Domain Informational Vocabulary Extraction Experiences with Publication Pipeline Integration and Ontology Curation

Amit Gupta [1]*; Weijia Xu [1]†; Pankaj Jaiswal [2]‡; Crispin Taylor [3]§; Jennifer Regala [3]¶

[1]Texas Advanced Computing Center, University of Texas at Austin, Austin, Texas, USA
[2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA
[3]American Society of Plant Biologists, Rockville, Maryland, USA

ABSTRACT

We will present updates on an ongoing project DIVE (Domain Informational Vocabulary Extraction), a system designed for extracting domain information from scientific publications. DIVE implements an ensemble of text mining methods for biological entity extraction from article text. DIVE also attempts use the co-occurrence patterns of these entities to establish probable relationships between them. DIVE also features an improved web interface for expert user curation of extracted information, thereby providing a means for a constantly growing and expert curated body of domain information for an article corpus. We also discuss our experiences from successful integration of DIVE with the publishing pipeline for two prominent Plant Biology Journals (*The Plant Cell* and *Plant Physiology*) from ASPB (American Society of Plant Biologists). The extracted results are embedded at the end of the final proof of the published article to enhance its accessibility and discoverability. Furthermore, DIVE tracks expert user curation actions on its web interface for future training and improvement of the entity detection algorithm.

1 INTRODUCTION

Synthesizing information from a large corpus of journal articles or technical documents requires a great deal of time, non-trivial effort to understand and digest the contents and also demands significant expertise from the reader. Furthermore, journal articles are often the first textual appearance of new terms, concepts, ideas and discoveries that are without precedence. This furthermore exacerbates the already difficult task of extracting and preserving information contained in these articles. Over various scientific domains, these articles amount in the millions and with continuous publication they continue to grow at an astonishing rate. To keep up with the constant influx and volume of new information, computationally analyzing these growing corpora of technical documents seems like a natural solution.

To address this problem, we have designed and implemented a system called DIVE (Domain Informational Vocabulary Extraction) Weijia Xu (2016a). DIVE employs text mining methods for entity extraction, and utilizes cyberinfrastructure for online processing and service support. To detect entities of interest, DIVE uses an ensemble of methods from text mining like keyword dictionary matching, regular expression rules and cross checking against known ontologies. The results are stored in a relational database and made available through a web interface (Weijia Xu, 2016b) for expert user curation. We have also successfully integrated DIVE with the publication pipeline for two internationally recognized Plant Biology Journals (namely, *The Plant Cell* and *Plant Physiology*) from ASPB (American Society of Plant Biologists. The architecture of this integration is also presented. We also describe some advanced features added to DIVE like corpus level Association Analysis, Article Search and our Entity Ranking Model. We conclude with planned features under current development and scheduled to be rolled out in the near future.

*agupta@tacc.utexas.edu
†xwj@tacc.utexas.edu
‡jaiswalp@science.oregonstate.edu
§ctaylor@aspb.org
¶jregala@aspb.org

2 DIVE ARCHITECTURE

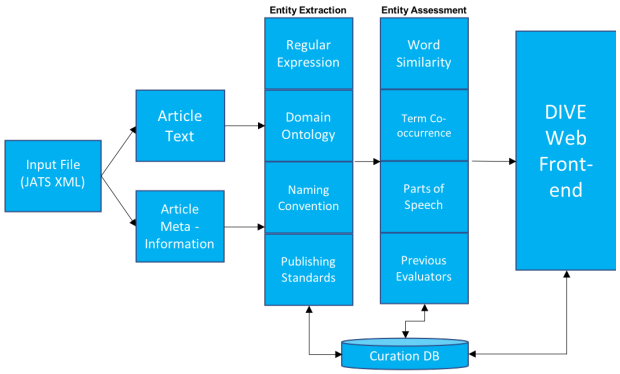

Fig. 1. DIVE Architecture and Processing Workflow

The architecture of DIVE is as shown in Figure (1). The input to DIVE is the article in a structured format like Journal Article Tag Suite (JATS) XML standard (NIH, 2018). There are essentially 3 stages in processing this file.

- Text Extraction: Here text is extracted from the structured format JATS XML. For efficient retrieval DIVE tracks both the pure text data as well as the structural data and other metadata (i.e Section information, formatting information, global position of each token etc) and maintains the adequate mappings.

