Smarter Wikis through Integrated Natural Language Processing Assistants

Bahar Sateli René Witte

Semantic Software Lab

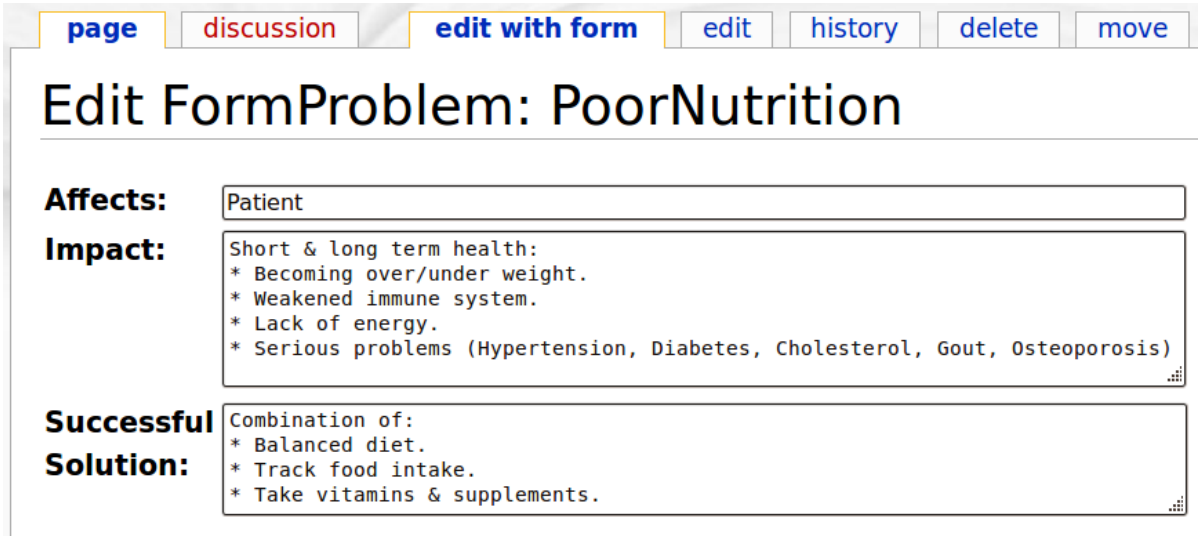
Concordia University, Montréal, QC, Canada

SMWCon Spring 2013 March 22nd, New York City, USA

Outline

I. Wikis as Collaborative Authoring Environments

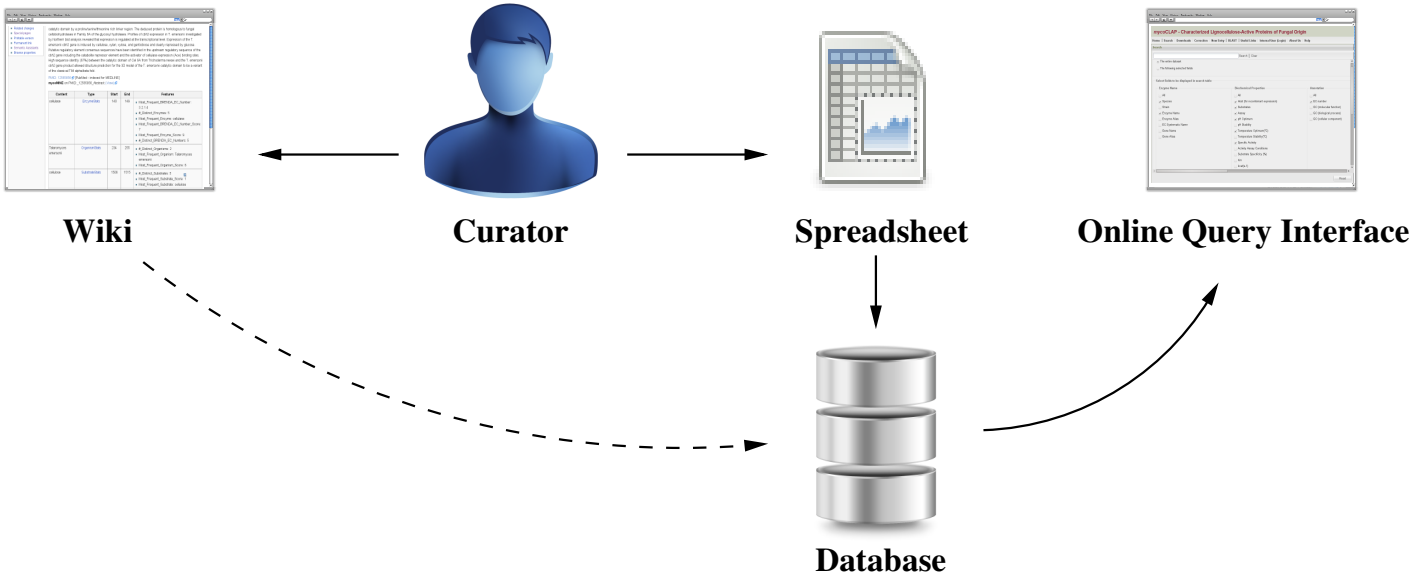
- Globalization of Software Development
  - Teams are spatially and temporally apart
  - Stakeholders with various backgrounds are involved
  - Requirements Specifications are written in natural language



