Case Study Open Access

Weijia Xu*, Amit Gupta, Pankaj Jaiswal, Crispin Taylor, Patti Lockhart, Jennifer Regala

Improving Publication Pipeline with Automated Biological Entity Detection and Validation Service

https://doi.org/10.2478/dim-2019-0003
received November 1, 2018; accepted March 1, 2019.

Abstract: With the increasing amount of digital journal submissions, there is a need to deploy new scalable computational methods to improve information accessibilities. One common task is to identify useful information and named entity from text documents such as journal article submission. However, there are many technical challenges to limit applicability of the general methods and lack of general tools. In this paper, we present domain informational vocabulary extraction (DIVE) project, which aims to enrich digital publications through detection of entity and key informational words and by adding additional annotations. In a first of its kind to our knowledge, our system engages authors of the peer-reviewed articles and the journal publishers by integrating DIVE implementation in the manuscript proofing and publication process. The system implements multiple strategies for biological entity detection, including using regular expression rules, ontology, 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 authors 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 in the future. We describe our framework and deployment in details. In a pilot program, we have deployed the first phase of development as a service integrated with the journals Plant Physiology and The Plant cell published by the American Society of Plant Biologists (ASPB). We present usage statistics to date since its production on April 2018. We compare automated recognition results from DIVE with results from author curation and show the service achieved on average 80% recall and 70% precision per article. In contrast, an existing biological entity extraction tool, a biomedical named entity recognizer (ABNER), can only achieve 47% recall and return a much larger candidate set.

Keywords: entity extraction, digital curation, digital library, machine learning, ontology, text mining, natural language processing

1 Introduction

Over the past decade, advances in information technology have brought profound continuing transformations in media publishing industry including greatly increased content format, empowering authors with self-publishing and expanding accessibility with open access (Björk, 2017). In scholarly journal publishing and academia research community, hundreds of new journal titles and topics are introduced in each decade (Tenopir, C. & King, 2014). One notable change is the increasing volume of digital content. Not only new journals are created in a digital format but also many publishers have begun to digitize previously printed journals. The digital version of articles has increased accessibility and speeded up delivery of content to readers. New functionality can be implemented using the digital publishing process to enhance the traditional scholarly communication channel (Ware & Mabe, 2015). Among many challenges brought by digital publishing, improving the accessibility and use of domain knowledge embedded in the journal article is a central focus of this paper. As new technology keeps accelerating scientific discovery, the number of new scientific publications continues to rise accordingly. To keep up with the constant influx and volume of new information, automatically

***Corresponding author: Weijia Xu,** Texas Advanced Computing Center, University of Texas, Austin, USA, Email: xwj@tacc.utexas.edu
Amit Gupta: Texas Advanced Computing Center, University of Texas, Austin, USA
Pankaj Jaiswal: Oregon State University, Corvallis, Oregon, USA
Crispin Taylor: American Society of Plant Biologists, Rockville, Maryland, USA
Patti Lockhart: American Society of Plant Biologists, Rockville, Maryland, USA
Jennifer Regala: American Society of Plant Biologists, Rockville, Maryland, USA

analyzing these growing corpora of technical documents is a natural solution. 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 problem, we have designed and implemented a system called domain informational vocabulary extraction (DIVE) and deployed it as a cloud-based service in collaboration with the American Society of Plant Biologists (ASPB) (Xu, 2016; Gupta, 2018).

ASPB is a professional society established in 1924 and devoted to the plant sciences. ASPB manages and publishes two premium journals in the field of plant biology, *Plant Physiology* and *The Plant Cell*. Both journals are highly cited and accessed by readers from all over the world. Over the years, the areas of interest of these journals have evolved and expanded to include cellular and molecular biology, genetics, development, evolution, physiology, and biochemistry. Like in many scientific fields, publication is an important form of scholarly communication. Researchers publish their research findings in academic journals to share novel approaches and maximize knowledge creations (Chang, 2008). Therefore, new ideas and new terminologies are constantly invented and presented without precedents through journal publications. Owing to its technical depth and rich informational content, a scientific publication often requires significant amounts of time and effort for readers, domain experts, and curators to fully comprehend and make intelligent judgments. In fact, journal articles are often the first textual appearance of new terms, concepts, ideas, and discoveries that are without precedence. Synthesizing information from a large corpus of journal articles or technical documents requires a great deal of time and nontrivial effort to understand and digest the contents and also demands significant expertise from the reader. Capturing and collecting key domain information embedded in the article in a timely manner can improve article management and increase their readability and accessibility.

DIVE uses text-mining methods for entity extraction and utilizes cyberinfrastructure (CI) for online processing and service support. The analysis tool uses an ensemble of methods including keyword dictionary matching, regular expression rules, and cross-checking against known ontologies. 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 current results. The updates are stored and managed via the relational database for future improvements to the entity detection process. The service includes several components: automated informational vocabulary extraction based on existing domain ontologies, experts' validation and curation, and integration of results using the formal publication process. The contribution of the DIVE includes integration of CI for publishing pipeline, open extensible framework, incorporation with existing domain ontologies and other information sources, and an interface bridging author and knowledge curation.

The system can be integrated and benefits 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. This 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 prepublication 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 (Cooper & Jaiswal, 2016). In this paper, we present our development and deployment experiences of DIVE services with ASPB. We detail the design and implementation of the system including the entity detection, extraction pipeline, and the web interface, and we also present a use case demonstration. Additional features are under development.

2 Background and Related Work

The work presented here is related to several topic areas including named entity recognition in general and biological domains, journal publishing practice, biological ontology development, and using CI as a service.

2.1 Entity Recognition Methods and Tools

Entity recognition has originated from a classic problem in database research for detection of duplicate records

known as entity resolution (Naumann & Herschel, 2010; Elmagarmid et al., 2007). In large database systems in the real word, there are duplicate representations of the same object, also known as "entities" in relational databases. These duplicate records may not be exactly the same or share a common key reflecting their connection in the system. The value of each record could only be partially matched to other records or presents certain errors that make detecting these duplicates a difficult task. A typical entity resolution process includes data preparation, which transforms data into a uniform model; field matching, which breaks records by field and defines similarity models between fields; and duplicate record detection, which utilizes data mining techniques such as clustering and learning classification, to identify possible duplicated records (Herzog et al., 2007). Over the years, there have been a number of studies in this area (Köpcke et al., 2010; Christen, 2012). Entity resolution applications have also broadened into fields such as social network analysis and web data mining (Bilgic et al., 2006; Bhattacharya & Getoor,2007; Getoor & Diehl, 2005).

In text mining, named entity recognition is a common task for extracting information (Pasca et al., 2006; Grishman & Sundheim, 1996). The goal of the task is to identify and classify phrases in the corpus to predefined categories, such as names of persons, organizations, locations, times and dates, numerical values, and percentages (Nadeau & Sekine, 2007). Developing systems to automatically extract named entities are motivated through several contests and challenges and has become a research topic since early 2000s (Sang & Meulder, 2003; Doddington et al., 2004; Santos et al., 2006). Recognition of named entities requires leveraging linguistics grammar-based models. Early works were centered on handcrafted and rule-based algorithms. Rule-based systems can exploit features within the specific language to improve the system performance (Shaalan, 2010). Although a rule-based system can be highly efficient for a specific domain, the successes rely on integration of domain knowledge, which can be expensive to develop and hard to transfer.

