# Text Mining Assistants in Wikis for Biocuration

Bahar Sateli, Caitlin Murphy, René Witte, Marie-Jean Meurs and Adrian Tsang
Concordia University

Biocuration 2012
The Conference of the International Society for Biocuration, Washington DC, USA, April 2-4, 2012

**Abstract**

Researchers need to extract critical knowledge from a massive amount of literature available in multiple and ever-growing repositories. The sheer volume of information makes the exhaustive analysis of literature a labor-intensive and time consuming task, during which significant knowledge can be easily missed.
In addition, the metadata generated from the curation process is typically not described in a standard language that would allow the knowledge to be machine-readable and reused in an Open Data context.
To address these problems, Natural Language Processing (NLP) and Semantic Web approaches are increasingly adopted in biomedical research. An ongoing challenge is to select appropriate technologies and combine them in a coherent system that brings measurable improvements to the users.

We present our ongoing development of a generic architecture for collaborative literature curation through a user-friendly wiki interface. Our architecture seamlessly integrates NLP capabilities in a wiki environment, allowing users - curators - to benefit from text mining techniques to discover knowledge embodied in the wiki. Content to be curated is first imported into the wiki system. In addition, domain-specific NLP pipelines, in our context developed based on the General Architecture for Text Engineering (GATE), need to be deployed. The curator's wiki interface then becomes enhanced with automated "Semantic Assistants" that collaborate with him on locating important knowledge in the wiki, much like a human assistant would. For example, concrete NLP pipelines can locate biomedical concepts, their relations, or other entities.

Technically, our wiki-NLP integration adds a user interface that is dynamically injected into the users' browser on-the-fly, allowing him to invoke any arbitrary NLP service that has been made available through the Semantic Assistants architecture. Once a user requests help from a specific assistant, the selected wiki content is sent to the designated NLP pipeline for analysis. Following a successful service execution, results are transformed by the architecture and added to the wiki's database. Thereby, all updated pages become immediately available to all curators for collaborative adjustment, modification and refinement of the results. Additionally, the wiki-based framework further facilitates the curation process by automatically versioning wiki pages and providing roll-back functionality in case of erroneous annotations.

In one concrete application example, we have integrated our architecture with MediaWiki, a widely-used wiki engine best known from the Wikipedia project. The transformation of the NLP pipelines' results into RDF triples is realized through the Semantic MediaWiki (SMW) extension. This standard, formal representation of the extracted knowledge permits the users to semantically query the wiki content, in addition to manual browsing.

To evaluate our integration, we deployed it within the Genozymes project in order to support biomedical literature curation for lignocellulose research. The NLP service used in the experiment was our mycoMINE, a pipeline that automatically extracts knowledge from the literature on fungal enzymes by using semantic text mining approaches combined with ontological resources. The results gathered from this experiment confirm the usability and the effectiveness of our approach.