Semantic Text Mining for Lignocellulose Research

Marie-Jean Meurs
Department of Computer Science and Software Engineering
mjmeurs@encs.concordia.ca

Caitlin Murphy
Centre for Structural and Functional Genomics and Department of Biology
cmurphy@gene.concordia.ca

Ingo Morgenstern
Centre for Structural and Functional Genomics and Department of Biology
imorgenstern@gene.concordia.ca

Nona Naderi
Department of Computer Science and Software Engineering
n_nad@encs.concordia.ca

Greg Butler
Department of Computer Science and Software Engineering
gregb@encs.concordia.ca

Justin Powlowski
Centre for Structural and Functional Genomics and Department of Chemistry and Biochemistry
powlow@alcor.concordia.ca

Adrian Tsang
Centre for Structural and Functional Genomics and Department of Biology
tsang@gene.concordia.ca

René Witte*
Department of Computer Science and Software Engineering
rwitte@cse.concordia.ca

Concordia University
Montréal, QC, Canada

ABSTRACT

Semantic technologies, including natural language processing (NLP), ontologies, semantic web services and web-based collaboration tools, promise to support users in dealing with complex data, thereby facilitating knowledge-intensive tasks. An ongoing challenge is to select the appropriate technologies and combine them in a coherent system that brings measurable improvements to the users. We present our ongoing development of a semantic infrastructure in support of genomics-based lignocellulose research. Part of this effort is the automated curation of knowledge from information on enzymes from fungi that is available in the literature and genome resources. Fungi naturally break down lignocellulose, hence the identification and characterization of the enzymes that they use in lignocellulose hydrolysis is an important part in research and development of biomass-derived products and fuels. Working close to the biology researchers who manually curate the existing literature, we developed ontological NLP pipelines integrated in a Web-based interface to help them in two main tasks: mining the literature for relevant information, and at the same time providing rich and semantically linked information.

*corresponding author

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis—*text mining, ontology*; J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and Genetics*

General Terms

Experimentation

Keywords

Text Mining, Semantic Computing, Lignocellulose Research

1. INTRODUCTION

Since the early decades of the 20th century, when the internal combustion engine rapidly replaced the steam engine, transport is almost totally dependent on fossil fuels. As the amount of available petroleum decreases, producing sustainable liquid fuels with low environmental impact is one of the major technological challenges the world is facing today. Industrialized and developing countries consider *biofuels*, fuels produced from biomass, as a promising alternative to fossil fuels.

There are many advantages of using biofuels in terms of economic, environmental and energy security impacts [8]: easily available from biomass sources, biofuels can be sustainable and contribute to reduce the carbon dioxide emissions. Most of the biofuels used today are produced from the fermentation of corn starch which requires substantial input of water, fertilizer and energy. According to the United Nations Environment Programme [6], the global use of biofuels will nearly double during the next ten years. Hence, improving efficiency and sustainability of the biofuels production is of great interest. Underutilized agricultural and forestry residues, such as agricultural waste, wood chips from pulp

Table 1: Semantic entities, applicable level (sentence, *S* or word(s), *W*), definitions and examples

Semantic Entity	Level	Definition	Example
ActivityAssayConditions	*S*	conditions at which the activity assay is carried out	disodium hydrogen phosphate, citric acid, pH 4.0, 37℃
Assay	*W*	name of the experimental assay	Dinitrosalicylic Acid Method (Somogyi-Nelson)
Enzyme	*W*	enzyme name	alpha-galactosidase
Gene	*W*	gene name	mel36F
Glycosylation	*S*	enzymatic process attaching glycans to organic molecules	N-glycosylation
Host	*W*	organism used to produce the recombinant protein	Escherichia coli
KineticAssayConditions	*S*	buffer, pH, temp. for the kinetic parameters determination	0.1M (disodium hydrogen phosphate, citric acid), pH 4.0, 37℃
Organism	*W*	organism name	Gibberella sp.
pH	*S*	pH mentions	The enzyme retained greater than 90% of its original activity between pH 2.0 and 7.0 at room temperature for 3h.
ProductAnalysis	*S*	products formed from the enzyme reaction and identification method	HPLC, glucose, galactose
SpecificActivity	*S*	specific activity of the enzyme on the substrate	11.9U/mg
Strain	*W*	strain name	F75
Substrate	*W*	substrate name	stachyose
SubstrateSpecificity	*S*	substrate specificity mentions	The Endogluccanase from Pyrococcus furiosus had highest activity on cellopentaose.
Temperature	*S*	temperature mentions	The enzyme stability at different pH values was measured by the residual activity after the enzyme was incubated at 25℃ for 3h.

and paper production and all the "green" garbage, are composed of lignocellulose, which is the most abundant organic material on earth.

