



# Smarter Wikis through Integrated Natural Language Processing Assistants

Bahar Sateli   René Witte

**Semantic Software Lab**

Concordia University, Montréal, QC, Canada

SMWCon Spring 2013

March 22<sup>nd</sup>, New York City, USA

# Outline

- 1 Introduction
- 2 Wiki-NLP Integration
- 3 Applications
- 4 Conclusion

# I. Wikis as Collaborative Authoring Environments

- ▶ Globalization of Software Development
  - ▶ Teams are spatially and temporally apart
  - ▶ Stakeholders with various backgrounds are involved
  - ▶ Requirements Specifications are written in natural language

page
discussion
edit with form
edit
history
delete
move

## Edit FormProblem: PoorNutrition

**Affects:**

Patient

**Impact:**

Short & long term health:  
 \* Becoming over/under weight.  
 \* Weakened immune system.  
 \* Lack of energy.  
 \* Serious problems (Hypertension, Diabetes, Cholesterol, Gout, Osteoporosis)

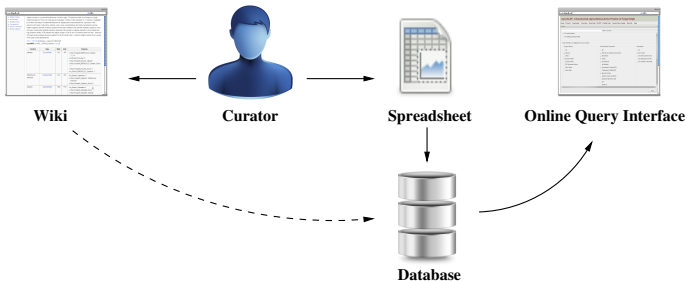
**Successful Solution:**

Combination of:  
 \* Balanced diet.  
 \* Track food intake.  
 \* Take vitamins & supplements.

- ▶ Often, the generated specifications are...
  - ▶ **Ambiguous** → inherited from using natural languages
  - ▶ **Inconsistent** → difficult to maintain manually
  - ▶ **Poor quality** → accounts for 50% of project failures

## II. Wikis as Knowledge Management Platforms

- ▶ Biomedical Literature Curation
  - ▶ 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

# Desiderata

In such contexts, we need a wiki that can:

- ▶ Detect various defects in its content  
*e.g., spelling mistakes, ambiguities, readability issues*
- ▶ Extract entities that are relevant to a user's interest or context  
*e.g., extract all person names mentioned in the wiki*
- ▶ Formally model the knowledge contained inside the wiki  
*e.g., generate Semantic MediaWiki markup from unstructured wiki text*
- ▶ Offer searching for content beyond keyword-based approaches  
*e.g., find all articles containing an enzyme name*
- ▶ Generate its own content  
*e.g., create summaries from long articles*

# Natural Language Processing (NLP)

- ▶ A branch of Artificial Intelligence
  - ▶ uses various techniques to process content written in natural language
- ▶ Multitude of NLP techniques
  - ▶ Named Entity Recognition
  - ▶ Quality Assessment
  - ▶ Summarization
- ▶ Various NLP APIs (e.g. OpenCalais, GATE, ...)

BBC News - Egypt crisis: Clashes in Cairo amid constitution row

Egypt crisis: Clashes in Cairo amid constitution row

Rival protesters have clashed outside the presidential palace in the Egyptian capital, Cairo, as unrest grows over a controversial draft constitution.

Stones were thrown and supporters of President Mohamed Morsi dismantled tents set up by anti-Morsi protesters.

Vice President Mahmoud Mekki has said a referendum on the draft will go ahead on 15 December despite the unrest.

But he indicated that changes could be made after the vote, saying the "door for dialogue" remained open.

He urged critics of the draft document to put their concerns in writing for future discussion.

The critics say the draft was rushed through parliament without proper consultation and that it does not do enough to protect political and religious freedoms and the rights of women.

The draft added to the anger generated by Mr Morsi passing a decree in late November which granted him wide-ranging new powers.