Machine learning techniques have been adopted to learn features automatically in recent years. Supervised learning methods require a curated training data in which named entities have been identified and properly labeled. A typical workflow starts with processing the raw text to generate various features, such as annotations, position, and part-of-speech tags. Supervised learning methods are then used to derive an inference model based on the training dataset. Various learning methods have been explored over the years, including the Hidden Markov Model (Wang et al, 2014), support vector machine (Saha et al., 2010), maximum-entropy Markov model (MEMM)-based systems (Saha et al., 2009), logistic expression-based systems (Ek et al., 2011), and conditional random field (CRF) (Sutton & McCallum, 2012; Majumder et al., 2012; McCallum & Li, 2003). Just recently, deep neural network for named entity recognition has also been proposed (Dernoncourt et al., 2017). NeuroNER uses a deep recurrent neuron network to identify and classify named entities. However, supervised learning methods require a considerable amount of training data, which may be difficult to achieve. Unsupervised and semi-supervised learning methods are also proposed (Bhagavatula et al., 2012; Thenmalar et al., 2015). For more details and other approaches, readers can refer to a recent survey (Goyal et al., 2018). Still, successful solutions are often tightly coupled with the underlying language models and domains. It remains as a challenging problem to port models and algorithms working well for one domain to another problem domain. Substantial amounts of efforts have focused on algorithm improvements in a well-tuned, domain-specific scenario in practice. Thus far, no universal algorithm exists that can work reasonably well across a broad swath of domains.

The flourish of different entity recognition methods has resulted in various tools and libraries for entity recognition in practice. Two well-known, open-source, state-of-the-art libraries for general domains are spaCy and Stanford Named Entity Recognizer. The spaCy is a python package that provides comprehensive natural language processing (NLP) support. In the latest version, the implementation also leverages convolutional neuron network to improve parsing and inference (Kiperwasser & Goldberg, 2016). Stanford Named Entity Recognizer is a java library and implements linear chain CRF sequence mode (Chen & Manning, 2014). Both libraries include prebuilt general models for English language models and support training of new models with customized labels from the training dataset. In addition to these two models, ensemble approaches combining multiple techniques have also been reported for specific use cases in practice. Liu and Zhou proposed a system using linear CRF and cluster-based approach for recognizing entities from English tweets (Liu & Zhou, 2013). Ensemble classifiers and rule-based approaches have been combined in entity recognition in other languages (Ekbal & Saha, 2011; Guanming et al., 2009). Rule-based approach has also been used in conjunction with CRF techniques for biomedical domains (Li et al., 2009).

Named entity recognition in the biomedical domain has been a major problem of interest along with the

research for general domain (Kim et al., 2004). Earlier, biological entity recognition focused on gene and protein detection (Tanabe et al., 2005; Shen et al., 2003; Mihăilă & Ananiadou, 2014). The categories of interests have expanded to general technical terminologies used in biology, including gene names, protein names, biological sequences, organisms, specimen, mutant, and taxonomy. Successful techniques developed for the general domain have been adopted over the years. A Biomedical Named Entity Recognizer (ABNER) is one of the earliest tools for extracting biomedical named entities using CRFs (Settles, 2005). The research efforts have been fueled by a number of challenges and evaluation datasets advancing application of NLP in biomedicine (Huang et al., 2015; Pyysalo et al., 2007). Past research has explored both supervised methods (Tsai et al. 2006; Campos et al., 2013; Ju et al., 2011), and the more recent one focuses on unsupervised methods to detect and annotate biological entities (Zhang & Elhadad, 2013). Recently, deep learning techniques have also been explored to better represent word embedding for biomedical NLP (Habibi et al., 2017; Chiu et al., 2016). Recent successes have been found in systems utilizing ensemble classifiers and models. Zhu and Shen combined support vector machine and CRF approaches (Zhu & Shen, 2012). Habibi et al. proposed an approach, long short-term memory network (LSTM), using both deep neuron network and statistical models (Habibi et al., 2017). Despite many efforts made over the years, named entities recognition in the biological domain is still a challenging problem, and it is hard to achieve good performance as in the general domain.

2.2 Journal Publishing Standards

Over the years, publishers have gradually adapted to using structured documents for online publications in order to enrich their information content. A digital publication can be curated with additional annotations and external links. A commonly used standard for this purpose is Journal Article Tag Suite (JATS) (Huh, 2014; Huh et al., 2014). JATS is a NISO standard used by National Center for Biotechnology Information (NCBI). JATS refers to a superset of well-defined XML elements and attributes that may be used to tag journal articles. An article model (also referred to as a tag set) may be formed by using a subset (or all) of tags available in JATS. It is possible to define whole collection libraries of various article models using this standard. JATS is therefore widely used to annotate articles when creating digital article repositories to be hosted on the Internet. These annotations are, however, largely created by hand, a practice that is not scalable to large collections.

By virtue of our closed-loop architecture, we are able to collect annotation feedback from users via our web interface. In other words, this means that we can plug this information back into the JATS standard and introduce richer annotation information directly from sources (reviewers/authors/editors) that qualify as the relevant domain experts for that article. This workflow may be neatly integrated into the review cycle for a specific publication venue (journal/conference). This is also a highly scalable approach with the ability to produce annotations with significant qualitative improvements, which can then be curated.

2.3 Biological Ontology Development

Ontology is a set of controlled vocabularies with semantic to provide a formal encoding of concepts within a domain. The concepts and relations captured in ontology are used to form foundations for knowledge representation and management. While domain experts often define domain ontologies manually, the process of building ontologies requires extraction of concepts and their relations from existing data that aligns well with the task of entity recognition. Therefore, results from entity extraction can also be used to populate existing ontologies (Etzioni et al., 2005; De Boer et al., 2006; Song et al., 2009). Consequently, vocabularies in the existing ontology also provide an excellent source and evidence for named entity recognition.

In biology, there exist a number of ontologies for different bodies of knowledge. Gene ontology is probably the most well-known biological ontology among others (Ashburner et al., 2000). Gene ontology defines terms and relations among genes across all species. It forms a hierarchical representation with three major categories: biological process, molecular function, and cellular component. The number of biological ontologies is growing rapidly over the years. New ontologies are developed for individual species or biological functions. Our work leverages ontologies maintained by Planteome, which is an international collaboration effort to develop, enrich, and use plant ontologies for data cross-references and annotations (Cooper L et al., 2017). The Planteome project (http://www.planteome.org) provides a suite of references and species-specific ontologies for plants and annotations to genes and phenotypes. Ontologies serve as common standards for semantic integration of a large and growing corpus of plant genomics, phenomics, and

genetics data. The reference ontologies include the Plant Ontology, Plant Trait Ontology, and the Plant Experimental Conditions Ontology developed by the Planteome project, along with the Gene Ontology, Chemical Entities of Biological Interest (ChEBI), Phenotype and Attribute Ontology, and others. The project also provides access to species-specific crop ontologies developed by various plant breeding and research communities from around the world. All the Planteome ontologies are publicly available and are maintained at the Planteome GitHub site (https://github.com/Planteome) for sharing and tracking revisions and new requests.

