Biodiversity Information Science and Standards : Conference Abstract

Print

Flora Prepper: Preparing floras for morphological parsing and integration

Jocelyn Pender ‡

‡ Agriculture and Agri-Food Canada, Ottawa, Canada

Corresponding author: Jocelyn Pender (pender.jocelyn@gmail.com)

Received: 27 Jun 2019 | Published: 04 Jul 2019

Citation: Pender J (2019) Flora Prepper: Preparing floras for morphological parsing and integration. Biodiversity Information Science and Standards 3: e37743. https://doi.org/10.3897/biss.3.37743

Abstract

The increased availability of digital floras and the application of optical character recognition (OCR) to digitized texts has resulted in exciting opportunities for flora data mining. For example, the software package CharaParser has been developed for the semantic annotation of morphological descriptions from taxonomic treatments (Cui 2012). However, after digitization and OCR processing and before parsing of morphological treatments can begin, content types must be annotated (i.e., text represents names, morphology, discussion or distribution). In addition to enabling morphological parsing, content type annotation also facilitates content search and data linkage. For example, by annotating pieces of a floral treatment, assertions from various floras of the same type can be combined into a single document (i.e., a "mash-up" floral treatment). Several products and pipelines have been developed for the semantic annotation, or mark-up, of taxonomic documents (e.g., GoldenGATE, FlorML; Sautter et al. 2012, Hamann et al. 2014). However, these products lack a combination of both ease of implementation (e.g., the ability to run as a script in a programmatic workflow) and the use of modern parsing methods, such as text mining and Natural Language Processing (NLP) approaches.

Here I present a pilot project implementing text mining and NLP approaches to marking-up floras implemented in a Python package. I will describe the success of the project, and summarize lessons learned, especially in relation to previous flora markup projects. Annotation of existing flora documents is an essential step towards building next-generation floras (i.e., mash-ups and enhanced floras as platforms) and enables automated trait extraction. Building an easy-to-use access point to modern text mining and NLP techniques for botanical literature will allow for more flexible and

responsive flora annotation, and is an important step towards realizing botanical data integration goals.

Keywords

Flora, text mining, natural language processing, Python, annotation

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
- Hamann T, Müller A, Roos M, Sosef M, Smets E (2014) Detailed mark-up of semi-monographic legacy taxonomic works using FlorML. Taxon 63 (2): 377–393. https://doi.org/10.12705/632.11
- Sautter G, Böhm K, Agosti D (2012) Semi-automated XML markup of biosystematic legacy literature with the GoldenGate editor. Biocomputing 2007 https://doi.org/10.1142/9789812772435_0037