A Web Application for Extracting Key Domain Information for Scientific Publications using Ontology

Weijia Xu Amit Gupta
Texas Advanced Computing Center
University of Texas at Austin
Austin, Texas USA
{xwj, agupta}@tacc.utexas.edu

Pankaj Jaiswal
Department of Botany and Plant Pathology
Oregon State University
Oregon, Portland, USA
jaiswalp@oregonstate.edu

Crispin Taylor Patti Lockhart
American Society of Plant Biologists
Rockville, Maryland, USA
{ctaylor, plockhart}@aspb.org

***Abstract*— We present demos of an ongoing project, domain informational vocabulary extraction (DIVE), which aims to enrich digital publications through entity and key informational words detection and by adding additional annotations. The system implements multiple strategies for biological entity detection, including using regular expression rules, ontologies, and a keyword dictionary. These extracted entities are then stored in a database and made accessible through an interactive web application for curation and evaluation by authors. Through the web interface, the user can make additional annotations and corrections to the current results. The updates can then be used to improve the entity detection in subsequent processed articles. Although the system is being developed in the context of annotating journal articles, it can also be beneficial to domain curators and researchers at large.**

Keywords—component; Information systems applications; Information integration; Ontology; Text Mining

I. INTRODUCTION

Due to its technical depth and rich informational content, a journal article often requires that readers, domain experts, and curators invest significant amounts of time and effort to fully comprehend and make intelligent use of its content. This can be especially true in emerging areas, where novel ideas and new terminologies may be presented without precedent. As new technologies accelerating scientific discovery and more content becomes available online, the number of new articles that must be read and understood continues to rise. Therefore, there is a pressing need to develop computational methods and tools that can enrich the information content of digital publications, improve its accessibility and utility, and facilitate the readers' understanding by creating links between journal articles and relevant database entities during the article production process. To address this issue, we present software developments from an ongoing project, DIVE, which features auto extraction of informational vocabulary, web based access and curation tools, and integration into the digital publication process.

The framework implements several strategies in entity extraction, including using regular expression rules, ontology and a keyword dictionary. The results of the extracted biological entities are then stored in a database and made accessible through an interactive web application for curation and evaluation by authors and other domain experts. Through the web interface, a user can make additional annotations and corrections to the initial result set. The updates are stored and managed via the relational database for future improvements.

We present application demos to illustrate this framework using a sets of plant biology articles. We detail the design and implementation of the system, including entity detection, the extraction pipeline, and the web interface; we also present a use case demonstration. We would like to engage publishers and biology data curators in discussion and feedback.

There are three major steps in processing the documents: text extraction, entity candidate extraction, and candidate assessment. The input for the text extraction process is the structured document tagged by Journal Article Tag Suite (JATS) [1]. The input document is processed into two data structures for textual data and structural data. This dual data structure allows for efficient text processing of the publication content while still being able to easily retrieve the meta-structure around a particular set of words during the subsequent steps of processing. To identify informational vocabulary candidates, our application implemented four sets of extraction rules: regular expression rules, word dictionary, publishing convention, and ontology rules. The ontology rules utilize five biological ontologies including gene ontology [2], plant ontology [3], plant trait ontology [4], plant environment condition ontology [5] and Chemical Entities of Biological Interest (ChEBI) [6]. The results from document processing are stored in a MySQL database and serve as data storage for the web application.

The web front end in our prototype is implemented using Django (v 1.8). Based on Python, the web front is easily programmable, extensible, and is pluggable with multiple popular databases. It forms the presentation layer of this system, relying on the back end code to run the entity extraction algorithms from the manuscript and to transfer the results in a JSON format.

The system can benefit the entire life cycle of the digital publication, from initial manuscript submission to publishing the article and presenting information to readers. At the initial manuscript submission stage, the manuscript can be processed to extract known key informational vocabulary, such as biological entities, as well as to identify potential new technical words. That information may be used by editors to identify appropriate reviewers for the manuscript. After the article has been accepted for publication, additional information about the key informational words, such as links to external repositories or reference sites, may also be embedded during the pre-publication production process to enrich the information content and accessibility. Publication curators may also

leverage the information for curation. New information defined and verified by experts may also be injected to other information resources, such as Planteome [7].

II. Application Features Overview

Let us use an example to illustrate the features of our web interface, thereby displaying the various views, layouts, and functions available. Our prototype includes 609 manuscripts from the journal *Plant Physiology*.

Figure 1. Collection paginated view

The publication list view (Figure 1) is a paginated list of all the articles with an external DOI reference and the article title.

