IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences

Bahar Sateli Marie-Jean Meurs Greg Butler
Justin Powlowski Adrian Tsang René Witte

Concordia University, Montréal, QC, Canada

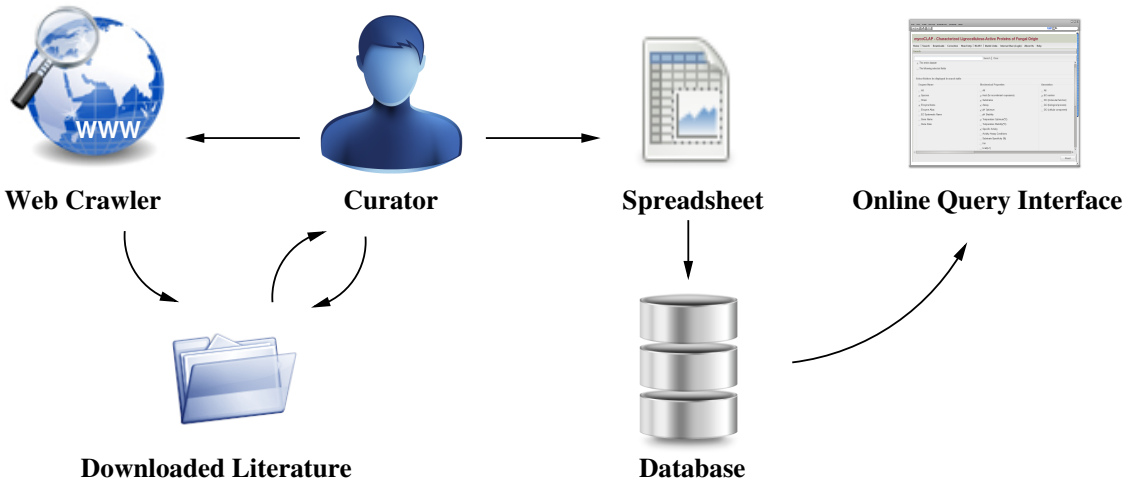
Semantic Software Lab

NETTAB 2012 Nov. 15th, Como, Italy

Outline

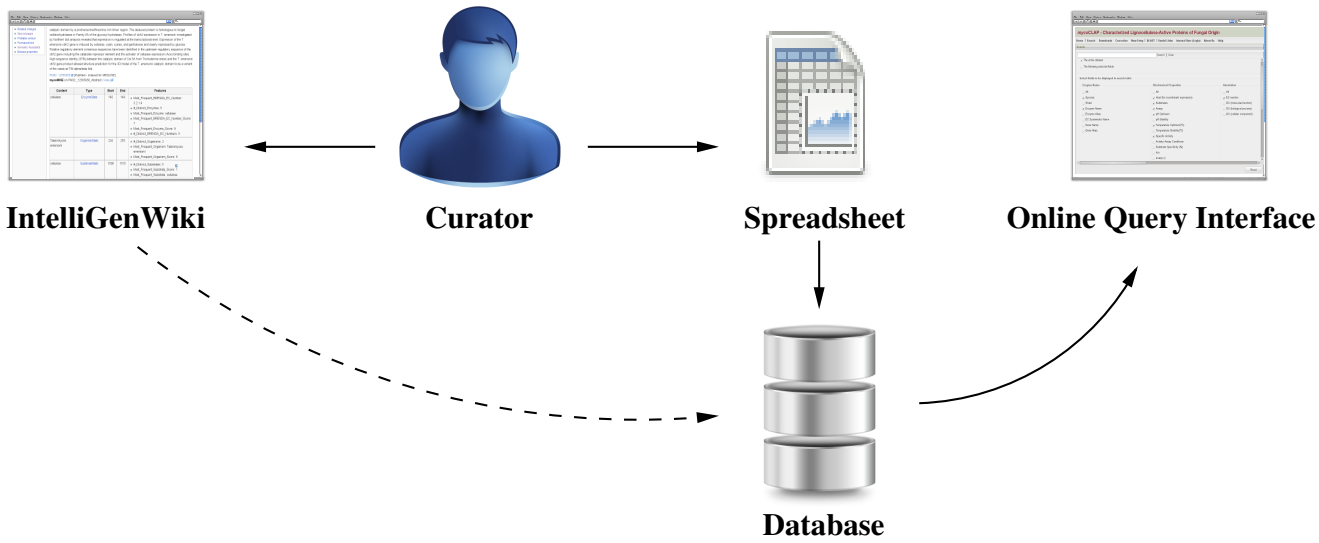
MOTIVATION: **Curation of Biomedical Literature**