'Breakthrough'

Egyptian Vice-President: 'Door open'

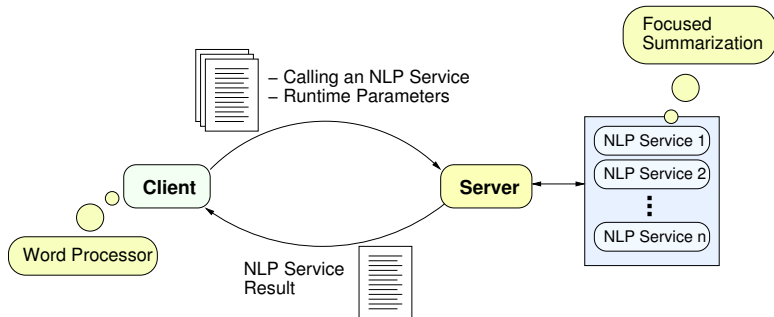
In a news conference broadcast live on state television, Mr Mekki said there was "real political will to pass the current period and respond to the demands of the public"

Entity List:

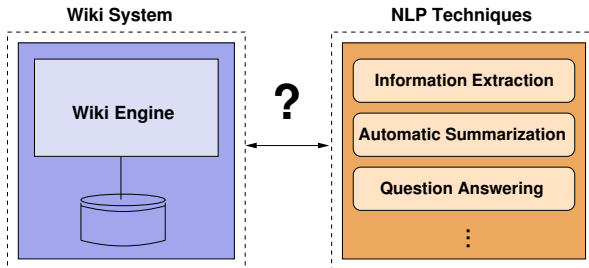
- ☐ Address
- ☒ Date
- ☐ FirstPerson
- ☒ JobTitle
- ☒ Location
- ☐ Lookup
- ☒ Organization
- ☐ Percent
- ☒ Person
- ☐ Sentence
- ☐ SpaceToken
- ☐ Split
- ☐ Title

## Semantic Assistants

- ▶ Service-oriented Architecture (SOA) [4]
- ▶ Publishes various NLP pipelines as W3C Standard Web services
- ▶ Open source framework (<http://www.semanticassistants.com>)



# Problem Statement



## Solve common problems

- ▶ Wikis' Loose Structure
- ▶ Information Overload

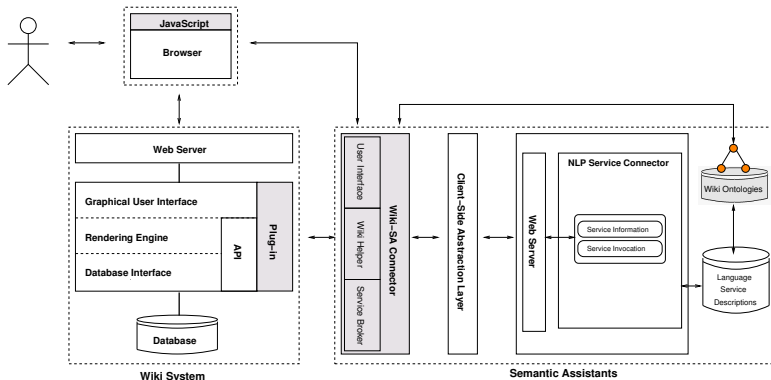
## Introduce new features

- ▶ Enable human-AI interaction
- ▶ Bring semantics to wiki content



## Wiki-NLP Integration Architecture

- ▶ Seamless integration of NLP capabilities within wikis [3]
- ▶ Open-source Software (AGPLv3)
- ▶ General, semantics-based architecture



## The NLP Interface

- ▶ The NLP user interface offers various text mining services
- ▶ Dynamically-generated interface
- ▶ Customizing services at runtime
- ▶ Collection-based Analysis

### Text Mining Assistants inside the wiki

The screenshot displays the Wiki-NLP interface. On the left, a sidebar shows navigation links: Main page, Community portal, Current events, Recent changes, Random page, and Help. The main content area shows the PubMed entry for PubMed:20709852, titled 'Characterization of a cellobiohydrolase (MoCel6A) produced by Magnaporthe oryzae'. Below the article, there is a 'Full Text' link and an 'Abstract' section. On the right, a snippet of the full text is visible, mentioning 'complete genome', 'oryzae GH-5 family', 'MoCel6A', 'saccharide', 'cellulose', 'ion of the CBD', 'ries and', 'p to 430 mM', 'd severely inhibited', 'as using MoCel6A', '6A showed', '4.5 and pH 6.0, and', 'pH 4.5, and', 'olytic activities by'.