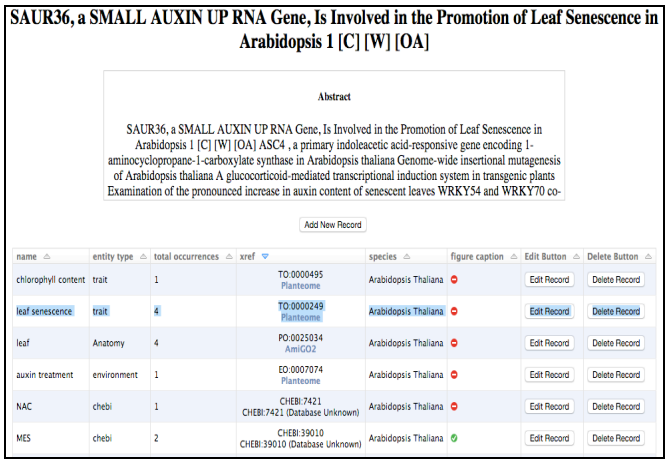

Figure 2. Interface for exploring entities in a publication

Figure 2 shows example of exploring entities extracted from a full article (i.e. 1002.xml in Figure 1). In the top of the page, the title and abstract of the article [8] are presented to give user some context of the manuscript. The list of entities found in this article are organized in a table. Each row includes information like name, type and number of occurrences in the article of an entity. The XRef column presents possible matches to existing ontology terms. If avalalible, links are also presented to other online databases with more information of that entity. For example, "leaf senescence" are mathced to a term in triat ontology with a link to the corresponding entry in the Planteome ontology database. In another example, "MES" are matched to a term in ChEBI but a link to the external database are missing at the time. The "species" column shows prediction on which species this entitiy is likely assoicated with based on the proximity of the term with the species name appeared in the article and/or indicated by the ontologies. The "Figure caption" column shows wheather the entity has been used within a figure caption in the article.

It is important to note that some entities can be matched to multiple onotology terms. Such cases are currently resolved based on the general priorites we assinged to each extraction rule during the extraction phase. However, for a particular article, the result may not always be the most approrirate one. The entity extraction phase can also generate phrases, while matching the ontology term alphabtetically, are not used for the purpose implied by the term. Those situations requires expert knolwedge and input.

Theorefore, each row also includes user control buttons for editing the record. Figure 3 shows an example of the Entitiy record editing interface. In this view, there are editable fields of this record where a user may correct or enter new values. A dynamic search box can be used to search and add new species into the species menu, if the appropriate species was not detected or inferred from the article. This search box uses an online service from NCBI to provide a very comprehensive list of options as the user dynamically types into it. Sentences of occurrence of this entity are extracted from the manuscript with the entity name highlighted in yellow. This again provides better, almost complete context information for this entity, as per the manuscript text.

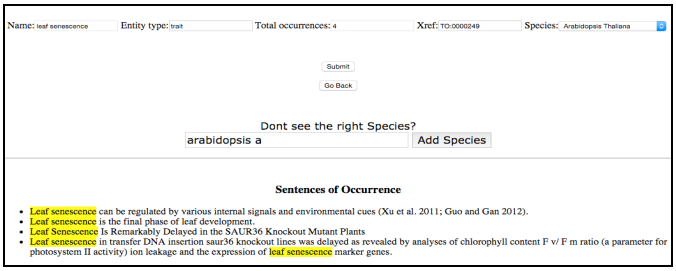

Figure 3. Interface for showing/editing entity details

The prototype is still under development, and we welcome feedback from domain researchers and publishing professionals for future developments and improvements.

Acknowledgment

DIVE is partially supported by CyVerse (NSF award DBI-0735191 and DBI-1265383) and the Gramene, a Comparative Plant Genomics Database (NSF award IOS-1127112).

References

[1] National Center for Biotechnology information. *Journal Article Tag Suite*. http://jats.nlm.nih.gov/, 2013.

[2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis A.P. (2000) Gene Ontology: tool for the unification of biology." *Nature genetics* 25, no. 1, pp 25-29.

[3] Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Zapata, F. (2005). Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comparative and Functional Genomics*, *6*(7-8), 388–397. http://doi.org/10.1002/cfg.496

[4] Arnaud, E, Cooper L, Shrestha, R, Menda, N, Nelson, R T, Matteis, L, Skofic M (2012) Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes in *KEOD*, pp220-5

[5] Plant Enviroment Condition Ontology, http://bioportal.bioontology.org/ontologies/PECO#

[6] Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, *36* (Database issue), D344–D350.

[7] Cooper, L. and Jaiswal, P. (2016) The Plant Ontology: A Tool for Plant Genomics. *Plant Bioinformatics: Methods and Protocols*, 89-114

[8] Hou, K., Wu, W., & Gan, S. S. (2013). SAUR36, a small auxin up RNA gene, is involved in the promotion of leaf senescence in Arabidopsis. *Plant physiology*, *161*(2), 1002-1009.