- Finding and extracting relevant knowledge from the domain literature
- Manually refining and updating bioinformatics databases



- Manual literature curation is
  - Expensive → requires domain experts
  - Labour-intensive → ever growing amount of scientific publications
  - Error-prone → critical knowledge can be easily missed

Approach: **IntelliGenWiki**

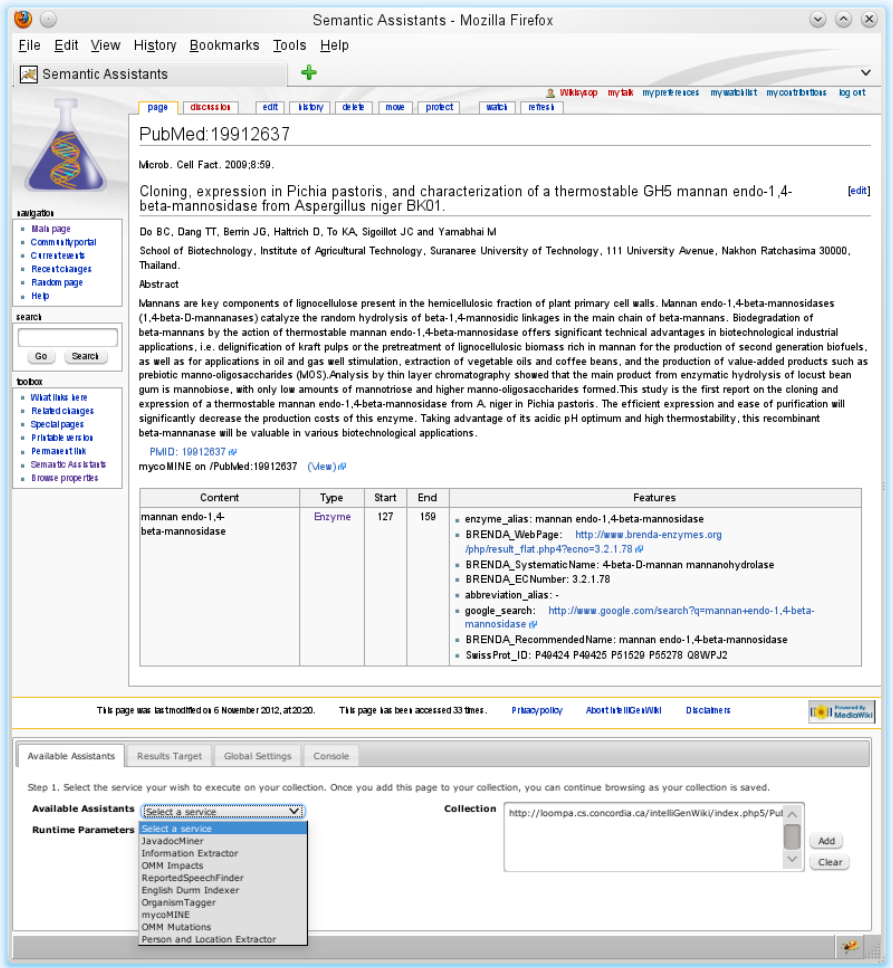


Enhanced Literature Curation Workflow Using IntelliGenWiki

- Text mining techniques integrated within the wiki environment
- Novel Human-AI collaboration patterns
- Producing semantic metadata
- Transform text into knowledge base

