Towards Evaluating the Impact of Semantic Support for Curating the Fungus Scientific Literature

Marie-Jean Meurs[1], Caitlin Murphy[2,3], Nona Naderi[1], Ingo Morgenstern[2,3],
Carolina Cantu[2,3], Shary Semarjit[2,4], Greg Butler[1,2], Justin Powlowski[2,4],
Adrian Tsang[2,3] and René Witte[1]*

[1] Department of Computer Science and Software Engineering
[2] Centre for Structural and Functional Genomics
[3] Department of Biology
[4] Department of Chemistry and Biochemistry
Concordia University, Montréal, QC, Canada
mjmeurs@encs.concordia.ca, cmurphy@gene.concordia.ca,
n_nad@encs.concordia.ca, {imorgenstern,ccantut,sshary}@gene.concordia.ca,
gregb@encs.concordia.ca, powlow@alcor.concordia.ca,
tsang@gene.concordia.ca, rwitte@cse.concordia.ca

Abstract. We present our ongoing development of a semantic infrastructure supporting biofuel research. Part of this effort is the automatic curation of knowledge from the massive amount of information on fungal enzymes that is available in genomics. Working closely with biologists who manually curate the existing literature, we developed ontological NLP pipelines, integrated through Web-based interfaces, to help them in two main tasks: spending less time to mine the literature for facts, while also being provided with richer and semantically linked information. An ongoing challenge is to measure precisely how much the developed semantic technologies benefit the end users and what their overall impact on the quality of the curated data is. We present preliminary evaluation results that show a significant reduction in manual curation time.

1 Introduction

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 based fuels. Extracting sugars from cellulose to produce biofuels requires to break down cellulose by using specific molecules called enzymes. Therefore, in the current race for replacing petroleum based fuels with renewable biofuels, discovering the most efficient enzymes for the cellulose degradation is a key challenge.

The largest knowledge source available to biofuel researchers is the PubMed bibliographic database, containing more than 19 million citations from over 21,000 life science journals. PubMed is linked to other databases, like *Entrez Genome*, which provides access to genomic sequences or *BRENDA, The Comprehensive Enzyme Information System* [9], which is the main collection of enzyme functional

[*] corresponding author

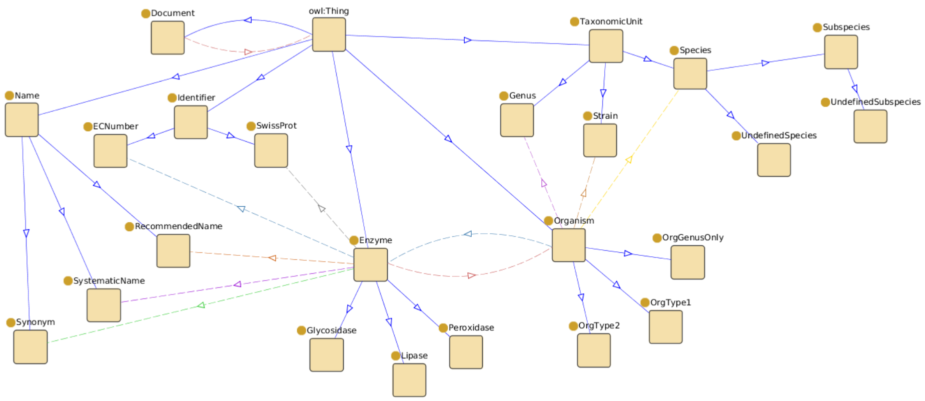

Fig. 1. Domain Ontology: Organism and Enzyme Entities

data available to the scientific community. A biology researcher querying PubMed using keywords collects an often long list of relevant papers. The way to analyze this collection is reading all the abstracts and sometimes the full text papers: this task is time consuming, difficult to handle and significant knowledge can be easily missed.

To address this problem, Natural Language Processing (NLP) and Semantic Web approaches are increasingly adopted in biomedical research [2, 10]. The work-in-progress we present in this paper focuses on the automatic extraction of knowledge from the massive amount of information on enzymes in fungi available from genome research. Text mining systems, like the one we developed here, are typically evaluated with *intrinsic* metrics, such as precision and recall. However, while these metrics can give insight into the accuracy of a system, they do not necessarily correspond to their *extrinsic* performance [1, 4]: How much does the system actually improve the tasks performed by users? Thus, in this work we are interested in also evaluating the impact of our semantic systems on the work performed by our biologists and the quality of the curated data.

2 Project Context and System Architecture

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

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 provides the thorough knowledge required for guiding research and experiments. The biology researchers involved in this curation are filling the mycoCLAP database [8], 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 integrating the various knowledge resources.