The sustainable conversion of lignocellulose into fermentable sugars for biofuel production requires the use of biological catalysts, called enzymes. Commercial biomass-degrading enzymes that are currently available are not efficient in lignocellulose degradation. Therefore, in the current race for replacing petroleum based fuels with renewable biofuels, discovering efficient enzymes for the cellulose degradation is a key challenge. In this context, researchers who aim to identify, analyze and develop these specific enzymes need to extract and interpret valuable and relevant knowledge from the huge amount of documents that are available in multiple, ever-growing repositories.

The largest knowledge source available to biological researchers is the PubMed bibliographic database [19], provided by the US National Center of Biotechnology Information, which contains more than 19 million citations from more than 21000 life science journals. PubMed is linked to other databases, like *Entrez Genome*, which provides access to genomic sequences, and *BRENDA, The Comprehensive Enzyme Information System* [20], which is the main collection of enzyme functional data available to the scientific community. A biology researcher querying PubMed using keywords collects an often long list of potentially relevant papers. The way to analyze this collection is reading all the abstracts and sometimes the full text papers: this task is time consuming and significant knowledge can be easily missed.

The work-in-progress we present in this paper focuses on the automatic extraction of knowledge from the massive amount of information on fungal enzymes available from the literature. In our approach, NLP pipelines brokered through web services support the extraction of relevant mentions and their enrichment with additional features.

This paper is organized as follows. The next section introduces related work, followed by background information in Section 3. Section 4 describes the architecture of the proposed system and presents the implementation of some of the components which support the curation through NLP Web services. Finally, Section 5 presents the corpus we are building, the annotation process and reports on the preliminary results.

2. RELATED WORK

To address the above mentioned challenges, NLP and Semantic Web approaches are increasingly adopted in biomedical research [1, 2, 21]. During the last decade, several systems combining text mining and semantic processing have been developed to help life sciences researchers in extracting knowledge from the literature:

Textpresso [14] enables the user to search for categories of biological concepts and classes relating two objects and/or keywords within an entire literature set;

GoPubMed [9] supports the arrangement of the abstracts returned from a PubMed query;

iHOP [12] converts the information in PubMed into one navigable resource by using genes and proteins as hyperlinks between sentences and abstracts;

BioRAT [3] extracts biological information from full-length papers;

Bio-Jigsaw [11] is a visual analytics system highlighting connections between biological entities or concepts grounded in the biomedical literature;

MutationMiner [24], based on a GATE pipeline [5, 7], automates the extraction of mutations and textual annotations

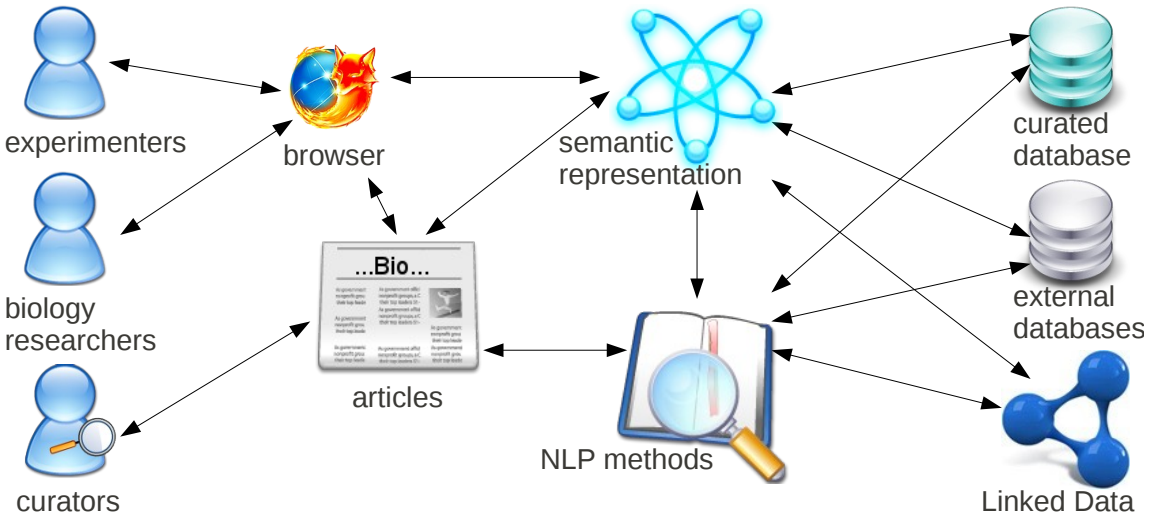

Figure 1: Integrating Semantic Support in Curation, Analysis, and Retrieval

describing the impacts of mutations on protein properties from full text scientific literature;

and Reflect [18] is a Firefox plugin which tags gene, protein and small molecule names in any web page.