Approach: **IntelliGenWiki**

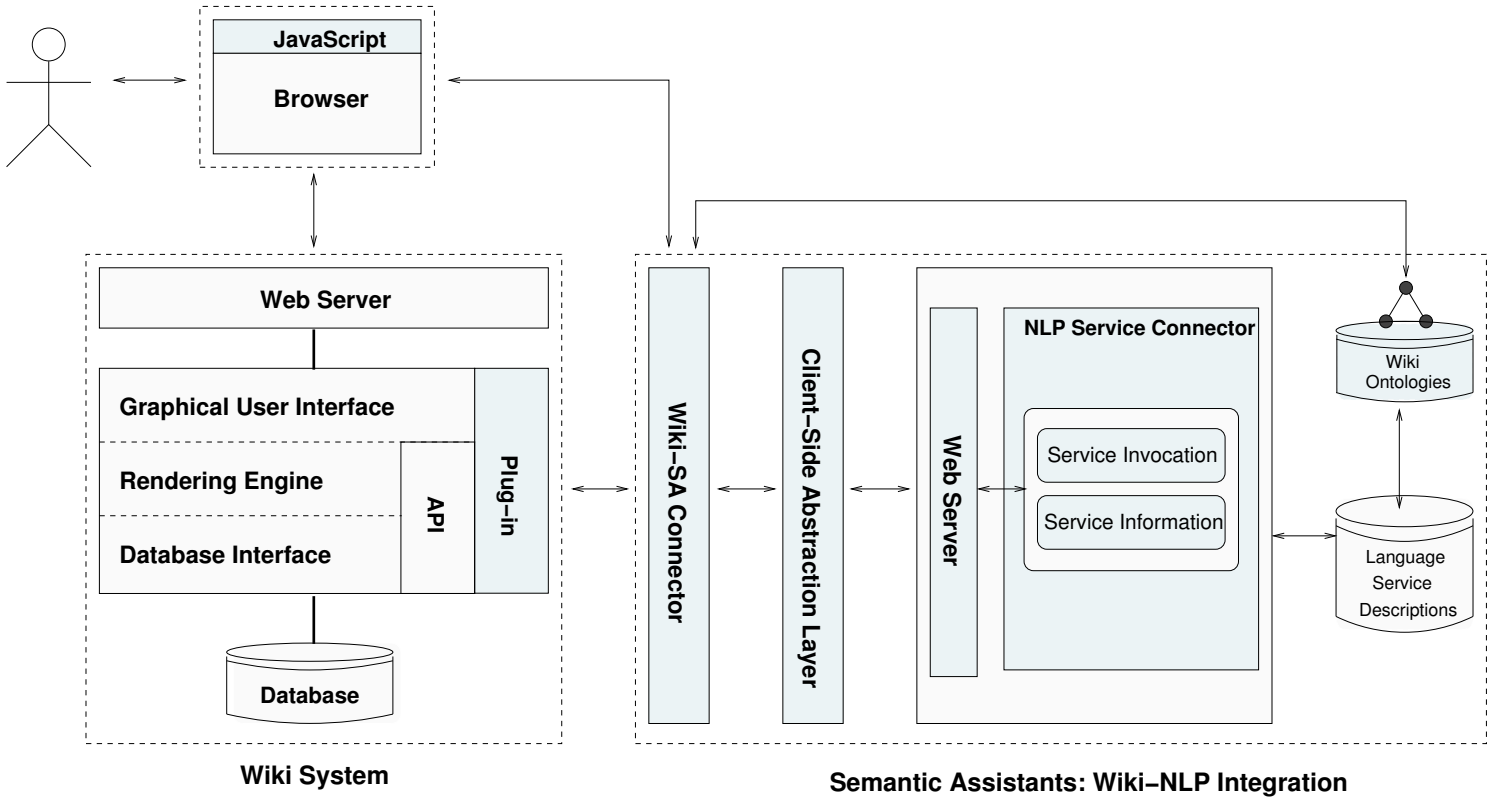
- Adopts the "Wiki" paradigm
  - Accessible via a web browser
  - Simple syntax (markup)
  - Open collaboration
- Based on the MediaWiki engine
  - Open source
  - Highly scalable
  - Extensible: Semantic MediaWiki
- Integrated Text Mining *Assistants*
- Provides semantic capabilities
  - Formalization of knowledge
  - Producing machine-readable content
- Open source software (AGPL3)