2.4 NLP as a Service

Since the number of journal articles published each year has significantly increased in the past decades, information management services, such as those provided through libraries and content creators, are seeking new ways to improve the content accessibility. New text-mining methods are adopted to provide better search experiences for users by offering more search categories and rankings. Many researchers now gradually rely on search service providers, such as Google Scholar, more than using traditional interfaces provided by content managers to identify new relevant articles. Along with publications, there are also new concepts; data and experiment details need to be annotated and cross-referenced for better utilizations. There is an ongoing need of providing a centralized environment for comprehensive information and knowledge access.

Within the field of plant biology, there are several ongoing efforts on integrating comprehensive information from diverse sources and making them accessible through a web portal interface for targeted research communities. Arabidopsis Information Portal (Araport) is a project dedicated to *Arabidopsis* research (https://www.araport.org/) (Swarbreck et al., 2007; Krishnakumar et al., 2014). The project integrates information on *Arabidopsis* genes and their function annotations from various data sources and makes all information accessible through a web portal. Other examples include Gramene (www.gramene.org) and Planteome (Tello-Ruiz et al., 2017; Cooper L et al., 2017). The Gramene database is freely available for download and used as long as Gramene is cited as the source. This includes the tools available at Gramene including but not limited to RiceCyc, CMap Viewer, Gramene Mart, and the Genome Browser (Tello-Ruiz et al., 2017). A core-processing task common to above projects of information integration is the need of NLP and text mining. In practice, these needs are met through customized software development process due to the limited transferability described in the previous section.

With the increasing amount of new data and expanding vocabularies, a scalable computing service that can be adapted to different domains and use cases can directly facilitate this information integration processes. CI, which refers to large shared online research environments, has been increasingly used in open science research and enables breakthrough discovery in many domains, including academic libraries for accommodating data and analysis within their services and collections. *Science as a service* is a new approach that has emerged along with the data-driven science (Grossman et al., 2016). This approach is a step up from the previous infrastructure-as-a-service (IaaS) model, through which the physical and/or virtual resources are allocated to users upon request. The IaaS model has limited interactive analysis support and does not provide support of graphical user interface. It limits usability of the CI resources and distances itself from users in noncomputational fields. In science-as-a-service model, common analysis tasks can be abstracted and deployed as a service module. The service module has an easy to use interface for end users and is backed up by powerful remote computing resources for fast performance. Furthermore, analysis can often be facilitated with comprehensive user environment and visualization support. One such example is the CyVerse project. CyVerse provides life scientists with powerful computational infrastructure to handle huge datasets and complex analyses, thus enabling data-driven discovery. It has an extensible platform that provides data storage, bioinformatics tools, image analyses, cloud services, application programming interface (APIs), and more (Merchant et al. 2016; Goff et al. 2011). Through CyVerse APIs, hundreds of applications have been developed to simplify the process of running analysis with remote CI. However, providing NLP as a service using CI is rarely seen. Here, we are motivated to develop a framework for named entity recognition service that can later be adapted across different domains.

3 Framework Design and Implementation

We propose a pluggable and flexible framework that can ingest available articles while fronted by a web service readily available to users. The framework helps curators to identify new concepts of interest emerged

from latest publications. This framework can facilitate integration of different approaches of entity recognition and migration across different domain use cases. The proposed architecture enables us introducing successful innovations in the detection pipeline into this framework. Our approach closes the feedback loop from users through recording author's annotation inputs, which can be used to improve processing workflow for future iterations.

3.1 Motivation

Our work is directly motivated by the need of providing better access to new journal publications. However, existing entity extraction methods are not suitable for this problem. General named entity recognition methods often have technical challenges in detecting entities that lie in boundary, long-range semantic constructs, co-reference resolution, and numerous others. There are several additional factors causing underperformance of the entity recognition: 1) intrinsic complexity of biomedical entities' structure; 2) ever increasing new terminologies and concepts; and 3) lack of well-curated training datasets. Entities in biological journals can have complicated structure and naming schemas. Nested structure and abbreviations are also common in biological literature. So far accuracy has been improved largely by cross-referencing with carefully curated dictionaries and ontologies that often require significant efforts to maintain. At the same time, the definition of named entities in biology are constantly evolving and expanding with new relevant phrases that are meaningful in the context of the domain of the article (e.g., names of methodology, equipment, and drug). Owing to these reasons, although existing named entity recognition methods may detect a lot of existing biological entities, recognized entities are often different from what authors think important. This shortcoming is further confirmed with comparison results between our approach and ABNER, a program specifically designed for biological entity extraction (Section 4).

Furthermore, our work not only is motivated by recognized entities from new publications but also is used as a tool to discover and curate new vocabulary. The quality of curated training data has direct impact on the efficiency of model inferred from the learning process. However, curating high-quality training dataset to keep up with increasing volume of articles and concepts is another time-consuming process that requires substantial efforts from human experts. It is an ongoing research area explored jointly by the domain scientists and the NLP community in hope of yielding newer algorithms that are less dependent on human-curated sources.

The proposed framework addresses these limitations through an ensemble method approach. By integrating multiple models for detection, the system can detect more concepts embedded within the documents. The framework is designed to utilize use case-specific knowledge and rules to improve the detection. Domain-specific knowledge can be integrated into the framework as *Rules*. For the use case presented here, we have leveraged existing biological ontologies and publication formatting information. The framework also includes user interface to enable human expert validation and curation. All curation and correction made by experts are tracked and stored in a database for improving entity detections in the future.

3.2 Entity Detection Architecture Overview

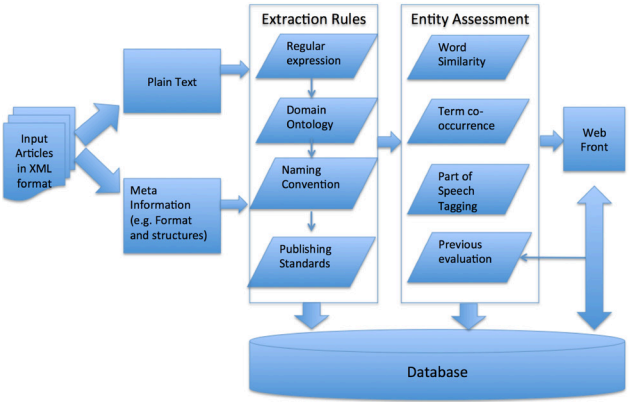


Figure 1. Overview of proposed journal processing framework

The processing workflow for identifying biological entities from journal articles is illustrated in Figure 1. There are three major steps: text extraction, entity candidate extraction, and candidate assessment. For this specific use case, the input is a structured document, which contains content and metadata of a journal article. The entity candidate extraction phases contain a number of implemented rules and models to infer potential phrases of interests. These candidates are to be further assessed for importance and validated through a graphical web user interface. All results, including entity inference, assessment, and human validation results, are stored in a relational database.