3. BACKGROUND

Before we describe our overall architecture and the text mining pipelines, we briefly introduce the user groups involved, the semantic entities we analyse and the resources we use.

3.1 User Groups

The identification and the development of effective fungal enzyme cocktails are key elements of the biorefinery industry. In this context, the manual curation of fungal genes encoding lignocellulose-active enzymes provides the thorough knowledge necessary to facilitate research and experiments. Researchers involved in this curation are building sharable resources, usually by filling dedicated databases containing the extracted knowledge from the curated literature.

The users of our system are filling and using the mycoCLAP database[1] [15], which is a searchable database of fungal genes encoding lignocellulose-active proteins that have been biochemically characterized. The *curators* are therefore the first user group of our system. The *biology researchers* who make decision about the experiments to conduct and the *experimenters* executing them represent two further user groups. They are mainly interested in the ability of combining multiple semantic queries to the curated data, thereby semantically integrating the various knowledge resources.

3.2 Semantic Entities

The system we are developing has to support the manual curation process; therefore, the semantic entities have been defined by the curators according to the information they need to store in the mycoCLAP database.

Entities include information that are of particular interest for the researchers, such as organisms, enzymes, assays, genes, kinetic properties, substrates, and environmental conditions. The list of the semantic entities along with the level they apply (sentence or word level), their definition and an instance example is provided in Table 1.

About half of these entities are detected at the word level (e.g., enzyme or organism names) and the other half consists of contextual properties captured at the sentence level (e.g., pH and temperature contexts). The entity set was built in the perspective of providing instances of the ontological representation of the domain knowledge. The enzyme names are sought as well as the names of their source organisms with strains. The enzymes involved in the lignocellulose degradation have specific biochemical properties, such as optimal temperature and pH, temperature and pH stability, specific activity, substrate specificities as well as kinetic parameters. These experimentally determined properties describe the enzyme's function and nature. Their mentions are captured from the literature along with the laboratory methods (assay) used and the experimental conditions (activity and kinetic assay conditions). In addition to these properties, the extraction of mentions describing the enzymatic process (glycosylation) and the products formed (product analysis) is performed to finalize the knowledge of the reaction.

3.3 Semantic Resources

In terms of knowledge sources, the system relies on external and internal resources and ontologies. The *Taxonomy database*[2] [10] from NCBI is used for initializing the NLP resources supporting organism recognition. BRENDA[3] [20] provides the enzyme knowledge along with SwissProt/UniProtKB[4] [22]. References to the original sources are integrated into the curated data, which allows us to automatically create links using standard Linked Data techniques: e.g., links from an organism mention in a research paper to its corresponding entry in the NCBI Taxonomy database or from an enzyme name to its EC number in BRENDA.

