Mapping of glossary terms from the Flora of North America to the Plant Ontology enhances both resources

Ramona Walls[1,*] Hong Cui[2], James A. Macklin[3], Chris Mungall[4], Laurel Cooper[5], Dennis Stevenson[1] and Pankaj Jaiswal[5]

[1] New York Botanical Garden, Bronx, New York, USA

[2]School of Information Resources and Library Science, University of Arizona, Tucson, Arizona, 85719 USA

[3]Research Branch, Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada

[4]Lawrence Berkeley National Lab, Berkeley, California, USA

[5]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA

1 INTRODUCTION

Traditional taxonomic literature can provide a wealth of data, but access to that data is limited by its free-text format. Taxonomic treatments such as the Flora of North America (FNA Editorial Committee 1993) consist of terse descriptions of the characters used to identify taxa, such as:

> **"**…Leaves usually alternate or opposite, sometimes in basal rosettes, rarely in whorls; rarely stipulate, usually petiolate, sometimes sessile…"

Converting taxonomic descriptions to computer-readable format makes them available for automatic retrieval and large-scale analyses. Ontologies such as the Plant Ontology (PO) play a central role in automatic annotation, by providing semantic meaning for the words in a description. We used automated and manual methods to map terms from the Categorical Glossary for the Flora of North America Project (http://128.2.21.109/fmi/xsl/FNA/home.xsl) to the PO.

2 METHODS

Terms from the pre-existing categories of "structure", "feature", or "nominative" were extracted from the FNA glossary, roughly corresponding to the PO class *plant anatomical entity*. An automated mapping to PO release 16 was done using Obol software (Mungall 2004). We manually checked the automated mapping, and removed any matches that were incorrect. Remaining glossary terms were either manually mapped to existing PO terms, classified as inappropriate for the PO, or marked to be added to the PO.

3 RESULTS AND DISCUSSION

839 terms were extracted from the FNA glossary, compared to 1080 terms in the *plant anatomical entity* branch of the PO. Using text matching, Obol mapped 264 FNA terms to 313 existing PO terms or synonyms, including 49 FNA terms that matched more than one PO term or synonym. Most duplicate matches arose because the PO has many synonyms in Spanish that are identical to the English term name. Only 30 Obol matches had to be removed, in cases where the FNA has multiple terms with the same name but separate meanings that should map to separate PO terms. A curator mapped the remaining FNA glossary terms to PO terms, based on the FNA and PO definitions.

A total of 193 FNA terms mapped to existing PO primary term names and 126 mapped to existing synonyms. 333 FNA terms had the same meaning as existing PO terms and have been added as synonyms to the PO, citing the FNA glossary as the source. 143 unique new terms will be added to the PO, corresponding to 180 FNA glossary terms. 118 FNA terms could not be mapped to PO terms, either because they were too vague (12 terms, e.g., FNA:*lamella*, which could apply to many different tissue types), because they are subcellular components and belong in the Gene Ontology (5 terms, e.g., FNA:*flagella*), or because they are better modeled as qualities (93 terms, e.g., FNA:*puncta* is better treated as the quality punctate).

The PO is fairly extensive in its coverage of plant anatomical entities, as many of missing terms are specialized structures found only in a few taxa. The PO benefits from this mapping through increased coverage of plant terminology. Text mining tools such as CharaParser (Cui 2012) that are being developed to mine taxonomic descriptions can now use the PO more effectively for automated text annotation and in return mine more candidate terms from the literature to further enrich PO. The mapping of FNA IDs to PO IDs is available at http://tinyurl.com/FNAPOmapping.

ACKNOWLEDGEMENTS

NSF-IOS: 0822201 to the Plant Ontology Project, and the Flora of North America Association.

REFERENCES

Cui, H. 2012. CharaParser for fine-grained semantic annotation of organism morphological descriptions. J. of Am. Soc. of Information Science and Technology. 63(4) doi:10.1002/asi.22618

FNA Editorial Committee, eds. 1993. *Flora of North America North of Mexico*. 16+ vols. New York and Oxford.

Mungall, Christopher J. 2004. "Obol: Integrating Language and Meaning in Bio-ontologies." *Comparative and Functional Genomics* 5 (6-7) (August 1): 509–520. doi:10.1002/cfg.435