3.2.1 Text Extraction

The text extraction processes the input structured document tagged by JATS. During this step, the input document is processed into two parts: textual content of

journal article and the structure of the journal document. The textual information data part is a list of string representation of the body of text included in the journal articles. The structural data include metadata information presented at the input document, such as section mark and special formatting mark. Each string is a sentence or line from the original document. A mapping is maintained between the textual value and metadata information by their global positions in the original documents. This dual data structure allows for efficient text processing of the publication content while still being able to easily retrieve the metastructure around a particular set of words during the subsequent steps of processing. To separate the textual context and presentation format and process separately also enables the framework to be extended in the future for additional input text formats. The subsequent processing can independently utilize each part separately or jointly based on model requirements and availability. Therefore, unstructured text without structure information can also be processed in the same framework.

3.2.2 Entity Candidate Detection

A feature of the processing framework is to support ensemble methods for entity detection. The framework is designed to be able to utilize existing generic models. Domain-specific detection can be encoded as rule-based models for detection. The detection rules can be defined based on various heuristics and requirements such as publishing requirements, naming conventions, and domain ontologies. New rules can be added on demand over time. Currently, there are four types of rules implemented in the DIVE: regular expression rules, word dictionary, publishing convention, and ontology rules.

Each rule can be defined as a regular expression and used for matching the candidate word. The regular expression rules utilize common naming conventions to identify biological entities, such as gene name, protein name, molecule structures, and chemical compound. The word dictionary rule consists of a predefined list of words that should be included or excluded in the candidate lists. The publication content is searched against the list at run time first. Both regular expression rules and word list are created based on inputs from biologists in the research team. The publishing convention rules are used to identify words that are in a special format, such as in italic, or in a particular component of the publication, such as a figure legend. The enclosing tags of the candidates are used to define each rule. The publishing convention rules are created based on the suggestions from editors at ASPB. Additional rules can be added by specifying additional tag values or by using naming conventions to detect entities like species names. The ontology rules utilize five biological ontologies including gene ontology (Ashburner et al., 2000), plant ontology (Jaiswal et al., 2005), plant trait ontology (Arnaud et al., 2012), plant environment condition ontology (Jaiswal & Cooper, 2018), and ChEBI (Degtyarenko et al., 2007). Rules are also detailed in the code.

3.2.3 Entity Candidate Assessment

By applying the extraction rules listed earlier, a set of entity candidates can be detected from the input document. Some candidates might be detected by multiple rules. Different detection rules also have different accuracy. Ontology file and dictionary-based approaches have the highest certainty. Candidates only identified by other rules need further validation. We currently implemented two automatic validation mechanisms. One is based on the previously validated results; the other one is based on co-location with other confirmed entities. However, the primary method of validation is by domain expert evaluation through the web interface, which is detailed in Section 3.2.

3.2.4 Entity Management and Versioning

The results from entity extraction processing workflow are stored in a relational database and served as data store for the web application. There are two major sets of table in the database, the *Files* table and the *Entities* table.

The *Files* table represents a publisher-generated XML file for a manuscript containing metadata information annotating various parts of the manuscript text. It has the following fields:

{Filename, Title, Abstract, Externally pointing DOI link}

The *Entities* table represents entities discovered by our entity extraction algorithms using various methodologies mentioned earlier (e.g., ontology, regex, and keyword-based retrieval). The basic information of each entity is stored with following fields:

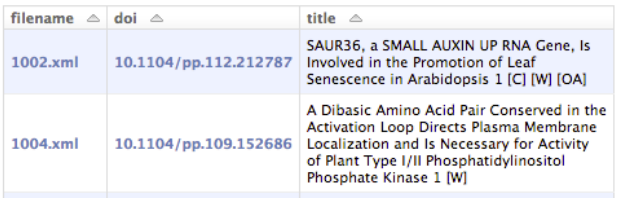
{Entity Name, Entity Type, XRef, Filename}

Entity name may consist of multiple words. An entity name can appear multiple times in the database if it exists

in different publications. XRef refers to a link to well-known comprehensive databases containing additional domain-specific information. There are also additional fields to track specific information about each entity, such as the species it is associated with and locations where it appears in the publication. Additionally, the database also records user revisions through following fields:

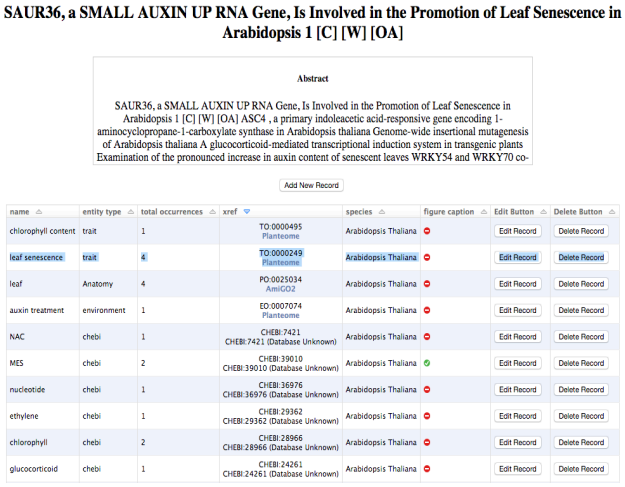
{Version, Timestamp, Change Type, IP}

We use version numbers to track changes made to an entity by algorithm or by the user. Timestamp and types of change of each update are recorded accordingly. The IP address is used to indicate who made the updates. Collectively, all of these help us track a chain of modify/delete events in the life of an extracted entity in the database. Such patterns can be retrieved and used as feedback to future algorithm iterations and learning. The database presently being used is SQLite. This is a flat file database that offers relational semantics. It was chosen because it makes our prototype a self-contained and easily deployable unit. This could be easily transitioned to any popular relational database such as PostgreSQL and MySQL for larger scale use cases.



filename	doi	title
1002.xml	10.1104/pp.112.212787	SAUR36, a SMALL AUXIN UP RNA Gene, Is Involved in the Promotion of Leaf Senescence in Arabidopsis 1 [C] [W] [OA]
1004.xml	10.1104/pp.109.152686	A Dibasic Amino Acid Pair Conserved in the Activation Loop Directs Plasma Membrane Localization and Is Necessary for Activity of Plant Type I/II Phosphatidylinositol Phosphate Kinase 1 [W]

Figure 2. Paginated view of collection



SAUR36, a SMALL AUXIN UP RNA Gene, Is Involved in the Promotion of Leaf Senescence in Arabidopsis 1 [C] [W] [OA]

Abstract

SAUR36, a SMALL AUXIN UP RNA Gene, Is Involved in the Promotion of Leaf Senescence in Arabidopsis 1 [C] [W] [OA] ASC4 , a primary indoleacetic acid-responsive gene encoding 1-aminocyclopropane-1-carboxylate synthase in Arabidopsis thaliana Genome-wide insertional mutagenesis of Arabidopsis thaliana A glucocorticoid-mediated transcriptional induction system in transgenic plants Examination of the pronounced increase in auxin content of senescent leaves WRKY54 and WRKY70 co-