[1]mycoCLAP: http://cubique.fungalgenomics.ca/mycoCLAP/

[2]NCBI Taxonomy: http://www.ncbi.nlm.nih.gov/Taxonomy/

[3]BRENDA: http://www.brenda-enzymes.org

[4]UniProtKB: http://www.uniprot.org/

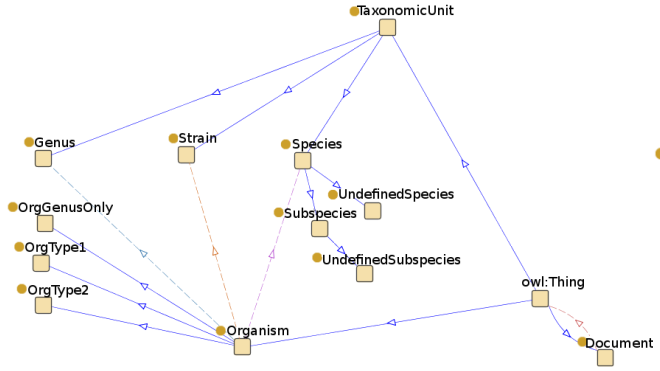

Figure 2: Organism ontology

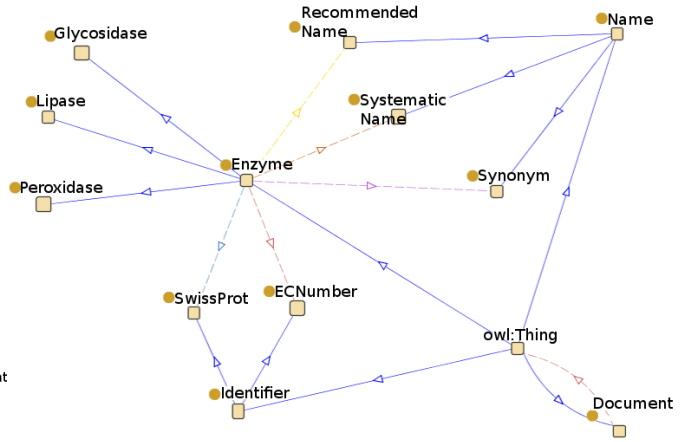

Figure 3: Enzyme ontology

4. SYSTEM DESIGN

In this section, we provide an overview of our system architecture, the semantic resources we deployed, and the text mining pipelines we developed.

4.1 System Architecture

With the different user groups and their diverging requirements, as well as the existing and continuously updated project infrastructure, we needed to find solutions for incrementally adding semantic support without disrupting day-to-day work. Our solution deploys a loosely-coupled, service-oriented architecture that provides semantic services through existing and new clients.

To connect the individual services and their results, we rely on standard semantic data formats, like OWL and RDF, which provide both loose coupling and semantic integration, as new data can be browsed and queried as soon as it is added to the framework (Figure 1). The use of the Semantic Assistants architecture [23] allows us to provide semantic analysis services directly within desktop applications, by leveraging standard SOAP web services and OWL service descriptions.

4.2 Ontologies

To facilitate semantic discovery, linking and querying the domain concepts across literature and databases, the entities are modeled in OWL ontologies, which are automatically populated from documents.

The system presented in this article makes use of two ontologies. Figure 2 shows the main entities in the organism ontology and Figure 3 depicts our custom built enzyme ontology, representing a subset of BRENDA's ontology. The ontologies are used both during the text mining process and for querying the extracted information [24].

4.3 Text Mining Pipelines

Our text mining pipelines are based on the *General Architecture for Text Engineering* (GATE) [7]. All documents first undergo basic preprocessing steps using off-the-shelf GATE components. Custom pipelines then extract the semantic entities mentioned above and populate the OWL ontologies using the OwlExporter [25] component. The same pipeline can be run for automatic (batch) ontology population, embedded in Teamware (described below) for manual annotation, or brokered to desktop clients through Web services for literature mining and database curation. The general workflow of the pipeline is depicted in Figure 4.

4.3.1 Preprocessing

The processing resources (PRs) composing the first part of the system pipeline are generic and independent from the domain. Some of these resources are based on standard components shipped with the GATE distribution. In particular, the JAPE language allows to generate finite-state language transducers that are processing annotation graphs over documents. After initializing the document, the *LigatureFinder PR* finds and replaces all ligatures, like *fi*, *ff* or *fl*, with their individual characters, thereby facilitating gazetteer-based analysis. The next PR is the *ANNIE English Tokenizer*, which splits the text into very simple tokens, such as numbers, punctuation characters and words of different types. Finally, the *ANNIE Sentence Splitter* segments the text into sentences by means of a cascade of finite-state transducers and the Hepple part-of-speech tagger that is included with GATE adds POS tags to each token.

4.3.2 Organism Recognition

The organism tagging and extraction relies on the OrganismTagger system.[5] The OrganismTagger is a hybrid rule-based/machine-learning system that extracts organism mentions from the biomedical literature, normalizes them to their scientific name, and provides grounding to the NCBI Taxonomy database [16].

The OrganismTagger also comes in form of GATE pipeline, which can be easily integrated into our system. It integrates the NCBI Taxonomy database, which is automatically transformed into NLP resources, thereby ensuring the system stays up-to-date with the NCBI database. Additionally, the organism ontology (Figure 2) formally describes the linguistic structure of organism entities at different levels of the taxonomic hierarchy [16]. The OrganismTagger pipeline provides the flexibility of annotating the species of particular interest

[5]The open-source OrganismTagger system, available at http://www.semanticsoftware.info/organism-tagger