IntelliGenWiki User Interface

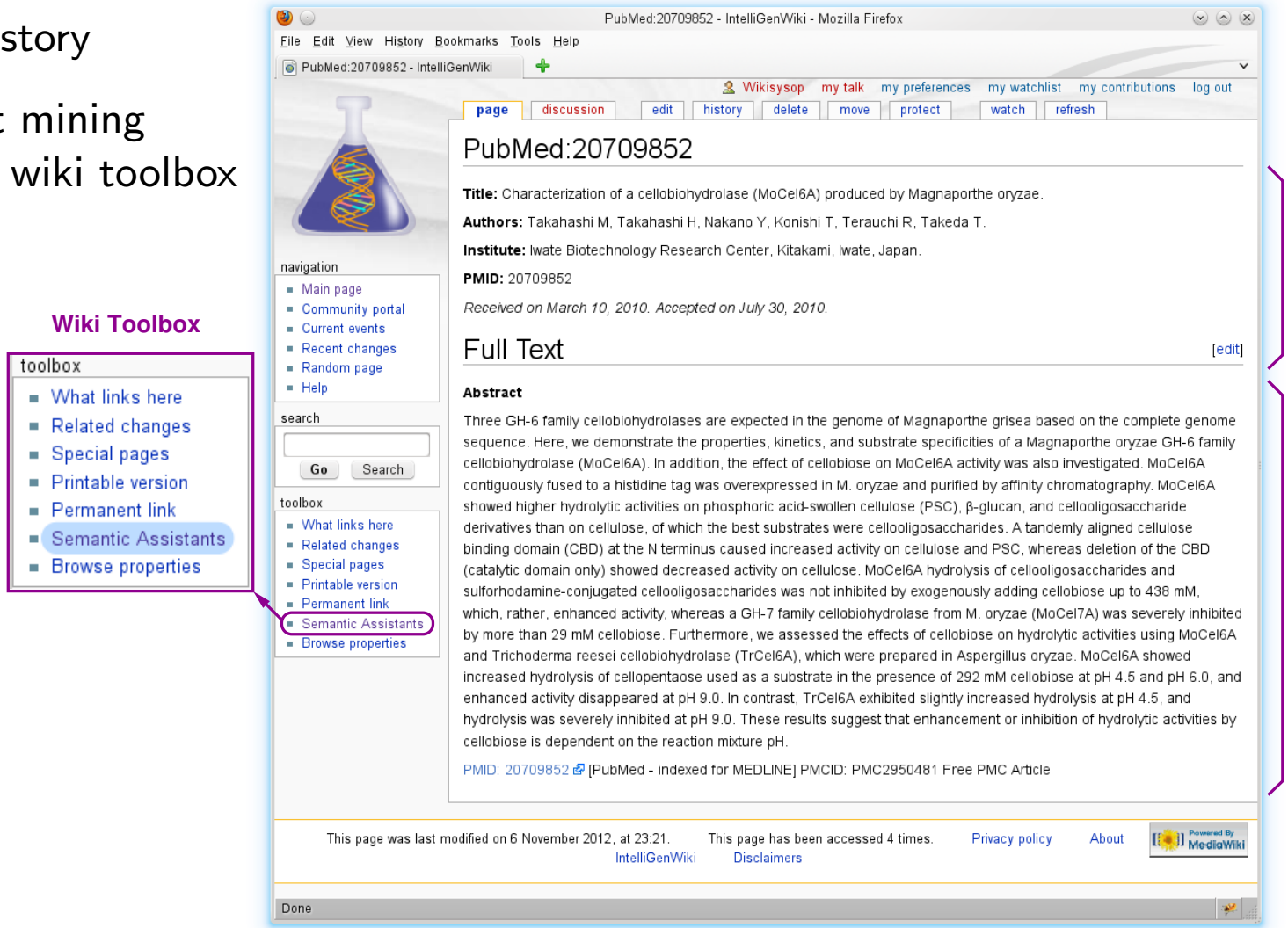
System Overview

- **Front-end:** Semantic MediaWiki
- **Back-end:** Wiki-NLP Integration [Sateli and Witte, 2012]
  - Comprehensive architecture based on the Semantic Assistants Framework [Witte and Gitzinger, 2008]
  - Seamless integration of various NLP capabilities *within* a wiki environment



