BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e38627
doi: 10.3897/biss.3.38627

OPEN ACCESS

Conference Abstract

Using Wikidata and Metaphactory to Underpin an Integrated Flora of Canada

Joel Sachs[‡], Jocelyn Pender[‡], Beatriz Lujan-Toro[‡], James Macklin[‡], Peter Haase[§], Robin Malik[§]

‡ Agriculture and Agri-Food Canada, Ottawa, Canada
§ Metaphacts GmbH, Walldorf, Germany

Corresponding author: Joel Sachs (joel.sachs@agr.gc.ca)

Received: 29 Jul 2019 | Published: 08 Aug 2019

Citation: 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: e38627.
https://doi.org/10.3897/biss.3.38627

Abstract

We are using Wikidata and Metaphactory to build an Integrated Flora of Canada (IFC). IFC will be integrated in two senses: First, it will draw on multiple existing flora (e.g. Flora of North America, Flora of Manitoba, etc.) for content. Second, it will be a portal to related resources such as annotations, specimens, literature, and sequence data.

Background

We had success using Semantic Media Wiki (SMW) as the platform for an on-line representation of the Flora of North America (FNA). We used Charaparser (Cui 2012) to extract plant structures (e.g. “stem”), characters (e.g. “external texture”), and character values (e.g. “glabrous”) from the semi-structured FNA treatments. We then loaded this data into SMW, which allows us to query for taxa based on their character traits, and enables a broad range of exploratory analysis, both for purposes of hypothesis generation, and also to provide support for or against specific scientific hypotheses.

Migrating to Wikidata/Wikibase

We decided to explore a migration from SMW to Wikibase for three main reasons: simplified workflow; triple level provenance; and sustainability.

Simplified workflow: Our workflow for our FNA-based portal includes Natural Language Processing (NLP) of coarse-grained XML to get the fine-grained XML, transforming this XML for input into SMW, and a custom SMW skin for displaying the data. We consider the coarse-grained XML to be canonical. When it changes (because we find an error, or we improve our NLP), we have to re-run the transformation, and re-load the data, which is time-consuming. Ideally, our presentation would be based on API calls to the data itself, eliminating the need to transform and re-load after every change.

Provenance: Wikidata's provenance model supports having multiple, conflicting assertions for the same character trait, which is something that inevitably happens when floristic data is integrated.

Sustainability: Wikidata has strong support from the Wikimedia Foundation, while SMW is increasingly seen as a legacy system.

Wikibase vs. Wikidata

Wikidata, however, is not a suitable home for the Integrated Flora of Canada. It is built upon a relatively small number of community curated properties, while we have ~4500 properties for the Asteraceae family alone. The model we want to pursue is to use Wikidata for a small group of core properties (e.g. accepted name, parent taxon, etc.), and to use our own instance of Wikibase for the much larger number of specialized morphological properties (e.g. adaxial leaf colour, leaf external texture, etc.) Essentially, we will be running our own Wikidata, over which we would exercise full control. Miller (2018) decribes deploying this curation model in another domain.

Metaphactory

Metaphactory is a suite of middleware and front-end interfaces for authoring, managing, and querying knowledge graphs, including mechanisms for faceted search and geospatial visualizations. It is also the software (together with Blazegraph) behind the Wikidata Query Service. Metaphactory provides us with a SPARQL endpoint; a templating mechanism that allows each taxonomic treatment to be rendered via a collection of SPARQL queries; reasoning capabilities (via an underlying graph database) that permit the organization of over 42,000 morphological properties; and a variety of search and discovery tools.

There are a number of ways in which Wikidata and Metaphactory can work together, and we are still exploring questions such as: Will provenance be managed via named graphs, or via the Wikidata snak model?; How will data flow between the two platforms? Etc. We will report on our findings to date, and invite collaboration with related Wikimedia-based projects.

Presenting author

Jocelyn Pender

Presented at

Biodiversity_Next 2019

References

- Cui H (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of the American Society for Information Science and Technology 63 (4): 738-754. https://doi.org/10.1002/asi.22618
- Miller M (2018) Wikibase for Research Infrastructure—Part 1. https://medium.com/@thisismattmiller/wikibase-for-research-infrastructure-part-1-d3f640dfad34. Accessed on: 2019-4-04.