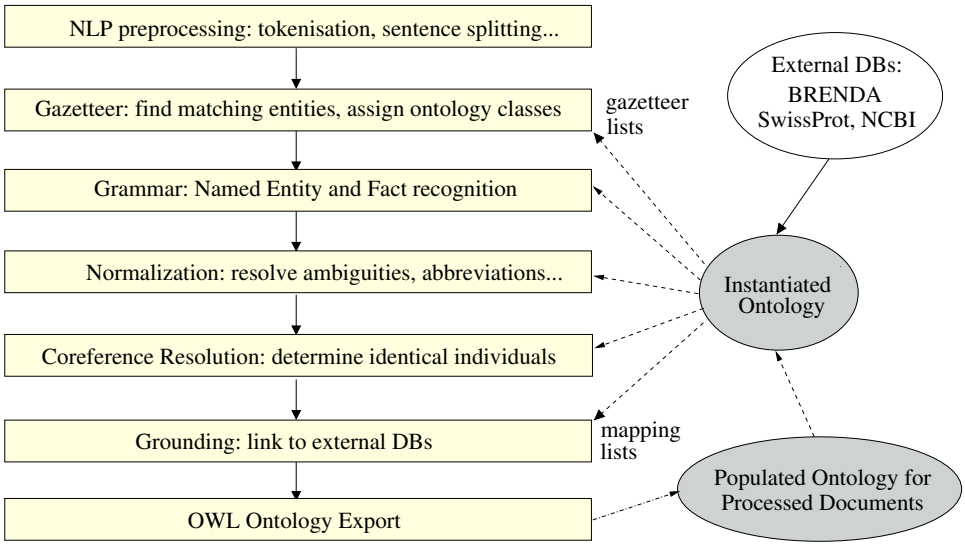

Figure 4: Natural Language Processing Workflow

to bio-engineers on different corpora, by optionally including detection of common names, acronyms, and strains.

4.3.3 Enzyme Recognition

Despite the standards published by the Enzyme Commission [13], enzymes are often described by the authors under various formats, from their 'Recommended Name' to different synonyms or abbreviations. Our enzyme recognition process is rule-based: Gazetteer and mapping lists are automatically extracted from the BRENDA database, in addition to a mapping list of SwissProt identifiers extracted from the SwissProt database.

An enzyme-specific text tokenization, along with grammar rules written in the JAPE language, analyses tokens with the *-ase* and *-ases* enzyme suffixes. The gazetteers allow to find the enzyme mentions in the documents by applying a pattern-matching approach.

Some abbreviated forms of enzyme names are not found during the pattern matching step, usually because these forms are created by the authors. The following sentence:[6]

> *The extracellular endoglucanase (EG) was purified to homogeneity from the culture supernatant by ethanol precipitation (75%, v/v), CM Bio-Gel A column chromatography, and Bio-Gel A-0.5m gel filtration. The purified EG (specific activity 43.33 U/mg protein) was a monomeric protein with a molecular weight of 27 000.*

shows the example of the **EG** abbreviation for endoglucanase, which is not reported in BRENDA. Such abbreviations are meaningful only within the context of a single document. Therefore, our pipeline contains grammar rules identifying these author's abbreviations and performing coreference resolution on each document.

The mapping lists link up the enzyme mentions found in the document and the external resources. Through this grounding step, the system provides the user with the enzymes' *Recommended Names, Systematic Names, EC Numbers, SwissProt Identifiers* and the *URL* of the related Web pages on the BRENDA website.

4.3.4 Temperature and pH Contexts

Temperature and pH mentions are involved in several biochemical contexts, like the temperature and pH dependence/stability or the description of the activity and kinetic assay conditions. Examples are given in the following sentences:[6]

> **Temperature:** *The purified enzyme exhibited maximum activity at 55℃, with 84% relative activity at 60℃ and 29% activity at 70℃ under the assay conditions used.*

> **pH:** *The enzyme displayed an optimum activity at pH 5.0 and retained 80% activity at pH 3.0 and also at pH 8.0.*

Our GATE pipeline contains PRs based on JAPE rules and gazetteer lists of specific vocabulary that enable the detection of these key mentions at the sentence level.

4.3.5 Other entities

The detection of the other entities mentioned in Table 1 is currently implemented through gazetteer lists and grammar rules implemented in JAPE; with the exception of the strain mentions, which are detected by the strain feature provided by the OrganismTagger pipeline.