Overlaid on the interface is a dialog box titled 'Available Assistants'. It has tabs for 'Available Assistants', 'Results Target', 'Global Settings', and 'Console'. The 'Available Assistants' tab is active, showing a list of services: 'mycoMINE', 'IR Information Extractor', 'Information Extractor', and 'OrganismTagger'. A 'Collection' input field is also present, with 'Add' and 'Clear' buttons. The dialog box includes instructions: 'Step 1. Select the service your wish to execute on your collection. Once you add this page to your collection, you can continue browsing as your collection is saved.'

Below the dialog box, a smaller version of the same interface is shown, with a red arrow pointing from the 'Available Assistants' tab to the 'mycoMINE' service in the list.

## Transformation of Results

- ▶ From the Semantic Assistants server response to wiki markup

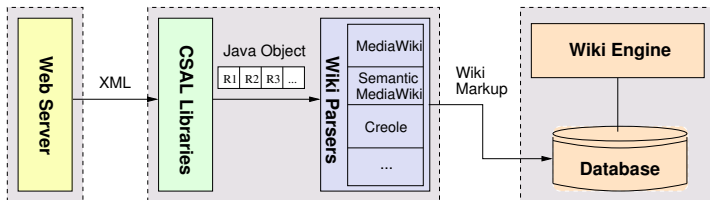
### Sample Service Invocation Response

```
<saResponse>
  <annotation type="Location" annotationSet="Annotation" isBoundless="false">
    <document url="http://localhost/wiki/sample_page">
      <annotationInstance content="Canada" start="16" end="22">
        <feature name="locType" value="country"/>
      </annotationInstance>
    </document>
  </annotation>
</saResponse>
```

#### Semantic Assistants

#### Wiki-SA Connector

#### Wiki System



## Presentation of Results

- Templating Mechanism, i.e., separating data model from its presentation

```

1  { | class="wikitable" style="height:50px"
2  ! width="200" | Content
3  ! width="80" | Type
4  ! width="50" style="text-align: center;" | Start
5  ! width="50" style="text-align: center;" | End
6  ! Features
7  |— valign="top" | {{{content}}} | style="text-align: center;" | ((Property:{{{type}}}|{{{type}}})) |
   style="text-align: center;" | {{{start}}} | style="text-align: center;" | {{{end}}} | {{{
   features}}}
8  |}

```

```

1  {{SemAssist-TableRow| content= Elizabeth Middleton | type=Person | start = 236 | end = 255 | features =
   gender:female}}

```

Content	Type	Start	End	Features
Elizabeth Middleton	Person	236	255	■ gender: female

## 1 Introduction

## 2 Wiki-NLP Integration

## 3 Applications

- Software Requirements Engineering
- Biomedical Literature Curation

## 4 Conclusion

# ReqWiki

- ▶ Semantic MediaWiki customized for collaborative Requirements Engineering (RE)
- ▶ Follows Unified Process (UP) Methodology
- ▶ Structures SRS artifacts using forms and templates

ReqWiki Group 16

navigation

- Main page
- Recent changes
- Help

documents

- Vision
- Use Case
- Supplementary Specification

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Semantic Assistants
- Browse properties

search

Go Search

Vision

Contents [show]

## 1. Introduction

*State purpose of the vision document and describe the purpose of the project/software solution*

## 2. Positioning

*Provide a statement summarizing the problem being solved by this project.*

### 2.1 Problem Statement

Create a stakeholder ← Links to semantic forms  
Create a problem

The problem of	StakeHolder	The Impact of Which is	A Successful Solution would be
Difficulty comparing nutrition values of similar aliment-products	System Users	- Inaccurate or raw-estimated food consumptions - Unable to make balanced food purchases to optimize healthy eating habits. - Poor nutrition	- Clarify nutrition-fact-labels - Food classification

Dynamic tables generated from semantic queries

### 2.2 Product Position Statement

# ReqWiki

- Semantic Forms as data entry point

The screenshot shows the ReqWiki interface. On the left is a sidebar with the 'REQ Wiki' logo, navigation links (Main page, Recent changes, Help), and document links (Vision, Use Case). The main content area has tabs for 'page', 'discussion', 'edit with form' (selected), 'edit', 'history', 'delete', and 'move'. The title is 'Edit FormProblem: PoorNutrition'. Below the title are three sections: 'Affects:' with a text box containing 'Patient'; 'Impact:' with a text box containing a list of health issues; and 'Successful Solution:' with a text box containing a list of solutions.