- 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

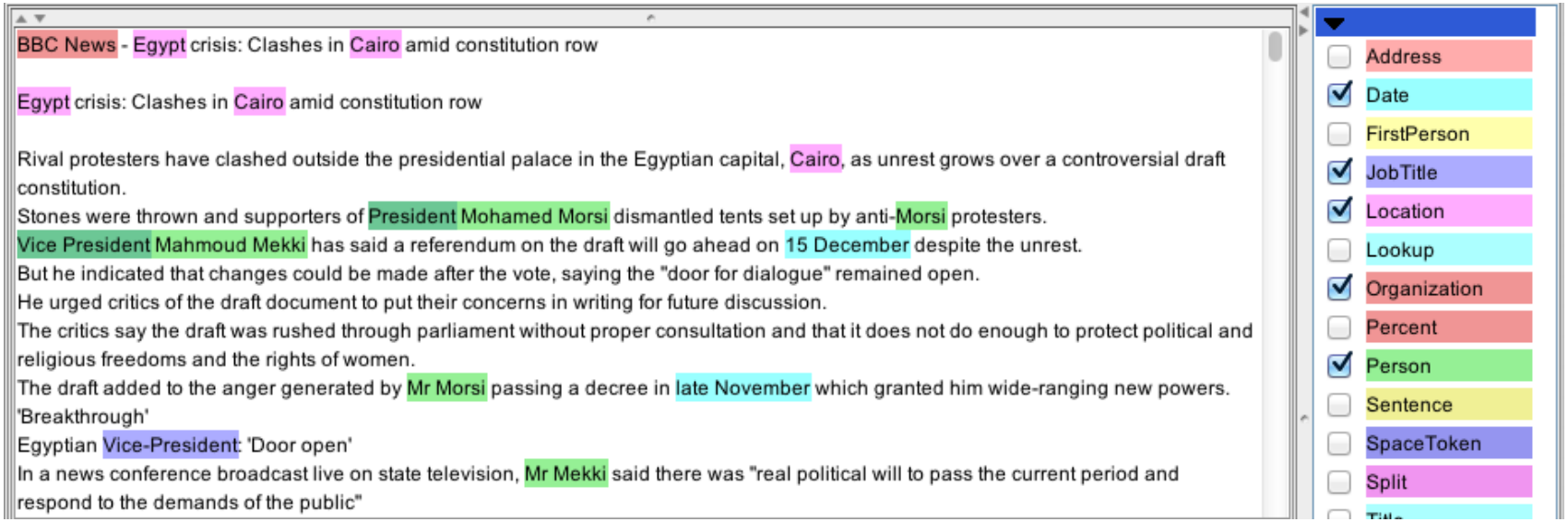
Desiderata

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

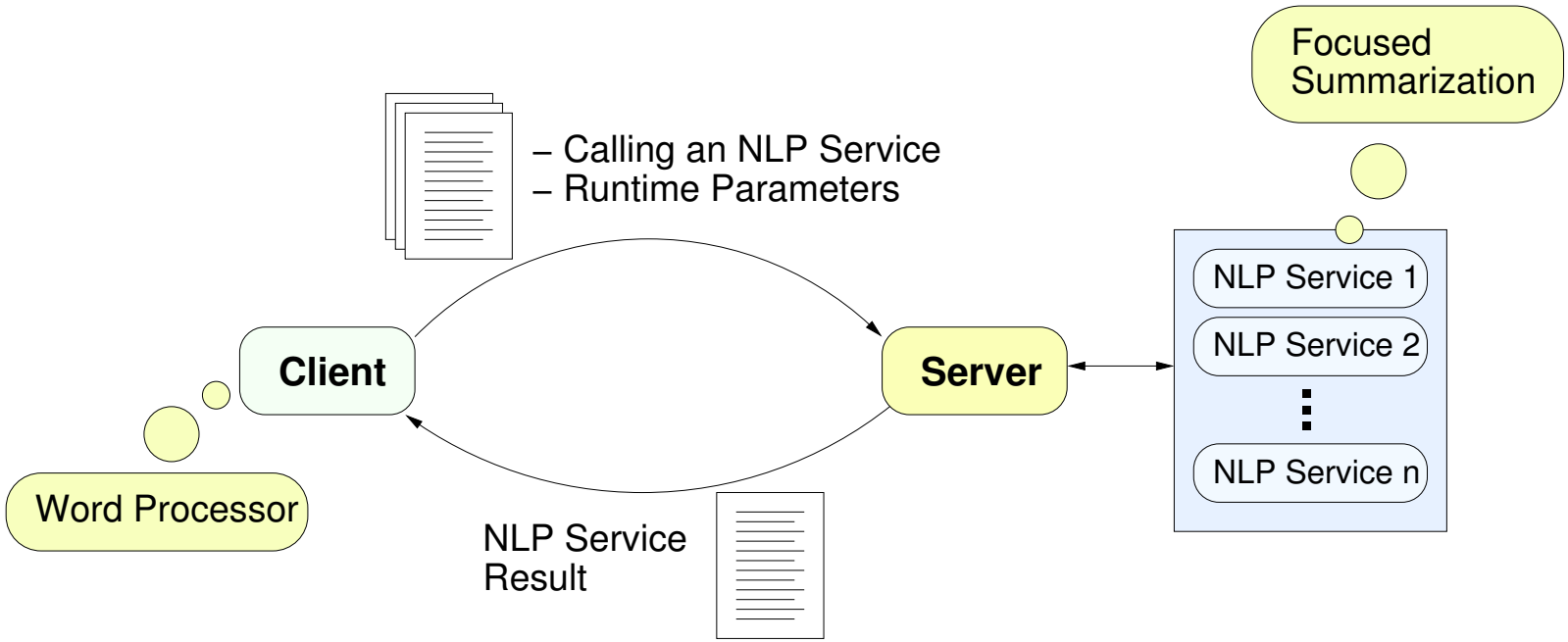
Natural Language Processing (NLP)

- 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, . . .)

