Information extraction from biomedical literature for community sharing.

Biomedical literature, including scientific articles, public health reports, books, become more and more available to user due to massive digitalization. Exploration and analysis of this rich source of data requires assistance of automatic tools capable of dealing with large volumes of texts.

In Bioinformatics core facility at LCSB, University of Luxembourg, we are developing a pipeline, called “Scraper” for processing publically available biomedical texts, abstracts, full texts, eventually books and electronic health records, starting from searching the web and downloading row files, to extraction and storing concepts and semantic relations between them into a knowledge base. We rely on “Reflect” (httpd://reflect.ws)[1] - a named entity recognition engine to identify biomedical concepts in the text. “GeniaTagger”[2] is used to obtain basic morphologic and syntactic information. The latter is completed by application of the “Stanford Syntactic Parser”[3] which converts sentences into syntactic trees representing dependencies between the words. Combined with a set of rules and dedicated patterns, this information allows for getting semantic interpretation of sentences and extraction of meaningful relationships between the concepts. Extracted knowledge is represented according to Genia ontology model (http://www.medlingmap.org/taxonomy/term/102).

During the Hackathon we would like to refine the current ontology and create converters from and to various formats, such as PubAnnotation, JSON-LD, OpenBel and SBML. The idea behind is to:
- allow information exchange between multiple databases which represent text mining results using triple store with a common model;
- facilitate automatic and curator’s query-based updates of disease maps.

[1] Pafilis, O’Donoghue, Jensen, Horn, Kuhn, Brown, Schneider. Reflect: Augmented Browsing for the Life Scientist. Nature Biotechnology, 27, 508-510, 2009.

[2] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii, Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005.

[3] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computa- tional Linguistics: System Demonstrations; 2014. p. 55–60.