**REQ Wiki**

navigation

- Main page
- Recent changes
- Help

documents

- Vision
- Use Case

page discussion **edit with form** edit history delete move

## Edit FormProblem: PoorNutrition

**Affects:** Patient

**Impact:** Short & long term health:  
 \* Becoming over/under weight.  
 \* Weakened immune system.  
 \* Lack of energy.  
 \* Serious problems (Hypertension, Diabetes, Cholesterol, Gout, Osteoporosis)

**Successful Solution:** Combination of:  
 \* Balanced diet.  
 \* Track food intake.  
 \* Take vitamins & supplements.

- Embedded traceability links with `{{#ask}}` inline queries

The problem of	StakeHolder	The Impact of Which is	A Successful Solution would be
PoorNutrition	Patient	Short & long term health: <ul style="list-style-type: none"> <li>■ Becoming over/under weight.</li> <li>■ Weakened immune system.</li> <li>■ Lack of energy.</li> <li>■ Serious problems (Hypertension, Diabetes, Cholesterol, Gout, Osteoporosis)</li> </ul>	Combination of: <ul style="list-style-type: none"> <li>■ Balanced diet.</li> <li>■ Track food intake.</li> <li>■ Take vitamins &amp; supplements.</li> </ul>

# ReqWiki

## ► Various NLP services

Available Assistants

Runtime Parameters

- Information Extractor
- Writing Quality
- English Durm Indexer
- Requirements QA Defects
- Requirements QA Stats
- Readability Metric Stats
- ReadabilityMetrics
- Person and Location Extractor

- Detect common defects and suggest solutions, where possible
- Automatically index the SRS documents

## UC/Manage Tasks

<b>Description</b>	The manager receives a customer service request. The manager directs the operation for creating, updating, deleting and querying a task. Some operations use either the automatic or manual task assignation functionality that were defined in the Supplementary Specification document.
<b>Level</b>	user-goal
<b>Primary Actor</b>	<a href="#">A / Manager</a>
<b>StakeHolders</b>	<a href="#">Manager</a> , <a href="#">Senior technician</a> , <a href="#">Junior technician</a>
<b>Pre-Conditions</b>	The manager must be identified and authenticated in the application
<b>Success end condition</b>	The task is created and assigned to the technicians with status Assigned. The tasks is updated and assigned to the technicians with status Assigned. The task is queried. The task is deleted.
<b>Failure end condition</b>	The task is created with status Submitted.
<b>Features</b>	<a href="#">Manage Task</a>

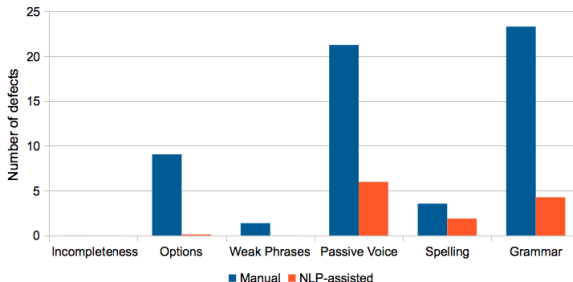
Writing Quality on UC/Manage\_Tasks ([View](#))

Content	Type	Start	End	Features
were defined	<a href="#">AtD</a>	236	248	<ul style="list-style-type: none"> <li>problem: Passive voice</li> <li>suggestion: -</li> </ul>
must be	<a href="#">AtD</a>	434	441	<ul style="list-style-type: none"> <li>problem: Passive voice</li> <li>suggestion: -</li> </ul>
is created	<a href="#">AtD</a>	521	531	<ul style="list-style-type: none"> <li>problem: Passive voice</li> <li>suggestion: -</li> </ul>
The tasks is	<a href="#">AtD</a>	587	599	<ul style="list-style-type: none"> <li>problem: Subject Verb Agreement</li> <li>suggestion: The tasks are, The task is</li> </ul>