- Entity Candidate Extraction: Here a few rules are employed for extract entity candidates from text tokens.
  - Regular Expressions - where Entity candidates can be selected based on patterns indicated by Regular Expression rules. Furthermore, new rules can be added on the fly to DIVE as and when required.
  - Keyword Dictionary - where tokens can be matched against a list of known words known to be entities or to even eliminate known words that are not to be considered as entities.
  - Publishing Standards - where rules from publishing conventions are used to identify possible entity candidates. For example if something important is italicized, bold or quotes. Ontology, where Entity candidates are cross-checked against known ontology rules. In the case of Plant Biology articles, DIVE uses GRAMENE (Tello-Ruiz *et al.*, 2018), Arabidopsis Information Portal (Araport, 2018), CHeBI (Hastings *et al.*, 2016) and Plant Ontology Foundry (2018).
- Entity Assessment: The detection techniques above provide varying degrees of accuracy. Once a list of entity candidates is identified, they have to be assessed for accuracy. We use previously validated results and co-location with other verified entities as methods to validate entities. The primary means of verification is the expert user verifying the entities (via the DIVE web interface, mentioned below).

All of the above is stored in a relational database that can be analyzed with complex queries later. DIVE also features a web interface, implemented in the Django Web Framework (Django, 2018), for expert user curation of extracted entity information. Here the user has the opportunity to Edit, Add and Delete entity information related to the article. The website itself is backed by the relational database with the entity information.

These design features allows for 2 things:

- As DIVE processes more articles from a specific scientific domain and its corpus size increases, the web interface provides means to create a constantly growing and expert curated body of domain information represented in that article corpus.
- Besides tracking the entity information itself, the backend database can track expert user actions on the web curation interface. This forms a perfect testbed to test and develop automated machine learning algorithms to both improve the entity recognition and to recommend curation changes to authors based on historical data of such changes.

3 ASSOCIATION ANALYSIS

We employ Association Analysis, a Machine Learning method, to detect possible relationships between detected entities based on their co-occurrence patterns in the text. We use the popular FP-Growth Algorithm (Han *et al.*, 2000) implemented in R for this analysis (Hahsler *et al.*, 2011). This algorithm basically infers the likelihood of 2 groups of entities occurring together based on their co-occurence patterns in the article text. Using such occurrence relationships between entity groups, frequently mirror or atleast serve as a hint towards the deeper interaction relationship that is often being explained in the article text. We also present a visualization of the top rules inferred by the algorithm to the user for

Fig. 2. Article Level Association Analysis

their edification in the web interface as seen in Figure (2). As this figure demonstrates, such pairs of these entity co-occurrence pattern can collectively show interesting patterns of possible interaction, worth analyzing by the expert user.

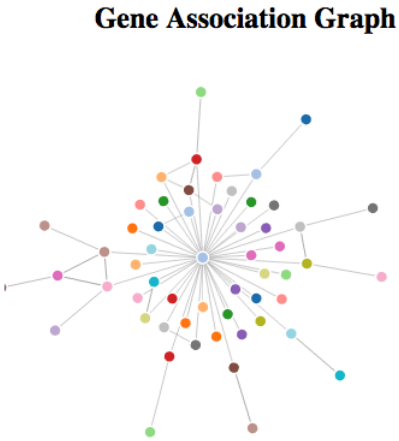

Fig. 3. Corpus Level Association Analysis for all Gene Entities

The algorithm may be scoped down narrow to the sentence level or scoped globally to the corpus level. Each level of granularity may reveal different entity relationships. In DIVE, these have been mostly scoped to the article level as shown in Figure (2). However, as the corpus being curated by DIVE grows larger, a global level association analysis might also prove insightful for the consumption of domain curation experts. As an exemplar, we can see a global level association analysis for roughly 2000 Plant Biology article corpus in Figure (3).

4 PUBLICATION PIPELINE INTEGRATION AND WEB SERVICE DESIGN

DIVE has been integrated into the publication pipeline of two Plant Biology Journals, namely *The Plant Cell* (ASPB, 2018b) and *Plant Physiology* (ASPB, 2018c) from ASPB (ASPB, 2018a). A company named Sheridan Journal Services develops and runs the publication and proofing software for these journals. The architecture of the integration is as shown in Figure (4). To enable this integration,

