OntoGene
www.ontogene.org

ODIN: An Advanced Interface for the Curation of Biomedical Literature

Fabio Rinaldi, Simon Clematide, Gerold Schneider, Martin Romacker, Thérèse Vachon

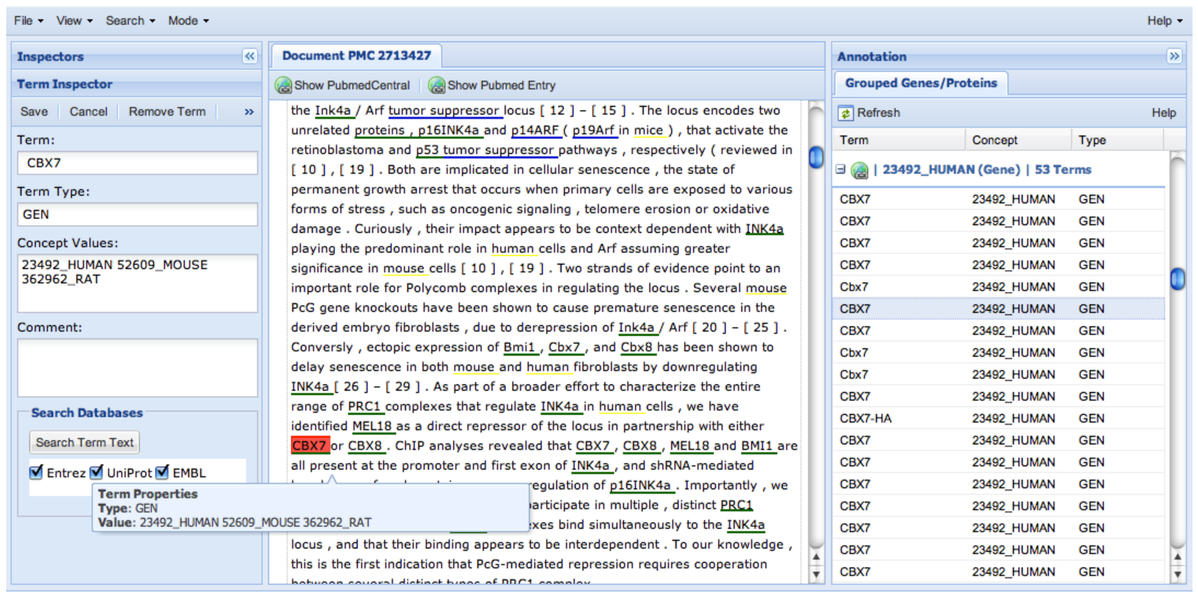
Introduction

ODIN (Ontogene Document INspector) is a system for interactive curation of biomedical literature, developed within the scope of the **SASEBio** project (Semi-Automated Semantic Enrichment of the Literature), as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group of Novartis Pharma AG. The purpose of the system is to allow a human annotator/curator to leverage upon the results of an advanced text mining system in order to enhance the speed and effectiveness of the annotation process.

The OntoGene system takes as input a document (e.g a full paper from PubMed Central) and processes it with a custom NLP pipeline, which includes Named Entity recognition and relation extraction. Entities which are currently supported include proteins, genes, experimental methods, cell lines, species. Entities detected in the input document are disambiguated with respect to a reference database (UniProt, EntrezGene, NCBI taxonomy, PSI-MI ontology). The annotated documents are handed back to the ODIN interface, which allows multiple display modalities.

ODIN: screenshot illustrating the inspection and editing functionalities.

The curator/annotator can view the whole document with in-line annotations highlighted, or can browse the extracted entities and be pointed back to the mentions of the entities within the original document. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change their extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences containing desired entities) is supported. Additionally, extensive logging functionalities are provided. All documents and entities are fully interlinked to reference databases, for the purpose of simplified inspection. Entities can be grouped in classes (e.g. by species) and actions can be applied to whole classes, for selective editing or removal.