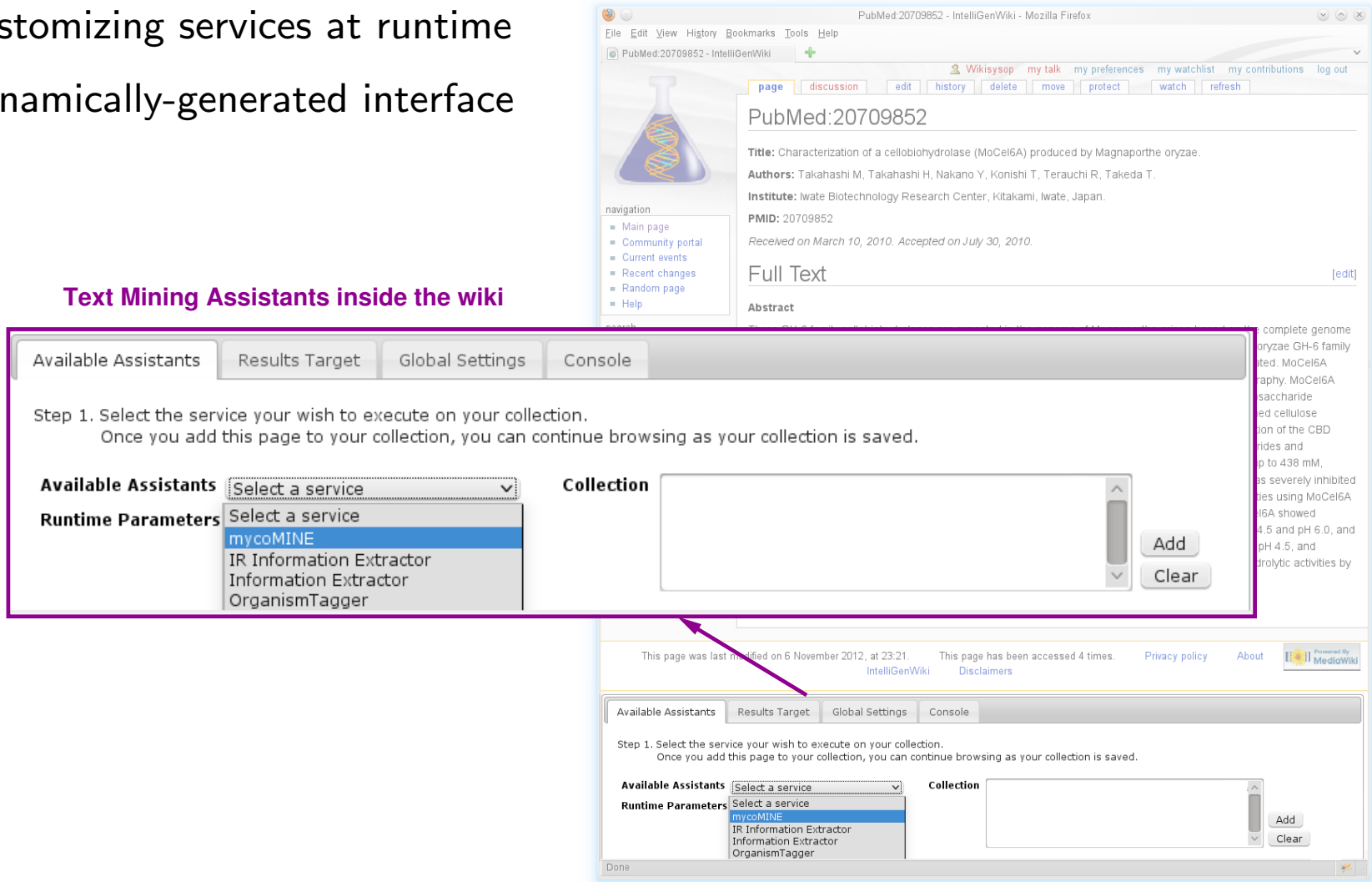
IntelliGenWiki Pages

- Each wiki page corresponds to a literature instance, e.g., abstract of a paper
- Revision History
- Inquire text mining services via wiki toolbox