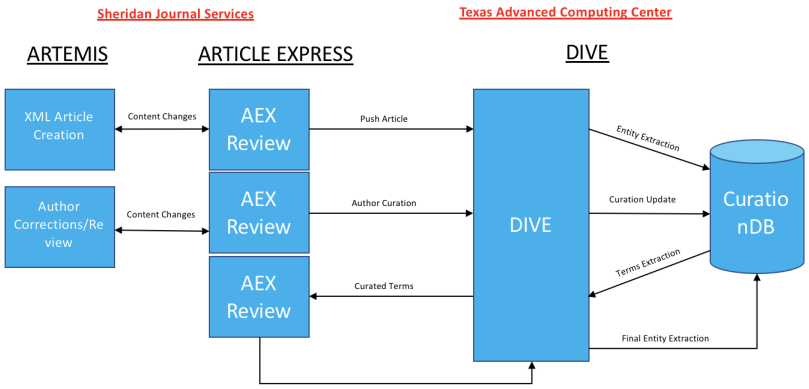

Fig. 4. Publication Pipeline Integration Architecture

DIVE functionality was exposed to the publication software as a web service with 2 endpoints.

- ***Article Endpoint***
  This endpoint receives two HTTP POST requests related to the article.
  - Article Push request: This request is to push a new article into DIVE. An article may also be pushed multiple times to DIVE during the publication process to incorporate proofing edits and corrections. This request contains the location of the cloud storage service from where this article may be retrieved. This request also contains metadata information about the article file being pushed for verification purposes.
  - Pull Curation request: This request is to pull a summary of curation information from DIVE about the article. It is usually done at the end of the proofing process and is embedded into the final proof of the article.
- ***Article Landing Page Endpoint***
  This endpoint is the landing page of the article. It is where the extracted entities for the article may be viewed and curated. This is where the authors are directed during the proofing process of their article to curate the terms. This page contains instructions of author curation actions and the list of entities extracted with metainformation. The authors may either verify its accuracy or do curation actions of edits, additions, deletions to this information. The extracted result is appended at the end of the final proof version of the publication with cross references to other known ontologies, to improve its accessibility and discoverability. These contributions are also tracked by the DIVE backend database and can serve to improve the information quality and improve future entity detection for DIVE.

5 ENTITY RANKING

When expert users arrive at their articles landing page, there is only a limited space to display information for curation. Furthermore, our experience was that too much information tends to confuse users and defeat the purpose of the curation interface. We therefore made a design decision to only display the most important entities (we display the 10 most important) to the users to curate, with a paginated option for the interested users to curate more entities. We designed a scoring function in DIVE to rank entities based on their type, whether or not they have a cross-reference to a mature external ontology like Gramene (Tello-Ruiz *et al.*, 2018).

Our three tiers of prioritization are:

- Genes
- Proteins
- Other entities like Plant Anatomies etc.

Within each tier an entity that is confirmed by an external ontology recieves higher priority and frequency of occurrence.

6 SEARCH INTERFACE

Expert users may also use our search interface to search for other articles within the corpus that contain an entity they are interested in. This can further help the articles discoverability amongst other users with overlapping domain expertise and interest. An example of the search interface is as shown in Figure (5). The articles are listed with their Title, Journal Name and other metadata like a doi link which leads to a site hosting a copy of the article.

7 FUTURE WORK

DIVE is under ongoing development and we have a few exciting features planned to be rolled out in the near future that would immensely benefit expert user curators from scientific domains. Some of them are summarize below:

- Based on the entities in the users article and/or the entities they have chosen to curate, DIVE will make article recommendations to the author by ranking articles on similarity. Such recommendations can also incorporate relationships discovered by Association Analysis.
- Based on historical curation action information for entities, appropriate recommendations will be made to users when they arrive on their articles landing page.
- We plan to incorporate full text indexing into search results to make it more comprehensive.