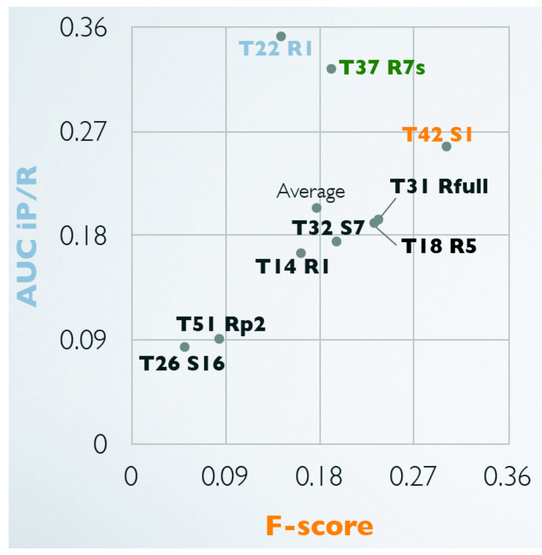


Another view of ODIN

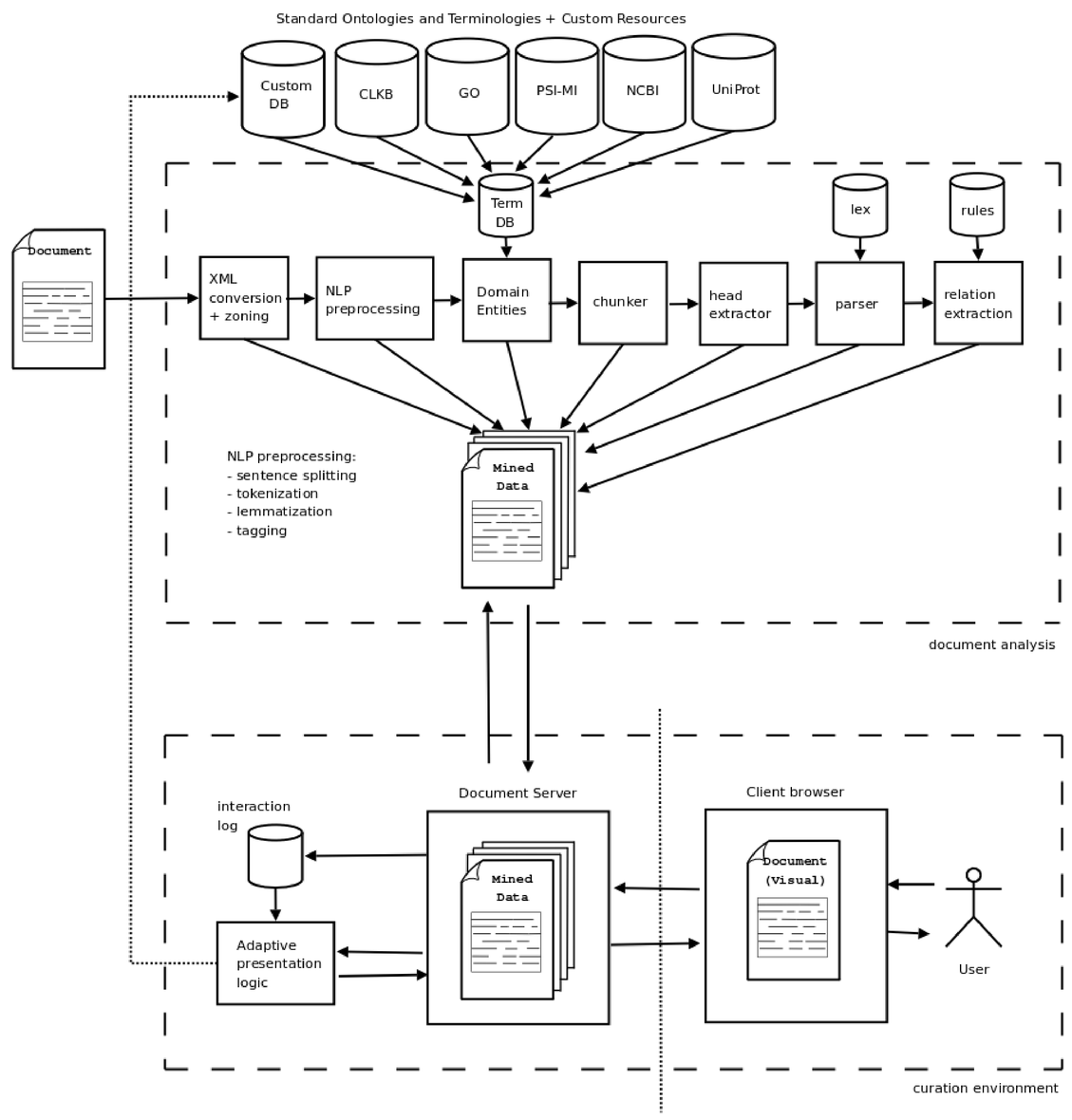
Evaluation: BioCreative competitions

In our initial application [1], relation mining was based on manually constructed cascading rules, which were organized modularly in order to support increasingly abstract types of queries. This approach has been validated through our participation in the BioCreative competitive evaluations of biomedical text mining systems. Our results in BioCreative II (2006), were among the best reported [2].

Later we developed a new approach based on a combination of machine learning and linguistic insight, which learns the syntactic paths expressing protein-protein interactions [3]. At BioCreative II.5 (2009) this approach served as the basis of our contribution. This resulted in the best run for the detection of protein-protein interactions (according to the 'raw' AUC iP/R metric). Our system was overall considered the best balanced system and among the best three [4] (look for system T37 in the graph to the right). In 2010, we have participated to the BioCreative III text mining challenge, achieving very competitive results in all of the tasks [5].



Architecture



References

[1] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker, An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA, BMC Bioinformatics, 7(Suppl 3):S3, 2006.

[2] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon, Ontogene in Biocreative II, Genome Biology, 9:S13, 2008.

[3] Gerold Schneider, Kaarel Kaljurand, Thomas Kappeler, Fabio Rinaldi. Detecting protein-protein interactions in biomedical texts using a parser and linguistic resources. CICLING 2009.

[4] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, Martin Romacker, "OntoGene in BioCreative II.5," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(3), pp. 472-480, 2010.

[5] Fabio Rinaldi, Gerold Schneider, Simon Clematide, Silvan Jegen, Pierre Parisot, Martin Romacker and Therese Vachon. OntoGene (Team 65): preliminary analysis of participation in BioCreative III. BioCreative III workshop, Bethesda, Maryland, September 13-15, 2010.

Funding

OntoGene is partially supported by the **Swiss National Science Foundation** (grants 100014 – 118396/1 and 105315_130558/1). Additional support is provided by NITAS/TMS, Text Mining Services, **Novartis Pharma AG**, Basel, Switzerland.

University of Zurich
Institute of Computational Linguistics

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

NOVARTIS

Contact: Fabio Rinaldi
rinaldi@cl.uzh.ch
fabio@ontogene.org