Add New Record

name	entity type	total occurrences	xref	species	figure caption	Edit Button	Delete Button
chlorophyll content	trait	1	TO:0000495 Planteome	Arabidopsis Thaliana		Edit Record	Delete Record
leaf senescence	trait	4	TO:0000249 Planteome	Arabidopsis Thaliana		Edit Record	Delete Record
leaf	Anatomy	4	PO:0025034 AmiGO2	Arabidopsis Thaliana		Edit Record	Delete Record
auxin treatment	environment	1	EO:0007074 Planteome	Arabidopsis Thaliana		Edit Record	Delete Record
NAC	chebi	1	CHEBI:7421 CHEBI:7421 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record
MES	chebi	2	CHEBI:39010 CHEBI:39010 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record
nucleotide	chebi	1	CHEBI:36976 CHEBI:36976 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record
ethylene	chebi	1	CHEBI:29362 CHEBI:29362 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record
chlorophyll	chebi	2	CHEBI:28966 CHEBI:28966 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record
glucocorticoid	chebi	1	CHEBI:24261 CHEBI:24261 (Database Unknown)	Arabidopsis Thaliana		Edit Record	Delete Record

Figure 3. Interface for exploring entities in a publication

3.3 Web Interface Design

We chose Django (v 1.10) to implement the web front end in our prototype. Based on Python, the web front is easily programmable, extensible, and 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. 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*.

The first view (Figure 2) is a paginated list of all the articles – i.e., xml files. There are additional columns pointing to an external DOI reference to a copy of the article itself and the article title.

The next view (Figure 3) is reached by clicking on the file name (e.g., *1002.xml* in Figure 2). The web interface runs the backend code for entity extraction and presents the results to the user based on the schema described in the previous section. The layout consists of the title at the very top and a scrollable text box that contains the abstract extracted from the manuscript. As these are provided, the user has some context knowledge of the actual manuscript accompanying the presented meta information that is extracted or generated by the backend algorithms. The XRef column also gives a link pointing to the Planteome ontology database.

Staying with the same example, the user control button for editing a record directs to the edit page where we see the layout of title and abstract at the top (Figure 4). These are again provided for context. Among other additions, we also see the editable fields of this record where a user may correct or enter new values. A dynamic search box can be used to search for 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.

3.4 Service Integration and Implementations

DIVE has been integrated into the publication pipeline of two plant biology journals, namely, *The Plant Cell*

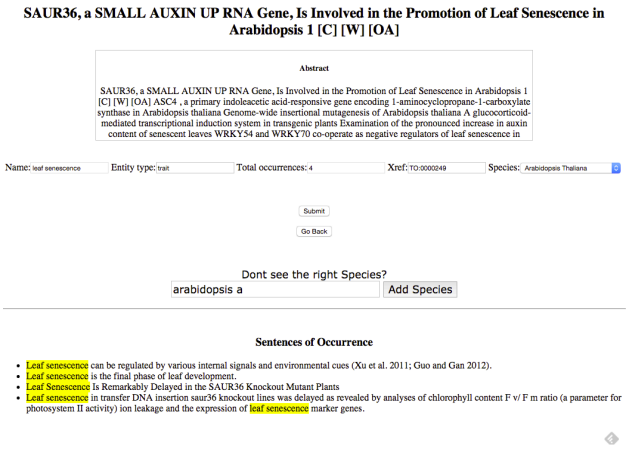


Figure 4. Interface for showing/editing entity details

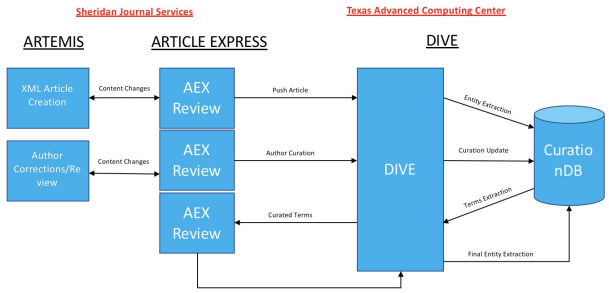


Figure 5. Integration architecture with publication pipeline

(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 shown in Figure 5. To enable this integration, DIVE functionality was exposed to the publication software as a web service with two endpoints.

- **Article Endpoint**
  This endpoint receives two HTTP POST requests related to the article.
  - Article Push request: This request is used 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 meta information. The authors may either verify its accuracy or do curation actions of edits, additions, or 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 future entity detection for DIVE.

3.5 Additional Features

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 among other users with overlapping domain expertise and interest. An example of the search interface is shown in Figure 6. 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.

We are working on supporting further analysis on the extracted results within our framework. One example is to analyze the associations among extracted entities. Figure 7 shows top 20 inference rules based on all ontology terms extracted from the collection. Each label indicates a frequent item set found in the collection. The directional arrow indicates an inference on co-occurrence between two item sets. The shade of the directional arrow indicates the confidence level of the rule. Such visual representations of inferred association between diverse entity types could tremendously aid a researcher in forming insights. This also has potential to be a similarity metric between articles that could help editors gage the novelty of a new article submission.

Figure 6. DIVE search interface

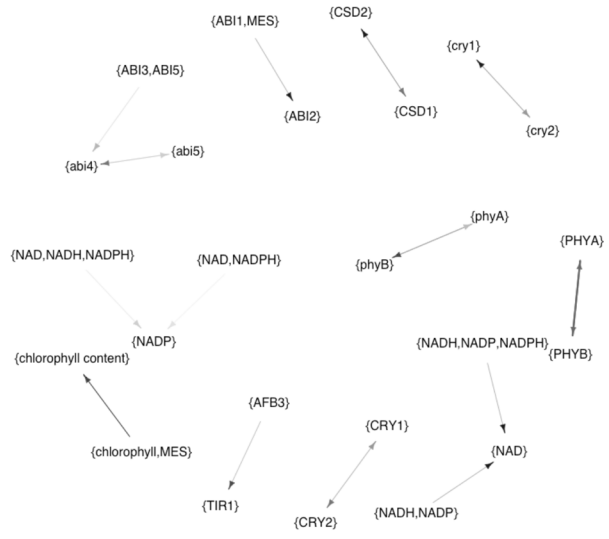


Figure 7. Top 20 inference rules from association analysis.

4 Usage Statistics and Evaluations

The DIVE web service integration has been deployed into production since April 2018 to process submissions to *The Plant Cell* and *Plant Physiology* journals. For this paper, we have collected 443 articles curated by DIVE before February 2019. Among these articles, 141 were submitted to *The Plant Cell* journal and 302 were submitted to *Plant Physiology* journal.

Table 1
Summary of Total Number of Entities Found, Presented, and Curated by Authors.