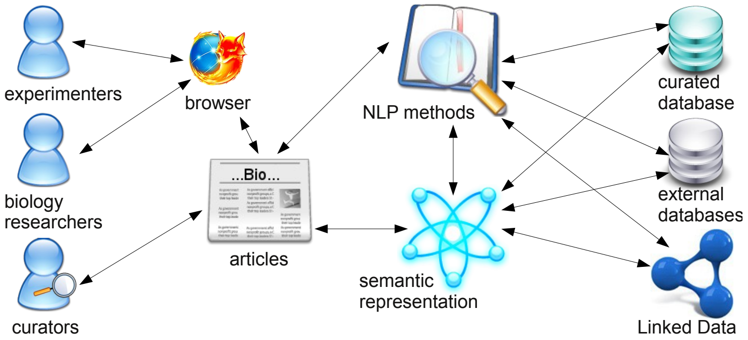

Fig. 2. Integrating Semantic Support in Curation, Analysis, and Retrieval

Semantic Entities. The system we are developing has to support the manual curation process; therefore, the semantic annotation types have been defined by the curators according to the information they need to store in the mycoCLAP database. Entities include information such as organisms, enzymes, assays, genes, kinetic properties, reactions, substrates, and environmental conditions. To facilitate semantic discovery, linking and querying these concepts across literature and databases, these entities are modeled in OWL ontologies, which are automatically populated from documents. As an example, Fig. 1 shows two main entities encoded in our ontology, *organisms* [13] and *enzymes*. The ontology is used both during the text mining process and for querying the extracted information.

Semantic Resources. In terms of knowledge sources, the system relies on external and internal processing resources and ontologies. The *Taxonomy database* [6] from NCBI is used for initializing the NLP resources supporting the organism recognition. BRENDA [9] provides the enzyme knowledge along with the UniProtKB/SwissProt [11]. References to the original sources are integrated into the curated data. This facilitates semantic connections through standard Linked Data techniques, e.g., from an organism mention in a research paper to its corresponding entry in the NCBI Taxonomy database.

System Architecture. With the large number of 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 these 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 (Fig. 2).

NLP services are provided by the Semantic Assistants architecture [12], which facilitates the publication of NLP pipelines through standard Web services with WSDL descriptions. Users can access these Semantic Assistants services from their desktop through client plug-ins for common tools, such as the Firefox Web browser or the OpenOffice word processor.

3 Text Mining Pipelines

Our text mining pipelines are based on the *General Architecture for Text Engineering* (GATE) [5]. All documents first undergo basic preprocessing steps using off-the-shelf components, such as tokenization, sentence splitting, and part-of-speech tagging. Custom pipelines then extract the semantic entities mentioned above and populate the OWL ontologies using the OwlExporter 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 curation.

Organism Recognition. The organism tagging and extraction relies on external resources that are automatically translated for reuse in our system, thereby providing users with the ability to update their installation when the NCBI Taxonomy database changes. Additionally, a custom built organism ontology, presented in Fig. 1, formally describes the linguistic structure of organism entities at different levels of the taxonomic hierarchy [13]. The GATE pipeline consists of modules for organism entity detection based on pattern matching to the NCBI reference taxonomy, providing scientific names and the NCBI Taxonomy Identifier. Strain mentions are extracted using a specific text tokenization and a machine learning based approach.

Enzyme Recognition. Despite the standards published by the Enzyme Commission [7], enzymes are often described by the authors under various formats. An enzyme-specific text tokenization, along with grammar rules written in the JAPE language, analyses tokens with the *-ase* enzyme suffix. Then, the enzyme entity recognition relies on automatically extracted knowledge from the BRENDA database. A pattern matching approach provides enzyme name identification. The detected enzyme mentions are associated with their *EC number*, their *Recommended Name*, their *Systematic Name* and their URL on the BRENDA website.

Temperature and pH Facts. Temperature and pH mentions are involved in several biological facts, like the temperature and pH dependence/stability or the description of the activity and kinetic assay conditions. 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 Intrinsic and Extrinsic Evaluation

As explained above, text mining systems require an evaluation showing their efficiency and effectiveness, both intrinsically and from an end user's point of view. In this section, we first discuss the development of the gold standard corpus and present preliminary evaluation results of our system.

4.1 The Manual Annotation Process

For the intrinsic evaluation, we are building a gold standard corpus of freely accessible full-text articles by manually annotating them using GATE Teamware [3], a Web-based management platform for collaborative annotation and curation. The annotation team is composed of four biology researchers. The researcher in charge of the curation task and an annotator having a strong background in fungus literature curation are considered as expert annotators. Their inter-annotator agreement is over 80%, hence their annotation sets are always defined as the most reliable sets during the adjudication process. The corpus is composed of ten papers related to a class of enzymes. Glycoside hydrolase papers and lipase papers each represent 40% of the articles, whereas 20% are related to peroxidases.

4.2 Intrinsic Evaluation: Precision and Recall

The correctness of our text mining pipelines is evaluated in terms of precision, recall and F-measure. The reference is provided by the manually annotated (gold standard) corpus. The preliminary results on the four most common entities (Enzyme, Organism, pH and Temperature) are shown in Table 1.

Table 1. Text Mining Pipelines: Precision, Recall and F-measure

	Strict (overlaps discarded)			Lenient (overlaps included)		
	Recall	**Precision**	**F-m**	**Recall**	**Precision**	**F-m**
Enzyme	0.64	0.55	0.59	0.78	0.67	0.72
Organism	0.84	0.81	0.82	0.88	0.83	0.85
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

4.3 Extrinsic Evaluation: Literature Mining and Annotation

The impact of the system on the *curation* and *annotation* tasks is evaluated in terms of required time (range and average) per paper and measured in minutes.

Paper selection. Since the beginning of the curation task, approximately 1000 papers have been examined. The time needed to examine an unannotated full paper and to make a decision about its selection for curation, without any semantic support, previously ranged from 2 to 3 minutes. With added support through the text mining services, the required time decreased to 1–2 minutes.

Paper curation. Among the 1000 examined papers, around 600 were already selected for curation. The time needed to curate an unannotated full paper, i.e., extracting salient facts for entry into the mycoCLAP database, ranged from 30 to 45 minutes for the fully manual workflow. With added semantic support through the text mining pipelines, the required time decreased to 20–30 minutes.

Paper annotation. For full paper annotation, we investigated the impact of different levels of semantic support on the time required to add annotations (Table 2). All sets have been manually annotated by four annotators. The 4 papers of the first set (SET 1) were annotated without any semantic support. The second set (SET 2) is composed of 3 papers, which have been pre-annotated by a degraded version of the system, using only generic tools, such as simple gazetteering list, resulting in lower precision and recall. The third set (SET 3) contains 3 papers, pre-annotated using the complete text mining pipelines, including the specialized tools and external resources as described above.

From the preliminary results, we can conclude that (1) there is a significant reduction of the average time required for paper selection, curation and annotation and (2) the level of support has a measurable impact as well.

Table 2. Average annotation time per paper with different levels of semantic support

set and level of semantic support		available tags	$\bar{t}$ (min)
SET 1	(no semantic support)	∅	90
SET 2	(partial semantic support)	enzyme, organism, pH, temperature	65
SET 3	(full semantic support)	enzyme, organism, pH, temperature	56

5 Conclusions

We presented our ongoing development of a semantic infrastructure for enzyme data management. In the context of biofuel research, our system targets the automatic extraction of knowledge on fungal enzymes from genome research literature. Preliminary experiments show that semantic support allows for a significant decrease in manual curation time. However, future work is needed to evaluate the impact of such a system on the quality of the curated data.

Acknowledgments. Funding for this work was provided by Genome Canada and Génome Québec.

References

1. Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Wang, X.: Assisted curation: does text mining really help. In: Pacific Symposium on Biocomputing. vol. 13, pp. 556–567 (2008)
2. Ananiadou, S., McNaught, J.: Text Mining for Biology And Biomedicine. Artech House, Inc., Norwood, MA, USA (2005)
3. Bontcheva, K., Cunningham, H., Roberts, I., Tablan, V.: Web-based Collaborative Corpus Annotation: Requirements and a Framework Implementation. In: New Challenges for NLP Frameworks. pp. 20–27. ELRA, Valletta, Malta (May 22 2010)
4. Caporaso, J.G., Deshpande, N., Fink, J.L., Bourne, P.E., Cohen, K.B., Hunter, L.: Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In: Pacific Symposium on Biocomputing. vol. 13, pp. 640–651. World Scientific Publishing (2008)
5. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proc. 40th Anniversary Meeting of the ACL (2002)
6. Federhen, S.: The Taxonomy Project. In: McEntyre, J., Ostell, J. (eds.) The NCBI Handbook, chap. 4. National Library of Medicine (US), National Center for Biotechnology Information (2003)
7. International Union of Biochemistry and Molecular Biology: Enzyme Nomenclature 1992. Academic Press, San Diego, California (1992)
8. Murphy, C., Powlowski, J., Wu, M., Butler, G., Tsang, A.: Curation of characterized glycoside hydrolases of fungal origin. Database 2011 (2011)
9. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., Schomburg, D.: BRENDA, the enzyme information system in 2011. Nucleic Acids Res. 39, (Database issue):D670–676 (2011)
10. Shadbolt, N., Berners-Lee, T., Hall, W.: The semantic web revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
11. The UniProt Consortium: The Universal Protein Resource (UniProt). Nucleic Acids Research 37(D), 169–174 (2009)
12. Witte, R., Gitzinger, T.: Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In: 3rd Asian Semantic Web Conference (ASWC 2008). LNCS, vol. 5367, pp. 360–374. Springer, Bangkok, Thailand (2009)
13. Witte, R., Kappler, T., Baker, C.J.O.: Ontology Design for Biomedical Text Mining. In: Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences, chap. 13, pp. 281–313. Springer (2007)