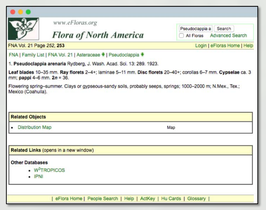
Bringing a Semantic MediaWiki Flora to Life

Jocelyn Pender‡, Joel L. Sachs‡, James A. Macklin‡, Hong Cui§, Andru Vallance|, Beatriz Lujan-Toro‡, Thomas Rodenhausen§, Melanie Belisle-Leclerc‡, Geoffrey Levin¶

‡ Agriculture and Agri-Food Canada, Ottawa, Canada § University of Arizona, Tucson, United States of America | Unaffiliated, Bristol, United Kingdom ¶ University of Illinois, Champaign, United States of America

Flora of North America



The existing FNA web presence

The **Flora of North America** (FNA) is a groundbreaking collaborative project (1993–ongoing). It assembles for the first time, in an authoritative publication, information on the names, taxonomic relationships, distributions and morphological characteristics of all plants native and naturalized found in North America north of Mexico.

Motivation

The existing web representation (efloras.org) of the FNA needs improvement. Our team has been working diligently to capitalize on existing **Natural Language Processing** (NLP) tools built for parsing of biodiversity data (Explorer of Taxon Concepts; Cui et al. 2016)

The aim is to present the FNA online in both machine and human readable formats (i.e., treatments can be viewed as easily as the data extracted from them can be queried). This would result in enhanced **mobilization** and **usability** of authoritative floristic treatment data, enabling data linkage to a Biodiversity Knowledge Graph (Page 2016), for instance.

Semantic MediaWiki

Semantic MediaWiki (SMW) is a free extension to MediaWiki, the open source software that powers Wikipedia and thousands of other wikis on the web (Wikistats, 2018).

SMW is a **knowledge management system**. It allows Wiki pages to be annotated **semantically** using SMW markup. Semantic annotations are stored by SMW and can be queried anywhere on the platform.

In SMW, an article on the country Canada might look like this:

The user sees	SMW markup on the backend is
… the capital city is Ottawa …	… the capital city is [[Has capital::Ottawa]] …
The user can then ask the SMW	
{{#ask [[Has capital::Ottawa]]}}	
SMW will return:	
Canada	

We've built a custom **FNA extension** that manages taxonomic concepts. Data can be easily published from SMW to the **Semantic Web using RDF or CSV** formats.

Current Status

We are happy to announce that our **public beta is now live.** The beta presents semantic properties such as treatment authors, taxon rank, taxon parent, print volume, elevation, phenology and habitat strings. It renders beautifully on any device!

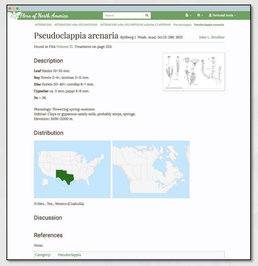
beta.semanticfna.org

A main feature of the FNA is its plethora of illustrations

Since our last announcement in 2016, we've worked closely with FNA editors to add an **emphasis on artwork**, **taxonomic hierarchy** browsing, and **parsed references** enabling reference author search, with BHL and PubMed linking.

Our **public development instance** is also live, and includes more fine grained properties (including parsed **morphology characteristics** for select volumes, this is ongoing!) and our latest deployed features.

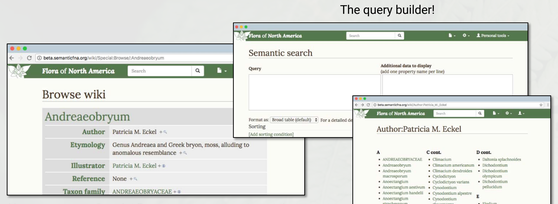
dev.semanticfna.org



The dev FNA SMW web presence

How to Use the FNA SMW

The query builder!



You can access Special:Browse on any page by adding it to the URL directly

A treatment author page listing all described species

Properties are the main way semantic data are stored in SMW. Explore any property by clicking on it at the base of a page (Facts about box), or by going to its Property page directly.

You can query properties and their values using special pages (via /wiki/**Special:SearchByProperty** accessed directly or by using the icon 🔍 or the query builder /wiki/**Special:Ask**).

Categories store pages into logical groupings. Explore by clicking categories found at the bottom of a treatment page or by going to a category page directly (/wiki/Category:Example).

We've developed **special pages** for properties of particular interest (i.e., treatment author pages, plate illustrator pages and references). Find these as links on treatment pages.

Facts about is a box at the base of treatment pages lists all known properties.

Clicking **Special:Browse** via the 👁 icon allows you to view properties and other information about a page.

FNA SMW for Hypothesis Testing

A major use case for the FNA SMW is the ability to **use aggregate floristic treatment data to test hypotheses**. Below is a simple demonstration of the power of parsed FNA treatment data.