Total entities retrieved by DIVE for all articles	22,747
Total entities displayed by DIVE to users for all articles	8358
Curation: total addition by authors	315
Curation: total edits by authors	1517
Curation: total deletion by authors	4218

4.1 Usage Statistics

In total, DIVE has learnt 22,747 entities from these journal articles. These include 11,964 proteins, 6611 genes, and 212 plant anatomy entities. On average, there are about 51 entities found per article. Although we initially presented all entities to authors, the number of entities has gradually changed to 10 in order to reduce curation workload for authors during this period. If there are more than 10 entities reported by DIVE, these entities are ranked based on both their appearance frequencies in the article and results of DIVE prediction. On the other hand, authors are required to curate up to 10 entities. However, some authors opted to curate more entities. We have tracked both entities presented to and actions from authors. The results are detailed in Table 1. Note that actions are tracked per article and have overlap with each other. For example, an author may delete an entity to add a new one instead of editing the entity. The author may perform add, edit, and delete actions over the same entity. So the number of curation action does not correspond to number of entities missing precisely.

Figure 8 (left) shows the monthly distribution of submissions to *The Plant Cell* and *Plant Physiology* from the ASPB publishing pipeline. Figure 8 (right) shows the aggregate ratio of number of articles from each journal submitted to the DIVE system in the same period. The number of submissions to *Plant Physiology* is roughly twice the number of submissions submitted to *The Plant Cell*.

Figure 9 shows the percentage of most common entity types from the DIVE corpus collection of ASPB journal articles. As seen, proteins and genes are the two most identified named entities from the article. The two types account for 83% of total entities identified. About 7% entities are verified by ontologies from Gramene and Araport. About 5% entities are identified as a chemical compound.

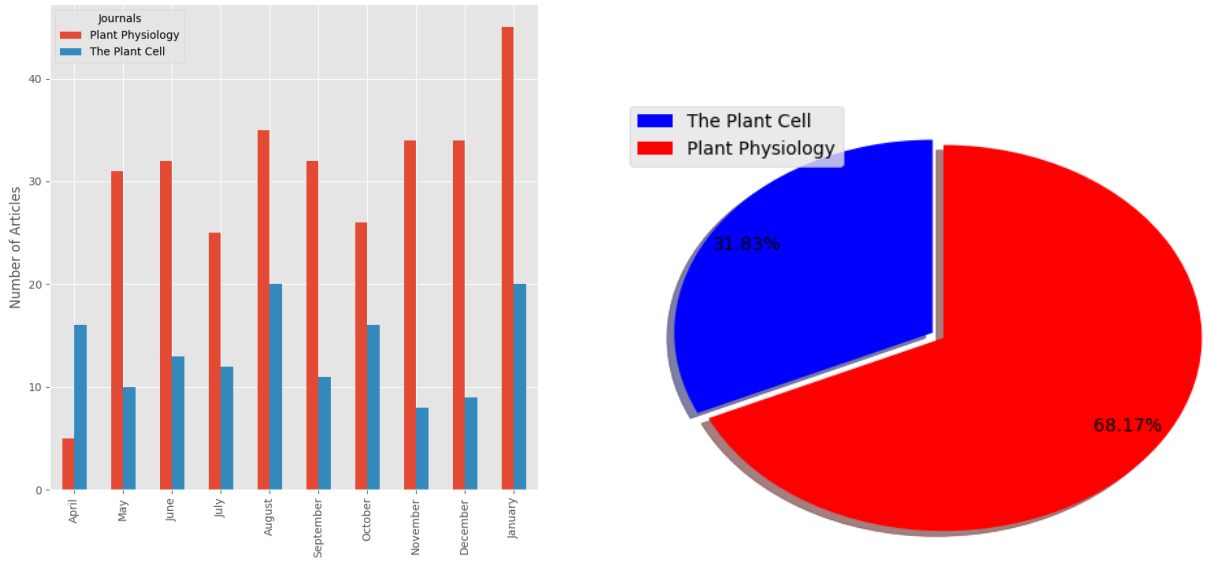


Figure 8. Distribution of DIVE articles by month (left) and by aggregation (right) for two journals published by ASPB.

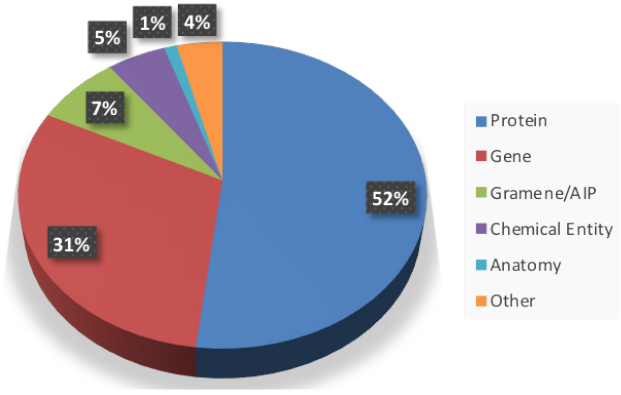


Figure 9. Distribution of recognized entities by type

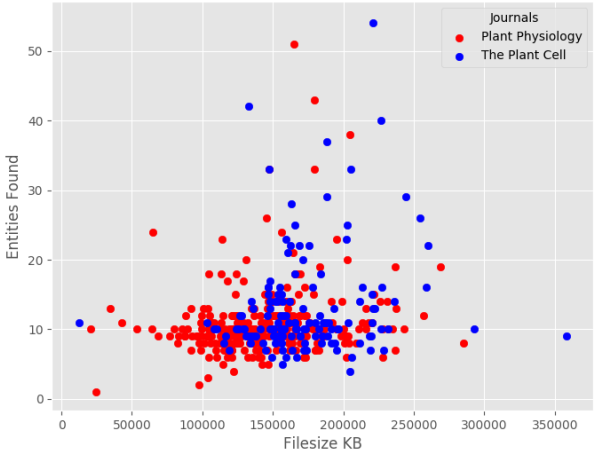


Figure 10. Number of entities learned with article file size

Figure 10 shows the relation between entities learned with respect to the article size (linearly correlated with its number of words) for both ASPB journals for the present corpus of articles. As expected, a general trend is that more entities are identified from longer submission. The figure also shows that *The Plant Cell* is usually longer than *Plant Physiology*. However, the number of entities from *The Plant Cell* seems fewer than that from *Plant Physiology* of a similar length.