## User Study I

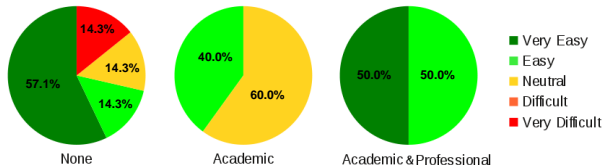
- ▶ Can text mining assistants help to improve requirements specifications? [1]
- ▶ 2 Software Engineering classes at Concordia University, Montréal
  - ▶ **Goal:** Identifying defects in manual vs. NLP-assisted requirements engineering
  - ▶ **NLP Services:** Spell checking, Readability Analysis, Passive Voice Detection, ...
  - ▶ **Measure:** Number of defects found in assignments
  - ▶ **Method:** Comparison against NLP-assisted quality assurance
- ▶ Results:



- ▶ **Conclusion:** ReqWiki NLP capabilities were indeed effective to significantly reduce SRS defects.

## User Study II

- ▶ How much NLP background do users need in order to use semantic capabilities?
- ▶ Same scenario as User Study I
- ▶ Anonymized questionnaire asking participants:
  - ▶ Their proficiency level in NLP
  - ▶ ReqWiki ease-of-use
- ▶ Results:



- ▶ **Conclusion:** Concrete NLP background is not required to make use of sophisticated semantic support provided in ReqWiki.

# IntelliGenWiki

- ▶ An intelligent semantic wiki for life sciences [2]
- ▶ Integrated bio-related NLP services for literature curation
- ▶ Offers basic semantic entity retrieval

**Wiki Toolbox**

- toolbox
  - What links here
  - Related changes
  - Special pages
  - Printable version
  - Permanent link
  - Semantic Assistants
  - Browse properties

**Navigation**

- Main page
- Community portal
- Current events
- Recent changes
- Random page
- Help

**Search**

Go Search

**Toolbox**

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Semantic Assistants
- Browse properties

**PubMed:20709852**

**Title:** Characterization of a cellobiohydrolase (MoCel6A) produced by *Magnaporthe oryzae*.

**Authors:** Takahashi M, Takahashi H, Nakano Y, Konishi T, Terauchi R, Takeda T.

**Institute:** Iwate Biotechnology Research Center, Kitakami, Iwate, Japan.

**PMID:** 20709852

Received on March 10, 2010. Accepted on July 30, 2010.

**Full Text** [edit]

**Abstract**

Three GH-6 family cellobiohydrolases are expected in the genome of *Magnaporthe oryzae* based on the complete genome sequence. Here, we demonstrate the properties, kinetics, and substrate specificities of a *Magnaporthe oryzae* GH-6 family cellobiohydrolase (MoCel6A). In addition, the effect of cellobiose on MoCel6A activity was also investigated. MoCel6A contiguously fused to a histidine tag was overexpressed in *M. oryzae* and purified by affinity chromatography. MoCel6A showed higher hydrolytic activities on phosphoric acid-swollen cellulose (PSC),  $\beta$ -glucan, and cellobiosaccharide derivatives than on cellulose, of which the best substrates were cellobiosaccharides. A tandemly aligned cellulose binding domain (CBD) at the N terminus caused increased activity on cellulose and PSC, whereas deletion of the CBD (catalytic domain only) showed decreased activity on cellulose. MoCel6A hydrolysis of cellobiosaccharides and sulforhodamine-conjugated cellobiosaccharides was not inhibited by exogenously adding cellobiose up to 438 mM, which, rather, enhanced activity, whereas a GH-7 family cellobiohydrolase from *M. oryzae* (MoCel7A) was severely inhibited by more than 29 mM cellobiose. Furthermore, we assessed the effects of cellobiose on hydrolytic activities using MoCel6A and *Trichoderma reesei* cellobiohydrolase (TrCel6A), which were prepared in *Aspergillus oryzae*. MoCel6A showed increased hydrolysis of cellobiose used as a substrate in the presence of 292 mM cellobiose at pH 4.5 and pH 6.0, and enhanced activity disappeared at pH 9.0. In contrast, TrCel6A exhibited slightly increased hydrolysis at pH 4.5, and hydrolysis was 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

This page was last modified on 6 November 2012, at 23:21. This page has been accessed 4 times. Privacy policy About