4.4 System Output and User Interfaces

The system output supports two different tasks: the manual annotation of reference papers needed for evaluation purpose and the database curation manually performed by the biologists. In the context of manual annotation, the original papers are enriched with the system output added as pre-annotations before being submitted to the human annotators. In the context of database curation, all text mining pipelines are brokered as NLP Web services through the Semantic Assistants framework [23]. Users can access

[6]Excerpt of: Badal C. Saha, "Production, purification and properties of endoglucanase from a newly isolated strain of Mucor circinelloides", 2004, doi 10.1016/j.procbio.2003.09.013

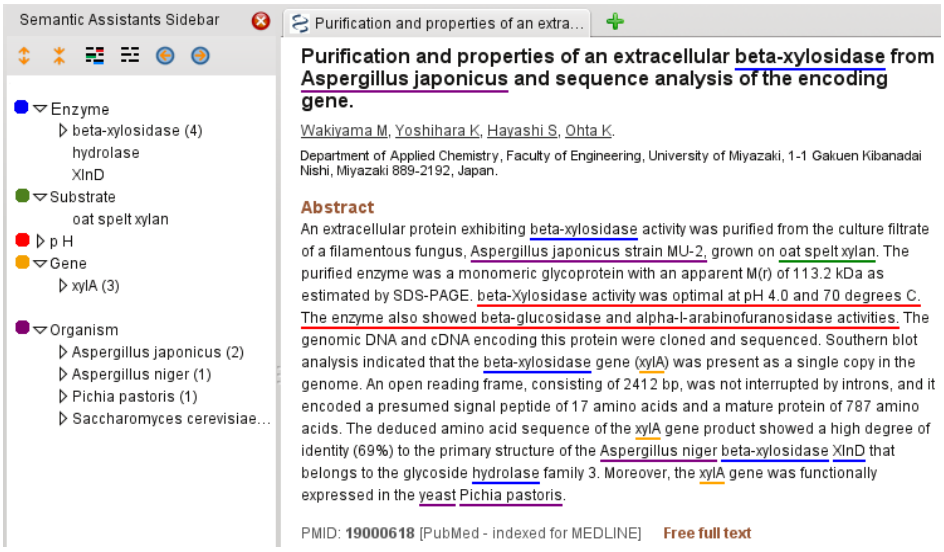

Figure 5: Text mining results displayed in Firefox through the Semantic Assistants Plug-In

these services from their desktop through client plug-ins for common tools, such as the Firefox Web browser (Figure 5) or the OpenOffice word processor. This provides the biologists using our system with the ability to quickly invoke semantic analysis services on scientific documents they browse online or edit in their text processor, without having to switch to an external text mining application.

External resources can be accessed from the user interfaces; the system output provides direct links to the relevant web pages, e.g., URLs of the Web pages related to the detected enzymes on the BRENDA web site or the found organisms on the NCBI Taxonomy web site.

5. EVALUATION AND RESULTS

In this section, we first discuss the development of the gold standard corpus and present preliminary results of our system.

5.1 Manual Annotation Process

For the intrinsic evaluation of our NLP pipelines, we are building a gold standard corpus of freely accessible full-text articles. These are manually annotated through GATE Teamware [4], a Web-based management platform for collaborative annotation and curation.

The tool reports on project status, annotator activity and statistics. The annotator's interface (see Figure 6) allows the curator to view, add and edit text annotations that are either manually created using the Teamware interface or pre-annotated. We make use of that ability by providing the annotators with documents we pre-annotate with our NLP pipelines throughout its development.

The annotation team consists of four biology researchers. The researcher in charge of the curation task and an annotator having a strong background in fungal literature curation are considered as **expert** annotators. The inter-annotator agreement between them is over 80% (F-measure), hence their annotation sets are always defined as the most reliable sets during the adjudication process.

5.2 Corpus

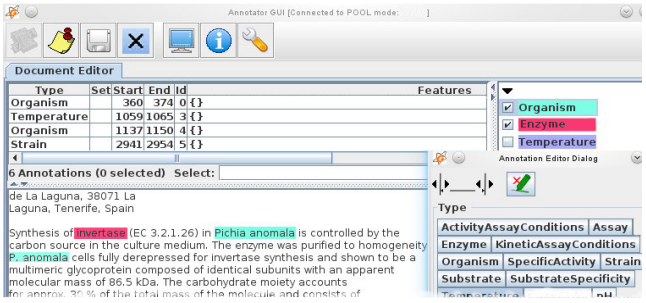

Figure 6: Teamware Annotator GUI

The corpus is composed of freely accessible full text articles containing critical knowledge and technical details the biology researchers aim to store in the mycoCLAP database specifically designed for their needs. The papers are related to a class of enzymes, among them the glycoside hydrolases, the lipases and the peroxidases. Glycoside hydrolase papers and lipase papers each represent 40% of the articles, whereas 20% are related to peroxidases. The current gold standard corpus is composed of ten full text papers that have been manually annotated by four biologists each.

At the word level, the two most common entities are enzymes and organisms, while the most common at the sentence level are pH and temperature. Table 2 shows these entities and their counts of occurrence in the current gold standard corpus. The goal for the current annotation task is to include fifty manually annotated papers in the gold standard corpus. This corpus will be available on demand.