4.2 Evaluation

To assess the effectiveness of DIVE results, we compared results with ABNER results for the same dataset. ABNER is a tool that automatically detects and tags biological entities such as proteins and genes in the natural language text (Settles, 2005). It is implemented in Java and also features a GUI interface where users can manually input sections of text for annotation. It uses CRFs (Lafferty et al., 2001), a supervised machine learning algorithm, that uses a probabilistic graphical model to enable it to detect and label relevant tokens. It comes packaged with two models, trained on the BioCreative (hereafter referred as ABNER_BIO) (Yeh et al., 2005) and NLPBA corpora (hereafter referred as ABNER_NLP) (Kim et al., 2004). A limitation of this method is that it requires a large amount of human annotated data to train the model, and these are relatively small corpora containing short stanzas. For this comparison, we have collected all 7095 entities approved by authors from the 443 articles and used them as the

Table 2
Results of Entity Recognition Against Author Curation as Ground Truth

	Total number of Entities	Total Entities in Ground Truth	Total Recall	Total Precision	Average Recall	Average Precision
DIVE	8358	5123	0.7221	0.6129	0.7993	0.6962
ABNER_BIO	136354	2814	0.3966	0.0206	0.4362	0.0208
ABNER_NLP	127443	3029	0.4269	0.0237	0.4722	0.0251
Ground Truth	7095	-	-	-		

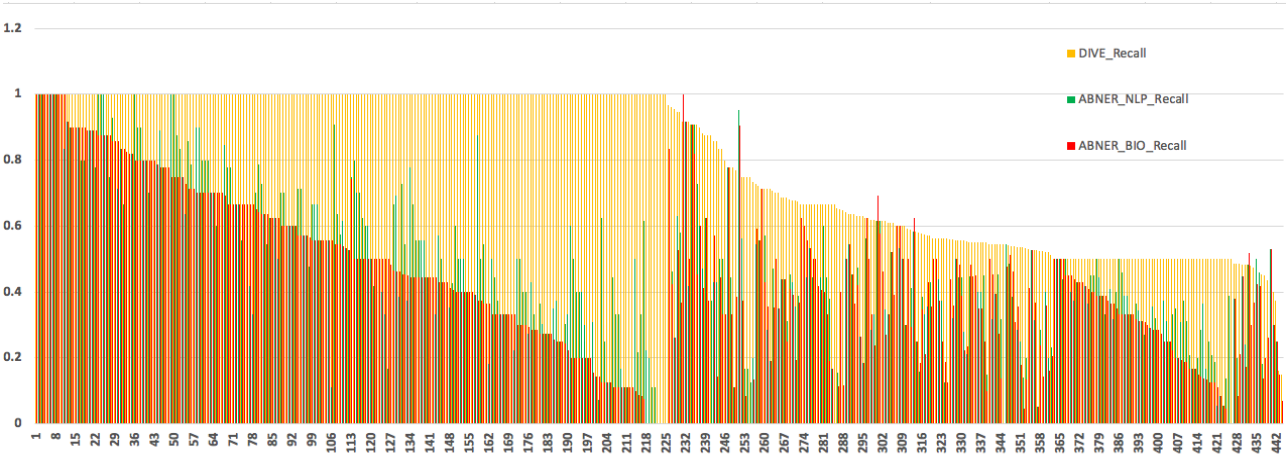


Figure 11. Recall comparison per article between DIVE, ABNER_NLP, and ABNER_BIO.

ground truth for these articles. We then extracted the text from all 443 articles as input for the ABNER program. We then checked how many entities are reported by DIVE, ABNER_NLP, and ABNER_BIO. We calculated recall and precision as follows:

$$\text{Recall} = \frac{\textit{Number of entities in Ground Truth that has been reported by each program}}{\textit{Number of entities in gold standard}}$$

$$\textit{Precision} = \frac{\textit{Number of entities in Ground Truth that has been reported by each program}}{\textit{Number of entities reported by each program}}$$

The results are detailed in Table 2. The "*total number of entities*" column shows the total number of entities in the ground truth data set, reported to author by DIVE, recognized by the ABNER_BIO model and the ABNER_NLP model. The true positives of each program are shown in the "*total entities in ground truth*" column. The "*total recall*" and "*total precision*" columns show recall and precision values over the entire dataset. The "*average recall*" and "*average precision*" are calculated averages per article. Both ABNER models found far fewer entities in the ground truth dataset than DIVE in total and on average. The results demonstrate that DIVE is more effective than ABNER to identify potential entities of importance to authors. Since the ABNER program does not offer a way to sort entities, we used total number of entities found by ABNER, which resulted in a very low precision score. This also reinforces our motivations that existing entities tools cannot be used for solving this problem directly. Additional features and functionalities must be developed to be used in practice.

Recall scores for all articles using three models are shown in Figure 11. Three columns are shown per article corresponding to recall results from DIVE (yellow), ABNER_NLP (green), and ABNER_BIO (red) models. To increase graph readability, we sorted 443 articles in the decreasing order of DIVE recall value (yellow), ABNER_BIO recall value (red), and then ABNER_NLP recall value (green). Figure 11 shows that there are only eight articles where ABNER outperforms DIVE. DIVE identified all entities in the ground truth dataset for 224 articles (50.5%) in contrast to only 12 articles (2.7%) using the ABNER_BIO model. DIVE identified at least half of the entities in the ground truth dataset for 425 articles (95.9%) in contrast to 173 (39.1%) articles using the ABNER_BIO model.

5 Discussion and Conclusions

In this paper, we present an application to extract domain-specific information from journal articles, identify additional information content, and deploy service to enrich the digital publication during article production. The application integrates multiple NLP methods for entity recognition and enables human curation to close the feedback loop. Based on practical usage statistics, our application outperforms the existing algorithm in entity recognition. Furthermore, our application is not just a new algorithm for entity detection. It can be adapted for curating new terminologies and vocabularies. The application is still in development, and we are gathering feedback from domain researchers and publishing professionals. 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, we see about 10 curation actions per article in our corpus. Based on their feedback, we are also working on improving the search features to incorporate full-text search and 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.

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 case for corpora in other domains as well (e.g., aerospace engineering) and will continue to expand in this area. We aim to expand DIVE for additional use case scenarios from many scientific domains to help scientists and researchers at large making sense of their large document corpora.

Acknowledgements: This research is partially supported by CyVerse (NSF awards DBI-0735191 and DBI-1265383) and Gramene, a comparative plant genomics database (NSF award IOS-1127112).

References

Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R. T., Matteis, L.,…. McLaren, G. (2012, October). *Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes.* In *KEOD* (pp. 220-225).

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Harris, M. A. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25–29.

Bhagavatula M., GSK S, Varma V. (2012, November). Named entity recognition an aid to improve multilingual entity filling in language-independent approach. *Proceedings of the First Workshop on Information and Knowledge Management for Developing Region* (pp. 3-10), ACM.

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD),1*(1), 1-36.

Bilgic, M., Licamele, L., Getoor, L., & Shneiderman, B. (2006, October). D-dupe: An interactive tool for entity resolution in social networks. In *2006 IEEE Symposium on Visual Analytics Science and Technology* (pp. 43-50). Baltimore, MD, USA.

Björk, B. C. (2017). Scholarly journal publishing in transition-from restricted to open access. *Electronic Markets, 27*(2), 101–109.

Campos, D., Matos, S., & Oliveira, J. L. (2013). Gimli: Open source and high-performance biomedical name recognition. *BMC Bioinformatics, 14*(1), 54.

Chang, C.-C. (2008). The value of knowledge created by individual scientist and research groups. *Journal of Scholarly Publishing, 39*(3), 274–293.

Chen, D., & Manning, C. (2014, October). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740-750). Doha, Qatar: Association for Computational Linguistics.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016, August). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th workshop on biomedical natural language processing* (pp. 166-174). Berlin, Germany: Association for Computational Linguistics.