IntelliGenWiki Disclaimers

Powered by MediaWiki

Paper  
Information

Paper  
Content

## Information Extraction (IE)

- ▶ 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	<ul style="list-style-type: none"> <li>■ enzyme_alias: cellobiohydrolase</li> <li>■ BRENDA_SystematicName: oligoxyloglucan reducing-end cellobiohydrolase</li> <li>■ BRENDA_EcNumber: 3.2.1.150</li> <li>■ abbreviation_alias: -</li> <li>■ google_search: <a href="http://www.google.com/search?q=cellobiohydrolase">http://www.google.com/search?q=cellobiohydrolase</a></li> <li>■ BRENDA_RecommendedName: oligoxyloglucan reducing-end-specific cellobiohydrolase</li> <li>■ SwissProt_ID: -</li> <li>■ BRENDA's page: <a href="http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.2.1.150">http://www.brenda-enzymes.org/php/result_flat.php4?ecno=3.2.1.150</a></li> </ul>
Magnaporthe oryzae	Organism	143	161	<ul style="list-style-type: none"> <li>■ NCBI_Taxonomy_WebPage: <a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=318829&amp;mode=info">http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=318829&amp;mode=info</a></li> <li>■ organism_scientific_name: Magnaporthe oryzae</li> <li>■ organism_alias: Magnaporthe oryzae</li> <li>■ google_search: <a href="http://www.google.com/search?q=Magnaporthe+oryzae">http://www.google.com/search?q=Magnaporthe+oryzae</a></li> <li>■ NCBI_Taxonomy_ID: 318829</li> </ul>

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?"*
- ▶ Exploit the semantic metadata generated by NLP services, e.g., *type* properties
  - ▶ Using Semantic MediaWiki 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 In vitro 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](http://www.fungalgenomics.ca))
  - ▶ **Goal:** Identifying and characterizing fungal enzymes
  - ▶ **Dataset:** 30 documents
  - ▶ **Users:** 2 expert biocurators
  - ▶ **NLP Service:** mycoMINE
  - ▶ **Measure:** Time spent on curation
  - ▶ **Method:** Comparison against time spent on manual curation

Average Curation Time

- ▶ Results:

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.

## Our Contributions

- ▶ Design of a cohesive Wiki-NLP integration architecture
- ▶ Extensible for other wiki engines
- ▶ Allows use of existing text mining techniques in your wiki
- ▶ Create machine-accessible information
- ▶ Performed the first extrinsic evaluation of an NLP integration within wikis
- ▶ Add another party to the wiki community: AI
- ▶ The groundwork for a multitude of new projects

## What you can do now

- ▶ Add NLP capabilities to your wiki for a variety of use cases
  - ▶ Find scenarios in which NLP assistance can be useful, e.g., Summarization
  - ▶ Develop the actual NLP pipelines, e.g., based on GATE<sup>1</sup>
  - ▶ Deploy the pipelines on a Semantic Assistants server
  - ▶ Alternatively, use the existing text mining services in our public server
- ▶ Download and deploy the Wiki-NLP integration
  - ▶ Deploy the Wiki-NLP servlet on a container, e.g., Tomcat or Jetty
  - ▶ Install the Semantic Assistants MediaWiki extension on your wiki
  - ▶ Configure the extension to point to the servlet endpoint

---

<sup>1</sup>General Architecture for Text Engineering, <http://www.gate.ac.uk>



## Related Publications



B. Sateli, E. Angius, S. S. Rajivelu, and R. Witte.

Can Text Mining Assistants Help to Improve Requirements Specifications?

In *Mining Unstructured Data (MUD 2012)*, Kingston, Ontario, Canada, October 17 2012.



B. Sateli, M.-J. Meurs, G. Butler, J. Powlowski, A. Tsang, and R. Witte.

IntelliGenWiki: An Intelligent Semantic Wiki for Life Sciences.

In *NETTAB 2012*, volume 18 (Supplement B), pages 50–52, Como, Italy, 11/2012 2012.  
EMBnet.journal, EMBnet.journal.



B. Sateli and R. Witte.

Natural Language Processing for MediaWiki: The Semantic Assistants Approach.

In *The 8th International Symposium on Wikis and Open Collaboration (WikiSym 2012)*, Linz, Austria, 08/2012 2012. ACM.



R. Witte and T. Gitzinger.

Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients.

In *3rd Asian Semantic Web Conference (ASWC 2008)*, volume 5367 of *LNCS*, pages 360–374, Bangkok, Thailand, Feb. 2–5, 2009 2008. Springer.

## Software Download and More Information

<http://www.semanticsoftware.info>