descriptions detailed in Table 1, we have currently extracted 2162 trait terms from which only 503 are represented in Phenotypic Quality Ontology (PATO).

Acknowledgment

Nathalie Nagalingum (California Academy of Sciences) and Eric Schuettpeltz (Smithsonian Institution) provided taxonomic descriptions and guided the sampling strategy, Annika Smith (Florida Museum of Natural History) contributed with term definitions. This work was supported by the National Science Foundation NSF-Building a Comprehensive Evolutionary History of Flagellate Plants (DEB-1541506), and NSF-Exploring Taxon Concepts (ETC) through Analyzing Fine-Grained Semantic Markup of Descriptive Literature (DBI-1147266). Funding for the Planteome project is provided by the National Science Foundation award IOS-1340112.

References

1. Burleigh JG, Alphonse K, Alverson AJ, et al (2013) Next-generation phenomics for the Tree of Life. PLoS Curr. doi: 10.1371/currents.tol.085c713acafc8711b2ff7010a4b03733

2. Deans AR, Lewis SE, Huala E, et al (2015) Finding Our Way through Phenotypes. PLoS Biol 13:e1002033 . doi: 10.1371/journal.pbio.1002033

3. Henning T, Plitzner P, Güntsch A, et al (2018) Building compatible and dynamic character matrices – Current and future use of specimen-based character data. Botany Letters. doi: 10.1080/23818107.2018.1452791

4. Cui H, Xu D, Chong SS, et al (2016) Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics 17:471 . doi: 10.1186/s12859-016-1352-7

5. Endara L, Cui H, Burleigh JG (2018) Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. Appl Plant Sci 6: . doi: 10.1002/aps3.1035

6. Walls RL, Athreya B, Cooper L, et al (2012) Ontologies as integrative tools for plant science. Am J Bot 99:1263–1275 . doi: 10.3732/ajb.1200222

7. Cooper L, Walls RL, Elser J, et al (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. Plant and Cell Physiology 54:e1–e1 . doi: 10.1093/pcp/pcs163

8. Endara L, Cole HA, Burleigh JG, et al (2017) Building the "Plant Glossary"- A controlled botanical vocabulary using terms extracted from the Floras of North America and China. Taxon 66:953–966 . doi: info:doi/10.12705/664.9

9. Garnier E, Stahl U, Laporte M-A, et al (2017) Towards a thesaurus of plant characteristics: an ecological contribution. Journal of Ecology 105:298–309 . doi: 10.1111/1365-2745.12698

10. Cooper L, Meier A, Laporte M-A, et al (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Research 46:D1168–D1180 . doi: 10.1093/nar/gkx1152

11. Hill K (1998) The Cycad Pages. In: The Cycad Pages. http://plantnet.rbgsyd.nsw.gov.au/PlantNet/cycad/. Accessed 8 Jun 2018

12. ABRS Flora of Australia Online. In: Flora of Australia Online. http://www.environment.gov.au/science/abrs/publications/flora-of-australia. Accessed 13 Jun 2018

13. Flora of China Editorial Committee (1994) Flora of China. Science Press and Missouri Botanical Garden Press, Beijing, St. Louis

14. Mickel JT, Smith AR (2004) The Pteridophytes of Mexico. Part I (Descriptions and Maps). The New York Botanical Garden Press, Bronx, NY

15. Flora of North America Editorial Committee (1993) Flora of North America North of Mexico. New York and Oxford

16. Earle CJ (2017) The Gymnosperm Database. In: The Gymnosperm Database. http://conifers.org/. Accessed 8 Jun 2018

17. efloras (2008) Moss Flora of China @ efloras.org. http://www.efloras.org/flora_page.aspx?flora_id=4. Accessed 11 Jun 2018