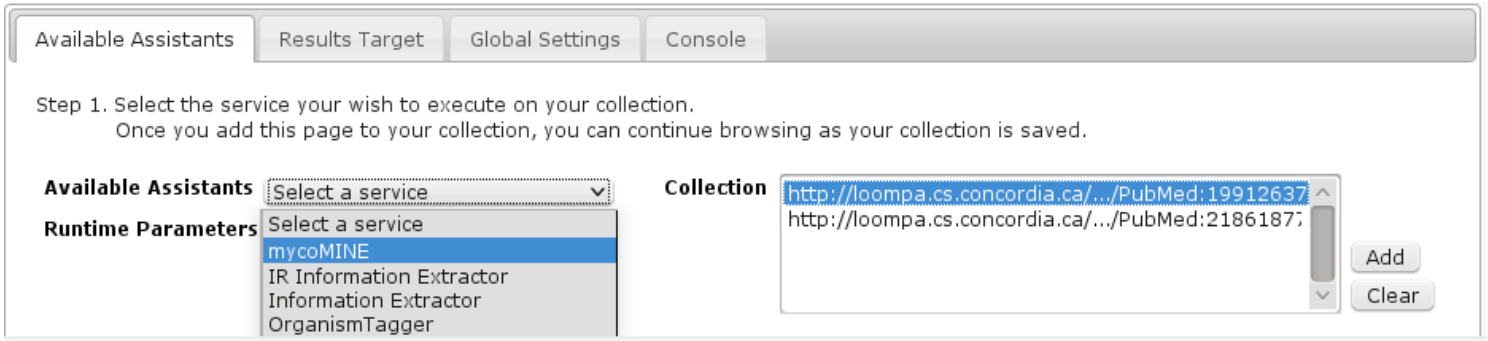
The NLP Interface

- The IntelliGenWiki NLP user interface offers various text mining services
- Customizing services at runtime
- Dynamically-generated interface



NLP Interface features

- Multi-document Analysis



- Flexible handling of results
  - Writing to the same page as the resource
  - Writing to a different page in the wiki
  - Writing to an external wiki

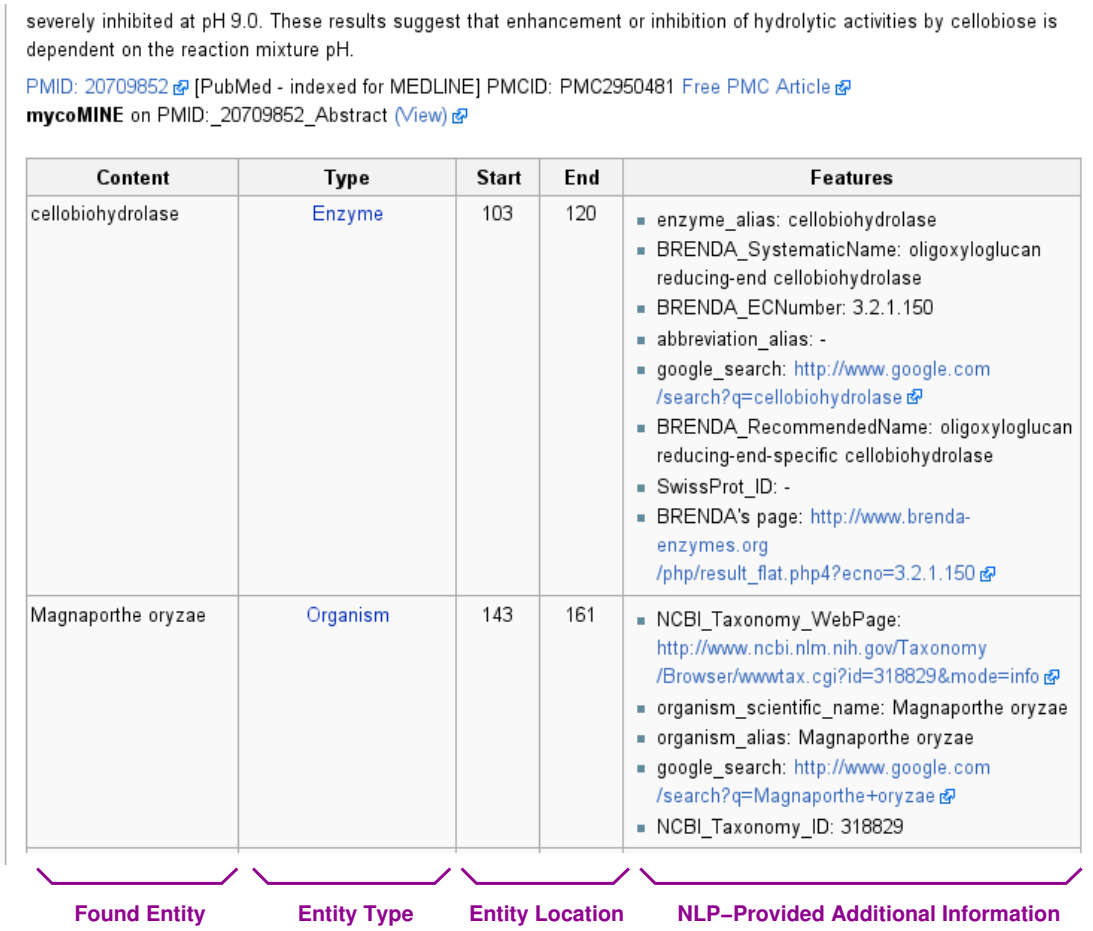
- Dynamic discovery of NLP services

Information Extraction

- Automatically extracting knowledge from text
- Various IE services
  - mycoMINE
  - OrganismTagger
  - Open Mutation Miner
  - ...
- Enrichment of literature content with semantic markup