Christen, P. (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. New York, USA: Springer Science & Business Media.

Cooper, L., & Jaiswal, P. (2016). The plant ontology: a tool for plant genomics. In Edwards D (Eds.), *Plant Bioinformatics* (pp. 89-114). New York, NY: Humana Press.

Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Jaiswal, P. (2017). The Planteome database: An integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research, 46*(D1), D1168–D1180.

De Boer, V., Van Someren, M., & Wielinga, B. J. (2006, June). Relation instantiation for ontology population using the web. In *Annual Conference on Artificial Intelligence* (pp. 202-213). Berlin, Heidelberg: Springer.

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A. Ashburner, M. (2007). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research,36*(Database issue), D344–D350. PMID:17932057

Dernoncourt, F., Lee, J.Y. & Szolovits, P. (2017). NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004, May). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (Vol. 2, p.

1), Lisbon, Portugal: European Language Resources Association (ELRA).

Ek, T., Kirkegaard, C., Jonsson, H., & Nugues, P. (2011). Named entity recognition for short text messages. *Procedia: Social and Behavioral Sciences, 27*, 178–187.

Ekbal, A., & Saha, S. (2011). A multi-objective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies. *Expert Systems with Applications, 38*(12), 14760–14772.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16, IEEE.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence, 165*(1), 91–134.

Getoor, L., & Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Explorations, 7*(2), 3–12.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers of Plant Science, 2*, 34.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review,29*, 21–43.

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics* (Vol. 1).

Grossman, R. L. , Heath, A. , Murphy, M. , Patterson, M. , & Wells, W. . (2016). A case for data commons: towards data science as a service. *Computing in Science & Engineering, 18*(5), 10-20. Saha, S. K., Narayan, S., Sarkar, S., & Mitra, P. (2010). A composite kernel for named entity recognition. *Pattern Recognition Letters, 31*(12), 1591–1597.

Guanming, Z., Chuang, Z., Bo, X., & Zhiqing, L. (2009, March). CRFs-based Chinese named entity recognition with improved tag set. In *2009 WRI World Congress on Computer Science and Information Engineering* (Vol. 5, pp. 519-522). Los Angeles, CA, USA.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics (Oxford, England), 33*(14), i37–i48.

Herzog, Thomas N., Scheuren, Fritz J., & Winkler., William E. (2007). *Data quality and record linkage techniques*. New York: Springer Science & Business Media.

Huang, C.-C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics, 17*(1), 132–144.

Huh, S. (2014). Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language. *Science Editing, 1*(2), 99–104.

Huh, S., Choi, T. J., & Kim, S.-H. (2014). Using Journal Article Tag Suite extensible markup language for scholarly journal articles written in Korean. *Science Editing, 1*(1), 19–23.

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. *International Journal of Genomics,6*(7-8), 388–397.

Jaiswal, P. & Cooper, L. (2018, February 18). Plant Environment Condition Ontology. Retrieved from http://bioportal.bioontology.org/ontologies/PECO#

Ju, Z., Wang, J., & Zhu, F. (2011, May). Named entity recognition from biomedical text using SVM. In *2011 5th international conference on bioinformatics and biomedical engineering* (pp. 1-4). Wuhan, China.

Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004, August). Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications* (pp. 70-75). Geneva, Switzerland.

Kiperwasser, E., & Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics, 4,* 313- 327.

Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2), 484-493.

Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M.,Town, C. D. (2014). Araport: The Arabidopsis information portal. *Nucleic Acids Research, 43*(Database issue, no. D1), D1003–D1009.

Lafferty, J., McCallum, A. & Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, L., Zhou, R., & Huang, D. (2009). Two-phase biomedical named entity recognition using CRFs. *Computational Biology and Chemistry, 33*(4), 334–338.

Liu, X., & Zhou, M. (2013). Two-stage NER for tweets with clustering. *Information Processing & Management, 49*(1), 264–273.

Majumder, M., Barman, U., Prasad, R., Saurabh, K., & Saha, S. K. (2012). A novel technique for name identification from homeopathy diagnosis discussion forum. *Procedia Technology, 6,* 379-386.

McCallum, A., & Li, W. (2003, May). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 188-191). Edmonton, Canada.

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biology, 14*(1), e1002342.

Mihăilă, C., & Ananiadou, S. (2014). Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online, 13*(2), S1.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management, 2*(1), 1–87.

Pasca, M., Lin, D., Bigham, J., Lifchits, A., & Jain, A. (2006, July). Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In *Proceeding AAAI'06 proceedings of the 21st national conference on Artificial intelligence* (Vol. 6, pp. 1400-1405), Boston, Massachusetts: AAAI Press.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics, 8*(1), 50.

Saha, S. K., Sarkar, S., & Mitra, P. (2009). Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of Biomedical Informatics, 42*(5), 905–911.

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proceedings of CoNLL-2003 (pp. 142-147), Edmonton, Canada.

Santos, D., Seco, N., Cardoso, N., & Vilela, R. (2006, May). Harem: An advanced ner evaluation contest for portuguese. In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed), *Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy.

Settles, B. (2005). ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics (Oxford, England), 21*(14), 3191–3192.

Shaalan, K. (2010). Rule-based approach in Arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT), 3*(3), 11-19.

Shen, D., Zhang, J., Zhou, G., Su, J., & Tan, C. L. (2003, July). Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 49-56). Sapporo, Japan.

Song, H. J., Park, S. B., & Park, S. Y. (2009, June). An automatic ontology population with a machine learning technique from semi-structured documents. Paper presented at the *2009 International Conference on Information and Automation* (pp. 534-539). Zhuhai, China.

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning,* 4(4), 267–373.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., & Radenbaugh, A. (2007). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research, 36*(suppl_1), D1009-D1014.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., & Wilbur, W. J. (2005). GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics, 6*(1), S3.

Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A.,...Ware, D. (2017). Gramene 2018: Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research,46*(D1), D1181–D1189.

Tenopir, C., & King, D. W. (2014). The growth of journals publishing. In Cope D.& Philips A. (Eds.), *The future of the academic journal* (pp. 159-178). New York, USA: Chandos Publishing.

Thenmalar, S., Balaji, J., & Geetha, T. V. (2015). Semi-supervised bootstrapping approach for named entity recognition. *arXiv preprint arXiv:1511.06833*.

Tsai, R. T. H., Sung, C. L., Dai, H. J., Hung, H. C., Sung, T. Y., & Hsu, W. L. (2006, December). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. In *BMC bioinformatics* (Vol. 7, No. 5, p. S11). BioMed Central.

Wang, Y., Yu, Z., Chen, L., Chen, Y., Liu, Y., Hu, X., & Jiang, Y. (2014). Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. *Journal of Biomedical Informatics, 47*, 91–104.

Ware, M., & Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing.

Yeh, A., Morgan, A., Colosimo, M., & Hirschman, L. (2005). BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics, 6*(1), S2.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics, 46*(6), 1088–1098.

Zhu, F., & Shen, B. (2012). Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. *PLoS One, 7*(6), e39230.