Conference Abstract

# Even Simple Habitat Ontologies are Hard to Use

Jocelyn Pender‡, Joel Sachs‡, Beatriz Lujan-Toro‡, James Macklin‡

‡ Agriculture and Agri-Food Canada, Ottawa, Canada

## Abstract

An essential component in describing, delimiting, and understanding the evolutionary context of a taxon is characterizing the habitats in which the taxon is found. We report on a simple habitat ontology that we have developed, and on our ongoing experience using volunteers to annotate legacy habitat descriptions with terms from the ontology.

Our botanical informatics group is building the Canadian Flora Commons, a knowledge platform to aggregate, integrate and facilitate collaboration on information about Canadian plants. Species pages in the Commons are seeded with structured data extracted from authoritative sources such as the Flora of North America (FNA), Flora of British Columbia, etc. In previous TDWG talks (e.g., Sachs et al. 2019), we described our workflow for extracting and structuring morphological data. To understand why habitat descriptions are different and pose a unique set of challenges, consider the following (from *Plectocephalus rothrockii* in FNA): "Damp soil near streams, roadsides, open pine-oak woodlands and forests". Here, the single field "habitat" is used to capture environmental conditions, canopy coverage, and taxonomic associations. We also find it often used for geology, climate, etc. Information in the habitat field is often detailed, but it is presented in free text with little editorial guidance, and comparison between treatments within a given flora and among floras is challenging.

Environment ontologies that could aid in the standardization of habitat descriptors exist, notably ENVO (ENVironment Ontology; Buttigieg et al. 2016). However, ENVO's goals have been primarily focused on describing the biomes, environmental features and environmental materials of molecular datasets, resulting in an ontology that thus far does

not serve our needs. To our knowledge, no habitat ontology exists that supports species-level use cases (but see the habitat classification scheme developed by the IUCN).

To address this, we developed a small and simple habitat ontology by examining over 3000 habitat descriptions across multiple families, and asked "what is the author trying to tell us?". In our taxonomic treatment authoring tool, being developed as part of another project, we will use this ontology to replace or supplement the single "habitat" field with multiple habitat dimensions ("soil type", "canopy coverage", etc.), some with controlled vocabularies (e.g. {open, closed, partial} for canopy coverage). We are also "translating" legacy habitat descriptions into instance data for the ontology. This is a time-consuming process and has the potential to be dependent on interpretations made by the translator. The crowdsourcing experiment described below is aimed at addressing the first issue and quantifying the second.

With our centre's support, we recruited a team of volunteers (6–8 at any given time), and taught them how to annotate habitat descriptions with WebProtegé (Horridge et al. 2014). We divided volunteers into two groups, with each group working with the same dataset, so that we could compare results.
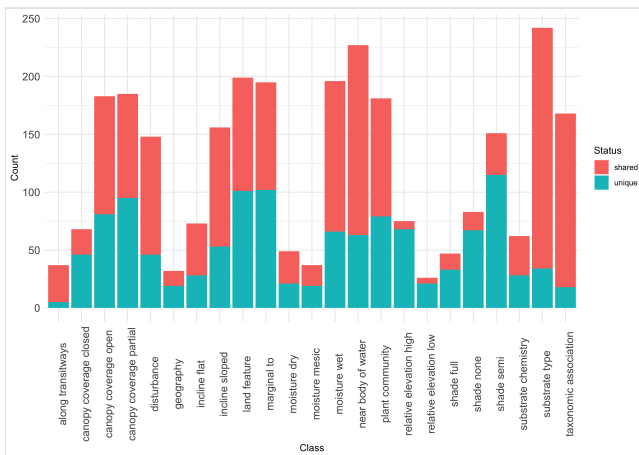


Figure 1.

The number of unique and shared annotations made by our volunteer habitat ontology group by class. Unique annotations are classes added to a habitat description by only one group (e.g., only one group added "canopy coverage = closed" to the phrase "pine forest").

While a purpose-built habitat ontology offers advantages over existing environment ontologies and a consensus was reached on habitat class definitions (e.g., moisture, elevation, canopy coverage), we discovered that it is difficult to achieve consensus on the *application* of habitat classes. Between the two groups, shared annotations represented 57% of the total annotations added to terms and phrases and unique annotations represented 43%. This aligns with previous efforts to build a controlled vocabulary for FNA treatments, where differences between term categorizations represented 49% of the effort

(Endara et al. (2017)). Amongst classes in our ontology, unique annotations varied between 11% and 76% (see Fig. 1).

Our talk will describe our findings, discuss the subjectivity of habitat classes and other difficulties we've encountered while building our ontology, and demonstrate the power of a habitat-driven search interface. This interface will live alongside parsed morphological descriptions (see dev.floranorthamerica.org). We invite collaboration towards increasing the robustness and applicability of the ontology.

## Keywords

ontology, crowdsourcing, environment, botany, floras

## Presenting author

Jocelyn Pender

## Presented at

TDWG 2020

## References

- Buttigieg PL, Pafilis E, Lewis S, Schildhauer M, Walls R, Mungall C (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics 7 (1). https://doi.org/10.1186/s13326-016-0097-6
- Endara L, Cole H, Burleigh JG, Nagalingum N, Macklin J, Liu J, Ranade S, Cui H (2017) Building the "Plant Glossary"—A controlled botanical vocabulary using terms extracted from the Floras of North America and China. Taxon 66 (4): 953-966. https://doi.org/10.12705/664.9
- Horridge M, Tudorache T, Nuylas C, Vendetti J, Noy N, Musen M (2014) WebProtégé: a collaborative Web-based platform for editing biomedical ontologies. Bioinformatics 30 (16): 2384-2385. https://doi.org/10.1093/bioinformatics/btu256
- Sachs J, Pender J, Lujan-Toro B, Macklin J, Haase P, Malik R (2019) Using Wikidata and Metaphactory to Underpin an Integrated Flora of Canada. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.38627