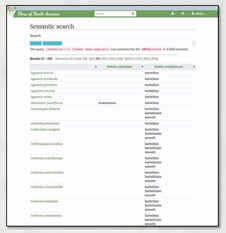
Plants with armature (thorns, bristles, trichomes, etc.) are often found at the edges of field and forests. An Asteraceae botanist may ask the question, are these plants more often reddish in color, to protect against light exposure?

Using #ask queries on the SMW platform, users can **query for datasets and export results** to various formats (JSON, CSV, RSS, RDF). In this example, the botanist may wish to query for armature and coloration.

As a first step, we decided to explore the relationship between bristle architecture and bristle coloration. Using the query to the right, we exported these two variables for 583 Asteraceae species to CSV.

```
{{#ask:
[[Armature::+]]
[[Coloration::+]]
[[Taxon rank::species]]
[[Taxon family::ASTERACEAE]]
|?bristle coloration
|?bristle architecture
}}
```

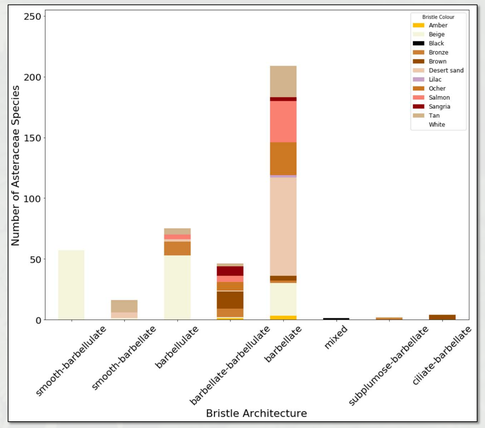
An #ask query used to select semantic data



The semantic search page where users can export data to CSV

The visualization below shows that **smoother bristles appear to be lighter in color than barbellate bristles**. Barbellate-barbellulate bristles are the most red in color.

A chi-squared test of independence ($p=8.1 \times 10^{-112}$) showed the two variables were significantly related. However, a further investigation into substantial bias introduced by authorship and data coverage would be required to reach a conclusion.



A histogram of bristle architecture groups, colored according to bristle coloration. Colors were aggregated into 12 groups, and bristle architecture types were aggregated into 8 groups. Included are data from 583 Asteraceae species of the FNA. Data were pulled from the SMW FNA platform, cleaned in Excel and Python, and visualized using the Seaborn package in Python.

Discussion

Significant data curation from PDF to parsed FNA text was required (i.e., exceptions to the rule tripping up programmatic formatting, nonsensical morphology parse etc.). Data curation challenges remain, however a main roadblock is how to *find* data issues without manually combing through thousands of treatments and properties.

Struggles with SMW software and its idiosyncrasies were unexpected but plentiful. While SMW does provide many essential data management features out-of-the-box, doing anything that SMW has not explicitly enabled is difficult. Additionally, a long list of maintenance tasks must be performed on a regular basis to iron out bugs.

Acquiring SMW web developer time was not obvious. We experienced challenges keeping the project moving, while working with our developer across the ocean.

In our opinion, the implications of a project like this are vast. The number of hypotheses that can be addressed using FNA data alone or in combination with other open data sources is *massive*, and we are beginning to develop a list. Please feel free to add your own: http://bit.ly/1S1nGjF

Linkages across data (i.e., formats and exchange protocols) are essential for supporting future botanical science, as well as potential enterprise and community purposes (e.g., agriculture). The FNA SMW project is a first step to integrating authoritative botanical information into the Semantic Web, a Biodiversity Knowledge Graph and other broadly applicable knowledge bases (e.g., Google Knowledge Graph).

Next Steps

1. Add additional data sources to our treatment pages (e.g., specimen data, plant images, etc.).
2. Build a complete Flora of Canada, by integrating authoritative and overlapping sources such as the FNA. Indeed, we are starting to perform the same process (NLP, import to SMW) on other flora sources (e.g., the Flora of Manitoba).

If you have ideas for data sources we should include, or electronically accessible floras we should parse, please let us know!

Other possible collaborations including building ontological infrastructure to support traversing of semantic data, and developing standardized morphological standard properties. Please contact us to collaborate!

References

Cui H *et al.* (2016) Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics 17: 471. https://doi.org/10.1186/s12859-016-1352-7

Page R. (2016) Towards a biodiversity knowledge graph. Research Ideas and Outcomes 2: e8767. https://doi.org/10.3897/rio.2.e8767

Contact us!

Jocelyn.Pender@canada.ca
Joel.Sachs@agr.gc.ca

Agriculture and Agri-Food Canada | Agriculture et Agroalimentaire Canada