Example:

[[hasType::Enzyme|cellobiohydrolase]]



severely inhibited at pH 9.0. These results suggest that enhancement or inhibition of hydrolytic activities by cellobiose is dependent on the reaction mixture pH.

PMID: 20709852 [PubMed - indexed for MEDLINE] PMCID: PMC2950481 Free PMC Article
mycoMINE on PMID:_20709852_Abstract (View)

Content	Type	Start	End	Features
cellobiohydrolase	Enzyme	103	120	▪ enzyme_alias: cellobiohydrolase ▪ BRENDA_SystematicName: oligoxyloglucan reducing-end cellobiohydrolase ▪ BRENDA_ECNumber: 3.2.1.150 ▪ abbreviation_alias: - ▪ google_search: http://www.google.com/search?q=cellobiohydrolase ▪ BRENDA_RecommendedName: oligoxyloglucan reducing-end-specific cellobiohydrolase ▪ SwissProt_ID: - ▪ BRENDA's page: http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.2.1.150
Magnaporthe oryzae	Organism	143	161	▪ NCBI_Taxonomy_WebPage: http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=318829&mode=info ▪ organism_scientific_name: Magnaporthe oryzae ▪ organism_alias: Magnaporthe oryzae ▪ google_search: http://www.google.com/search?q=Magnaporthe+oryzae ▪ NCBI_Taxonomy_ID: 318829

Found Entity | Entity Type | Entity Location | NLP–Provided Additional Information

Semantic Entity Retrieval

- Unadorned wikis offer only keyword-based search
- What if we want to *discover* what's contained in the wiki?
  - e.g., *"Which papers in this wiki mention an enzyme entity in their text?"*
- Solution: Querying the semantic metadata in the wiki
  - Search the wiki by semantic properties, e.g., entity *type*, generated by NLP services
  - Using special Semantic MediaWiki markup, called *inline queries*

```
{{#ask: [[hasType::Enzyme]]
   | ?Enzyme = Enzyme Entities Found
   | format = table
   | headers = plain
   | default = No pages found!
   | mainlabel = Page Name
}}
```

Property:Enzyme

Page Name	Enzyme Entities Found
PMID: 20709852	Cellobiohydrolase Cellulases endoglucanases β-glucosidases Invitrogen DNA polymerase

User Study

- Is the integration of text mining assistants in a wiki environment actually effective?
- User study within the Genozymes project context (www.fungalgenomics.ca)
  - **Goal:** Identifying and characterizing fungal enzymes
  - **Dataset:** 30 documents
  - **Users:** 2 expert biocurators
  - **NLP Service:** mycoMINE [Meurs et al, 2012]
  - **Measure:** Time spent on curation
  - **Method:** Comparison against time spent on manual curation
- Results:

Average Curation Time

Abstract Selection		**Full Paper Curation**	
no support	IntelliGenWiki	no support	IntelliGenWiki
1 min.	0.3 min.	37.5 min.	30.6 min.

- Conclusion: IntelliGenWiki was indeed efficient and reduced the paper selection and curation time by almost **70%** and **20%**, respectively.

Conclusion

What you can do now

- Install MediaWiki and Semantic MediaWiki extension
- Download and deploy the Wiki-NLP integration
- Use the existing text mining services in our public server
- Alternatively, setup your own Semantic Assistants services developed based on the GATE framework

What is next

- Cover other tasks, e.g.,
  - Quality assessment
  - Paper recommendation
  - Personalization
- Develop services for automatic import of literature, e.g., from PubMed
- Query the RDF in wiki from external applications

More Information

`http://www.semanticsoftware.info/intelligenwiki`

Acknowledgment

- Funding for this work was provided by NSERC, Genome Canada and Génome Québec.
- Caitlin Murphy and Sherry Wu, biocurators at the Centre for Structural and Functional Genomics (CSFG) at Concordia University, are acknowledged for their participation in the evaluation task.