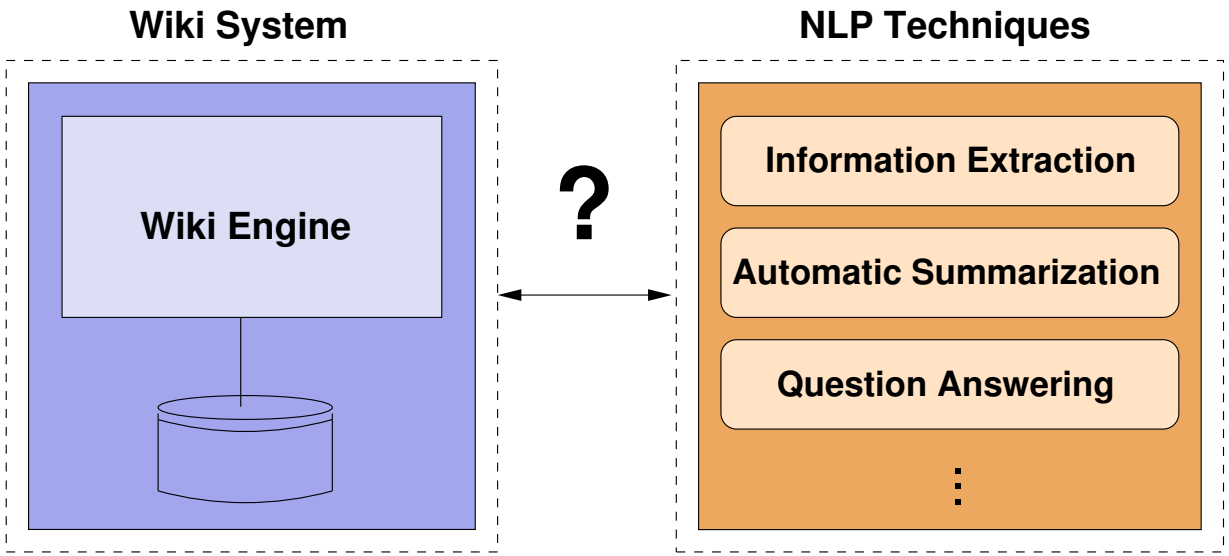


Semantic Assistants

- 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