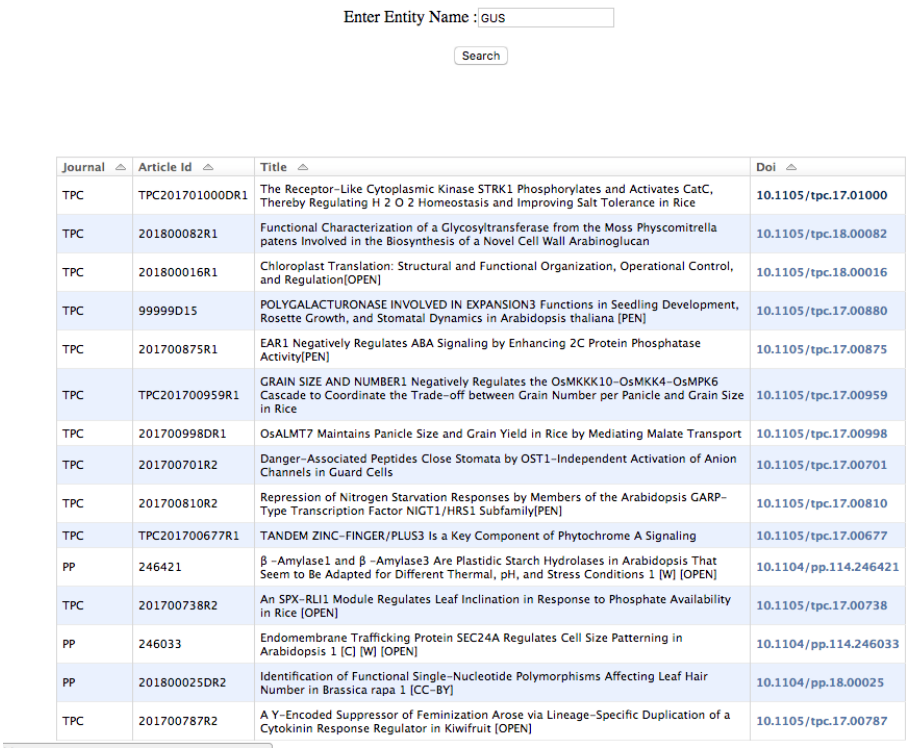

Fig. 5. DIVE Search Interface

8 CONCLUSION AND FUTURE WORK

Our early experience with deploying this solution in production with two internationally recognized Plant Biology Journals from ASPB has been promising. We are seeing enthusiastic participation by expert users and at present see about 10 curation actions per article in our corpus. Furthermore, although DIVE was developed for the use case of Plant Biology Journal Articles, it has been designed to be versatile and is quite readily adapted to document collections of any domain. We are presently investigating a use cases for corpora in other domains as well (ex. Aerospace Engineering) and will continue to expand in this area.

DIVE is under ongoing development and we have a few exciting features planned to be rolled out in the near future that would immensely benefit expert user curators from scientific domains. We are working on improving the search features to incorporate full text search, relationships uncovered by Association Analysis and are also investigating improvements to our entity detection algorithms. Other planned enhancements for expert users of DIVE include article recommendations and curation action recommendations. We aim to continue to build DIVE and deploy it important use case scenarios from many domains to benefit scientists and researchers at large make sense of their large document corpora.

ACKNOWLEDGEMENTS

DIVE is partially supported by CyVerse (NSF awards DBI0735191, DBI1265383) and by Gramene, A Comparative Plant Genomics Database (NSF award IOS 1127112).

REFERENCES

Araport (2018). Arabidopsis information portal. `https://www.araport.org/`.

ASPB (2018a). American society of plant biologists. `https://aspb.org/`.

ASPB (2018b). The plant cell. `http://www.plantcell.org/`.

ASPB (2018c). Plant physiology. `http://www.plantphysiol.org/`.

Django (2018). Django web framework. `https://djangoproject.com`.

Foundry, O. (2018). Plant ontology. `http://www.obofoundry.org/ontology/po.html`.

Hahsler, M., Chelluboina, S., Hornik, K., and Buchta, C. (2011). The arules r-package ecosystem: Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research*, **12**, 1977–1981.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, **29**(2), 1–12.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, **44**(D1), D1214–D1219.

NIH (2018). Journal article tag suite. `https://jats.nlm.nih.gov/`.

Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., Wei, S., Preece, J., Geniza, M. J., Jiao, Y., Lee, Y. K., Wang, B., Mulvaney, J., Chougule, K., Elser, J., Al-Bader, N., Kumari, S., Thomason, J., Kumar, V., Bolser, D. M., Naamati, G., Tapanari, E., Fonseca, N., Huerta, L., Iqbal, H., Keays, M., Munoz-PomerFuentes, A., Tang, A., Fabregat, A., DEustachio, P., Weiser, J., Stein, L. D., Petryszak, R., Papatheodorou, I., Kersey, P. J., Lockhart, P., Taylor, C., Jaiswal, P., and Ware, D. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*, **46**(D1), D1181–D1189.

Weijia Xu, Amit Gupta, P. J. C. T. P. L. (2016a). Enhancing information accessibility of publications with text mining and ontology. *International Conference on Biological Ontology*.

Weijia Xu, Amit Gupta, P. J. C. T. P. L. (2016b). A web application for extracting key domain information for scientific publications using ontology. *International Conference on Biological Ontology*.