5.3 Results

The performance of our text mining pipelines is evaluated in terms of precision, recall and F-measure. Here, the reference is provided by the gold standard corpus. Precision is defined as the number of correct tags hypothesized by the system divided by the total number of hypothesized tags.

Table 2: Entities and their counts in the current gold standard corpus

Entity	Counts
Enzyme	1493
Organism	984
pH	110
Temperature	115

Recall is defined as the number of correct tags hypothesized by the system divided by the total number of reference tags. The F-measure is the harmonic mean of precision and recall. For the 'strict' evaluation, we considers all partially correct responses as incorrect, while 'lenient' considers all partially correct (overlapping) responses as correct.

In this evaluation, we focus on the four most common entities (Enzyme, Organism, pH and Temperature) in our currently annotated corpus. The results of the text mining pipelines are shown in Table 3.

Table 3: Text Mining pipelines results on the gold standard corpus in terms of recall (R), precision (P) and F-measure (F)

	Strict			Lenient		
	R	**P**	**F**	**R**	**P**	**F**
Enzyme	0.74	0.65	0.70	0.88	0.77	0.82
Organism	0.87	0.88	0.87	0.91	0.92	0.91
pH	0.74	0.76	0.75	0.95	0.99	0.97
Temperature	0.64	0.67	0.65	0.90	0.93	0.91

5.4 Discussion

The OrganismTagger performance has previously been evaluated on two corpora, where it showed a precision of 95%–99%, a recall of 94%–97%, and a grounding accuracy of 97.4%–97.5% [16]. Since its result here are lower, we examined the error cases in more detail.

The manual annotation of organisms highlights all the textual mentions referring to an organism as indirect references, non-standard names (e.g., non-binomial names) or generic mentions. In some cases, correct results from the OrganismTagger were not manually annotated, leading to false positives. The following common sentence:

> *Soluble protein was determined according to the method of Lowry et al. (1951) using* ***bovine*** *serum albumin as standard.*

shows an example of such a case where the OrganismTagger correctly annotates *bovine* as an organism, whereas the expert annotators considered *bovine serum albumin* as a stand-alone expression.

In some other cases, human annotations are not detected by the OrganismTagger. For example, *Trichoderma viridie* and *M. Incrasata* or *cellulolytic fungi*[7] were manually annotated as organisms by the experts. These mentions are not detected by the OrganismTagger. In the first two cases, the cause is a spelling difference between the names of the organisms reported in the NCBI Taxonomy database and their mention in the article. In the last case, the annotation of a generic organism mention that is relevant within the context of our project is not an objective of the OrganismTagger system, which is designed to provide normalization with scientific names and grounding to the NCBI Taxonomy database.

Consequently, the results obtained by our pipeline on the organism recognition are lower than the published results of the OrganismTagger system. The text mining pipeline supporting our system needs to be enhanced in its ability to capture generic organism mentions and to discard stand-alone expressions containing organism names.

The results obtained on *Temperature* and *pH* sentence detection are much better in the lenient evaluation than the strict because of sentence splitter mistakes.

The enzyme recognition pipeline provides state-of-the-art performance. However, wrong detection of abbreviations and acronyms represent 92% of the false negatives found by our pipeline. Further work is needed to reduce this amount by improving the co-reference resolution with approaches as described in [17] and external resources, such as Allie[8] [26].

6. CONCLUSIONS

We presented our ongoing development of a semantic infrastructure for enzyme data management. As the first system specifically designed for lignocellulolytic enzymes research, it targets the automatic extraction of knowledge on fungal enzymes from the research literature. The proposed approach is based on text mining pipelines combined with ontological resources. Preliminary experiments show state-of-the-art results. Improving the consistency of the extracted knowledge by increasing the use of ontologies is one of the next goals for our system. Therefore, a key objective is the population of the overall ontology of the domain knowledge and its publication in Linked Data format.

The gold standard corpus of manually annotated papers will be made available, as well as the presented system.

The accessibility of the services through the Semantic Assistants framework allows the users to mine the semantically annotated literature from their desktop. Future work is needed to enable the interaction between selected users (e.g., curators) and the presented system in terms of data validation and knowledge acquisition.

In future work, we will further deploy our text mining pipelines to assess the quality of existing manually curated data in the databases. Measuring the overall impact of the semantic system on the scientific discovery workflow will be the target of an extrinsic study.

7. ACKNOWLEDGMENTS

Funding for this work was provided by Genome Canada and Génome Québec. Bahar Sateli is acknowledged for help on the Semantic Assistants resources. We also thank Carolina Cantu, Shary Semarjit and Sherry Wu who helped on the annotation task.

8. REFERENCES

[7]Examples from: Badal C. Saha, "Production, purification and properties of endoglucanase from a newly isolated strain of Mucor circinelloides", 2004, doi 10.1016/j.procbio.2003.09.013

[8]Allie, http://allie.dbcls.jp/

[1] S. Ananiadou and J. McNaught. *Text Mining for Biology And Biomedicine*. Artech House, Inc., Norwood, MA, USA, 2005.

[2] C. J. O. Baker and K.-H. Cheung, editors. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Springer, 2007.

[3] D. C. Bernard, B. F. Buxton, W. B. Langdon, and D. T. Jones. BioRAT: Extracting Biological Information from Full-length Papers. *Bioinformatics*, 20:3206–3213, 2004.

[4] K. Bontcheva, H. Cunningham, I. Roberts, and V. Tablan. Web-based Collaborative Corpus Annotation: Requirements and a Framework Implementation. In *New Challenges for NLP Frameworks*, pages 20–27, Valletta, Malta, May 22 2010. ELRA.

[5] K. Bontcheva, V. Tablan, D. Maynard, and H. Cunningham. Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering*, 10:349–373, 2004.

[6] S. Bringezu, H. Schütz, M. O'Brien, L. Kauppi, R. W. Howarth, and J. McNelly. Towards sustainable production and use of resources: ASSESSING BIOFUELS. Technical report, United Nations Environment Programme, 2009.

[7] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. *Text Processing with GATE (Version 6)*. University of Sheffield, Department of Computer Science, 2011.

[8] A. Demirbas. Political, economic and environmental impacts of biofuels: A review. *Applied Energy*, 86(Supplement 1):S108–S117, 2009.

[9] A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33(suppl 2):W783–W786, 2005.

[10] S. Federhen. The Taxonomy Project. In J. McEntyre and J. Ostell, editors, *The NCBI Handbook*, chapter 4. National Library of Medicine (US), National Center for Biotechnology Information, 2003.

[11] C. Görg, H. Tipney, K. Verspoor, W. Baumgartner, K. Cohen, J. Stasko, and L. Hunter. Visualization and Language Processing for Supporting Analysis across the Biomedical Literature. In R. Setchi, I. Jordanov, R. Howlett, and L. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6279 of *Lecture Notes in Computer Science*, pages 420–429. Springer Berlin/Heidelberg, 2010.

[12] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nature Genetics*, 36, 664:1061–4036, 2004.

[13] International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature 1992*. Academic Press, San Diego, California, 1992.

[14] H.-M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biol*, 2(11):e309, 09 2004.

[15] C. Murphy, J. Powlowski, M. Wu, G. Butler, and A. Tsang. Curation of characterized glycoside hydrolases of fungal origin. *Database*, vol.2011, 2011.

[16] N. Naderi, T. Kappler, C. J. Baker, and R. Witte. OrganismTagger: Detection, normalization, and grounding of organism entities in biomedical documents. *Bioinformatics*, 2011.

[17] N. Okazaki, S. Ananiadou, and J. Tsujii. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253, 2010.

[18] E. Pafilis, S. I. O'Donoghue, L. J. Jensen, H. Horn, M. Kuhn, N. P. Brown, and R. Schneider. Reflect: augmented browsing for the life scientist. *Nature Biotechnology*, 27:508–510, 2009.

[19] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. John Wilbur, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 38(suppl 1):D5–D16, 2009.

[20] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele, and D. Schomburg. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, 39:(Database issue):D670–676, 2011.

[21] N. Shadbolt, T. Berners-Lee, and W. Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

[22] The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 37(D):169–174, 2009.

[23] R. Witte and T. Gitzinger. Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In *3rd Asian Semantic Web Conference (ASWC 2008)*, volume 5367 of *LNCS*, pages 360–374, Bangkok, Thailand, 2009. Springer.

[24] R. Witte, T. Kappler, and C. J. O. Baker. Ontology Design for Biomedical Text Mining. In Baker and Cheung [2], chapter 13, pages 281–313.

[25] R. Witte, N. Khamis, and J. Rilling. Flexible Ontology Population from Text: The OwlExporter. In *The Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3845–3850, Valletta, Malta, 2010. ELRA.

[26] H. B. Y. Yamamoto, A. Yamaguchi and T. Takagi. Allie: a database and a search service of abbreviations and long forms. *Database 2011:bar013*, 2010.