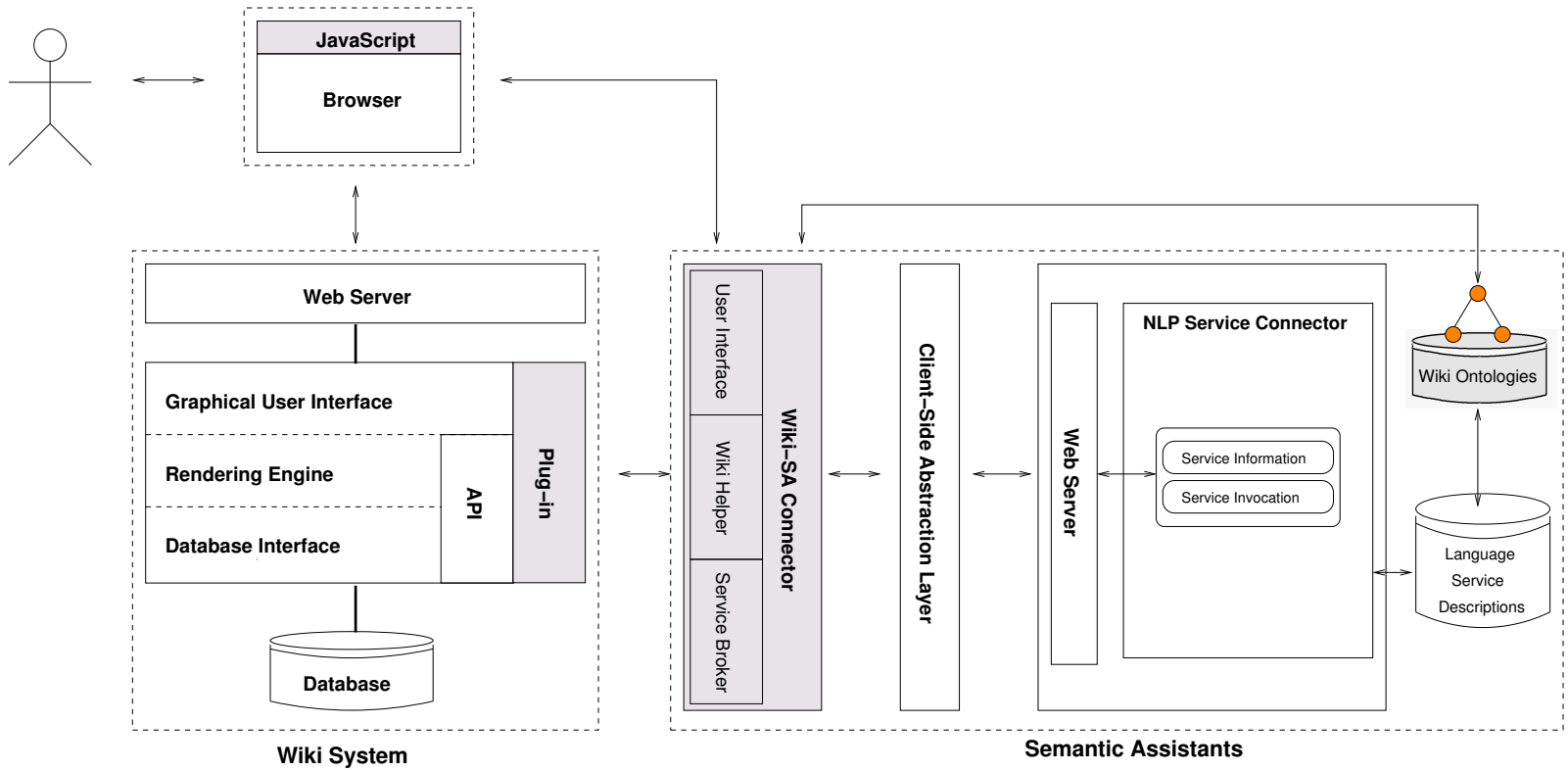
- Wikis' Loose Structure
- Information Overload

Introduce new features

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

Wiki-NLP Integration Architecture

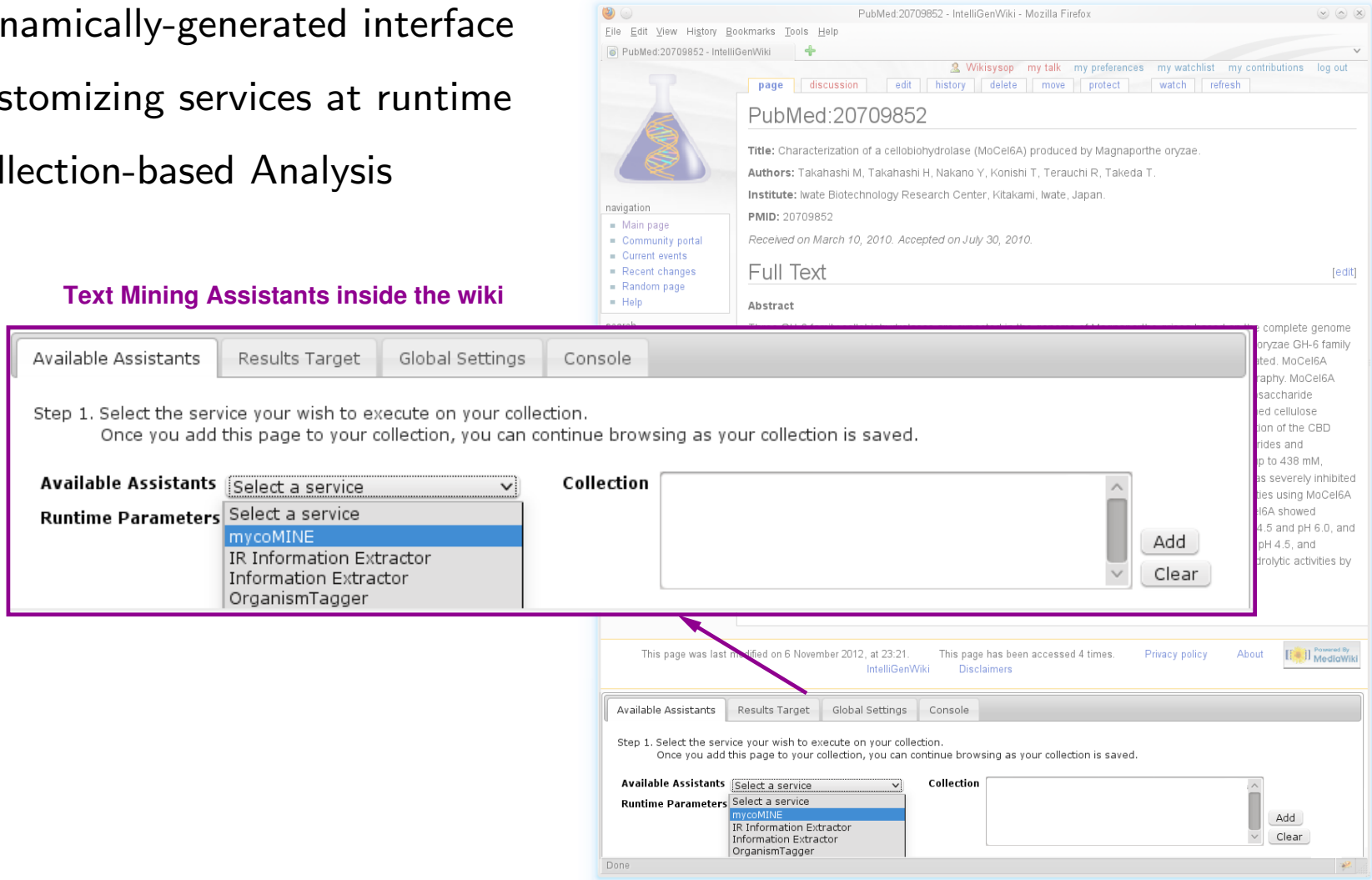
- Seamless integration of NLP capabilities within wikis [3]
- Open-source Software (AGPLv3)
- General, semantics-based architecture



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

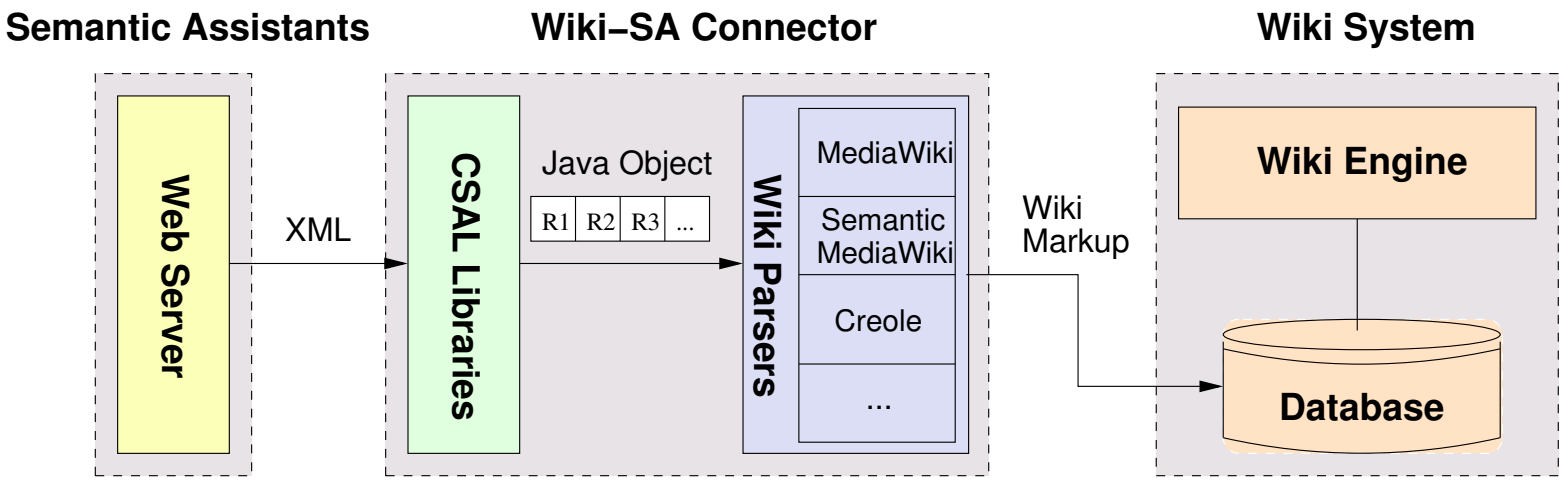


Transformation of Results

- 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>
```



Presentation of Results

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

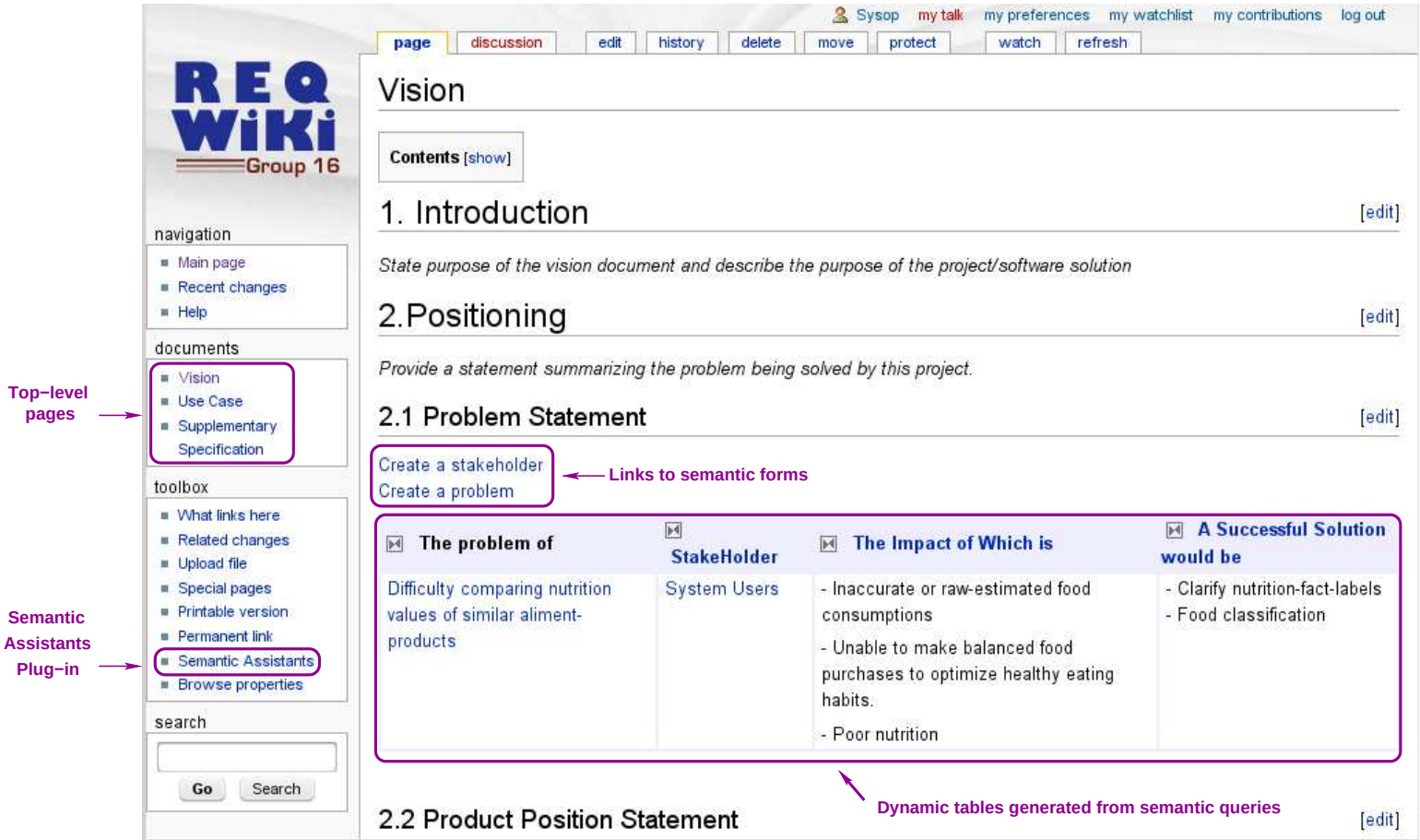
```
1 {| class="wikitable" style="height:50px"
2 ! width="200" | Content
3 ! width="80" | