IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences

Bahar Sateli[1], Marie-Jean Meurs[1,2], Greg Butler[1,2], Justin Powlowski[2,3], Adrian Tsang[2,4], René Witte[1]✉

[1]Department of Computer Science and Software Engineering; Concordia University, Montréal, Canada
[2]Centre for Structural and Functional Genomics; Concordia University, Montréal, Canada
[3]Department of Chemistry and Biochemistry; Concordia University, Montréal, Canada
[4]Department of Biology; Concordia University, Montréal, Canada

Motivation and Objectives

The rapid growth of the scholarly literature makes the management and curation of the available information a labor-intensive and time-consuming task for researchers, during which significant knowledge can be easily missed. To address this problem, efforts have been made to use Natural Language Processing (NLP) techniques as a means to (semi-)automatically improve the exhaustive analysis of the available information. In order to make these NLP techniques more end-user friendly and integrate them with knowledge management workflows, we developed IntelliGenWiki, a novel combination of a wiki system with state-of-the-art techniques from the NLP and Semantic Computing domains. Wikis are well known as an easy-to-use, collaborative platform for creating and organizing knowledge. For example, the Gene Wiki project (Huss III et al, 2010) applies community intelligence to the annotation of gene and protein functions. However, existing approaches rely on a manual analysis of the literature. With IntelliGenWiki, we aim to leverage the collaborative nature of wikis by introducing new Human-AI collaboration patterns: Our goal is to provide text mining assistants that work together with humans on literature analysis tasks, like curation or the generation of semantic metadata, which can be used in an Linked Open Data context. IntelliGenWiki is based on an open service-oriented architecture: it can be applied to different projects by deploying custom NLP analysis pipelines suitable for the specific task and domain. Here, we demonstrate the benefits of this approach within a collaborative literature curation context.

Methods

We first describe the general workflow for working with NLP assistants, followed by a description of the underlying architecture.

Workflow. IntelliGenWiki provides a standard wiki user interface. From any wiki page (Fig. 1, top), users can ask for "Semantic Assistants" from the menu (Fig. 1, left), which will result in a dynamically injected user interface from which assistants can be selected and executed (Fig. 1, bottom). The user can now select an appropriate assistant from the list and invoke it on one or multiple pages of the wiki, gathered in a so-called "collection". This will invoke the selected NLP pipeline on the set of wiki pages. The results (e.g., detected entities) are stored in the user's place of choice and made persistent in the wiki repository (Fig. 1, middle). Thereby, all updated pages become immediately available to all wiki users for collaborative adjustment, modification and further refinement of the results.

Architecture. Technically, IntelliGenWiki combines NLP analysis pipelines developed in the General Architecture for Text Engineering (GATE) (Cunningham et al, 2011) with MediaWiki, http://www.mediawiki.org (Last accessed: 26.09.2012), a widely-used wiki engine. These pipelines are published as standard web services through the Semantic Assistants framework (Witte and Gitzinger, 2008). The Wiki-NLP integration is based on a service-oriented architecture that seamlessly introduces these NLP web services into wiki systems (Sateli and Witte, 2012). This allows wiki users to benefit from text mining techniques directly within their wiki environment, without the need for switching to an external application. Additionally, we support the generation of semantic metadata from NLP analysis results. This metadata is formally represented in the wiki through the Semantic MediaWiki (SMW) extension: http://semantic-mediawiki.org/ (Last accessed on Sept 26, 2012). This formal representation of the available wiki knowledge can be exploited by exporting it in form of RDF triples. It can also be queried directly within the wiki using SMW inline queries. For example, users could write queries to retrieve

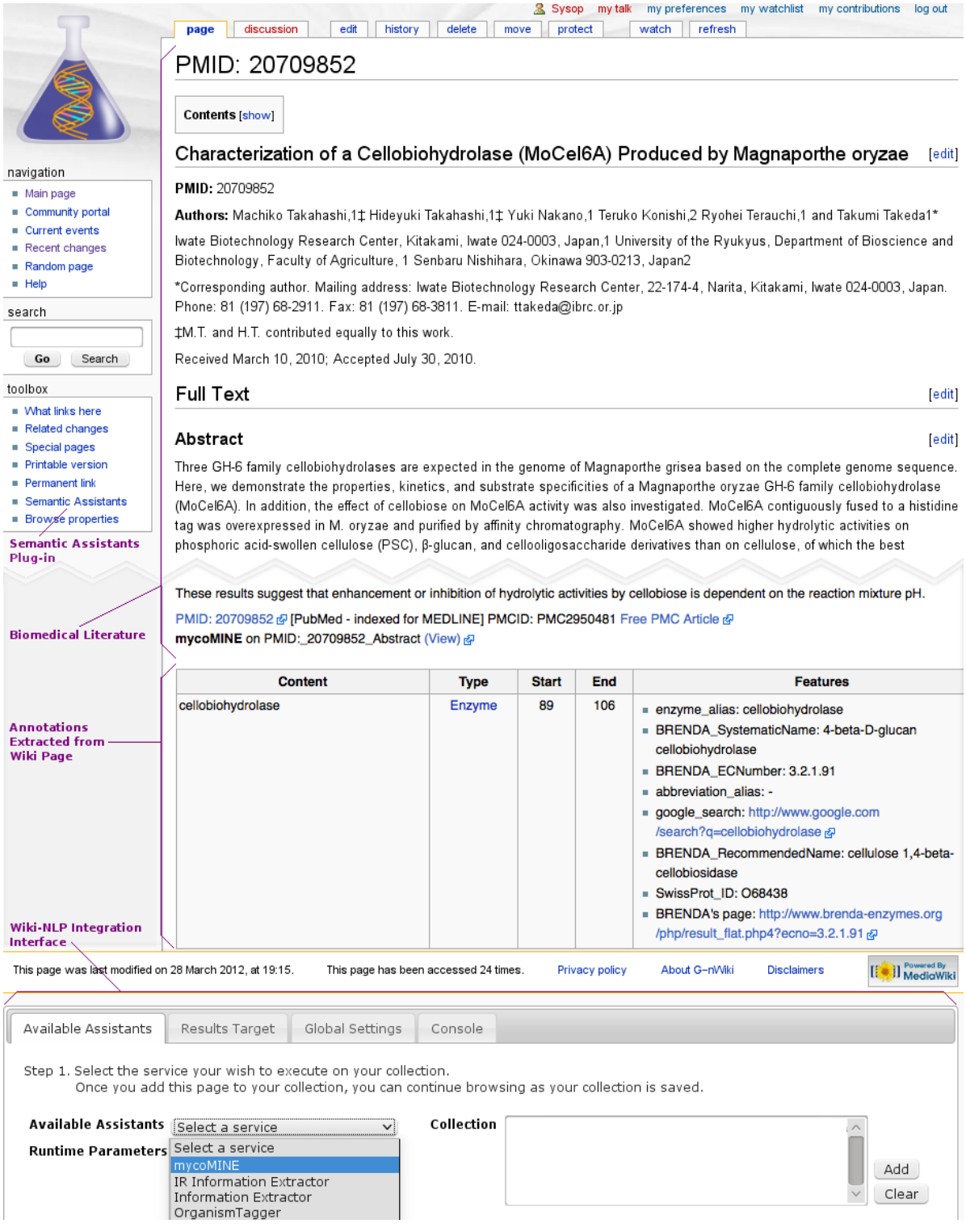

Fig. 1: The wiki interface with integrated text analysis services (bottom), showing automatically added, NLP-extracted entities (middle), together with original content (top)

literature that contains a certain type of entities, such as enzymes or organisms.

Results and Discussion

To test the effectiveness of NLP assistants in a wiki environment, we deployed an IntelliGenWiki installation within the Genozymes project: http://www.fungalgenomics.ca (Last accessed on Sept 20, 2012). The task we aimed to support in the project is biomedical literature curation for lignocellulose research. For this experiment, we deployed the mycoMINE NLP pipeline (Meurs et al, 2012), which automatically extracts knowledge from the literature on fungal enzymes by using semantic text mining approaches combined with ontological resources. We manually pre-filled the wiki with a corpus of 30 documents composed of PubMed abstracts and their corresponding full-text papers, selected by two expert biocurators. These biocurators provided us with their average time needed for curation without support on the same task. They performed the corpus curation through the wiki using mycoMINE to automatically extract relevant entities, and they kept track of the time spent on each document. The time for abstract selection (triage task) decreased from 1min. (without support) to 20sec. (using IntelliGenWiki), and from 37.5min (without support) to 30.6min (using IntelliGenWiki) for full paper selection (curation task), showing a productivity enhancement of 67% and 20%, respectively. The results gathered from this experiment confirm the usability and the effectiveness of our approach.

The IntelliGenWiki system, including the NLP integration back-end, is available as open source software from http://www.semanticsoftware.info/intelligenwiki.

Acknowledgements

Funding for this work was provided by NSERC, Genome Canada and Génome Québec. Caitlin Murphy and Sherry Wu are acknowledged for their participation in the evaluation task.

References

1. Cunningham H, Maynard D, et al (2011) Text Processing with GATE (Version 6), University of Sheffield, Department of Computer Science
2. Huss III J. W., et al (2010) The Gene Wiki: Community Intelligence Applied to Human Gene Annotation, Nucleic Acids Research 38, p. 633–639. doi:10.1093/nar/gkp760
3. Meurs MJ, Murphy C, et al (2012) Semantic Text Mining Support for Lignocellulose Research, BMC Medical Informatics and Decision Making 12(Suppl 1):S5. doi:10.1186/1472-6947-12-S1-S5
4. Sateli B and Witte R (2012) Natural Language Processing for MediaWiki – The Semantic Assistants Approach, In 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012). Linz, Austria.
5. Witte R and Gitzinger T (2008) Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients, In Asian Semantic Web Conference (ASWC 2008), Springer LNCS 5367, pp.360–374. doi:10.1007/978-3-540-89704-0_25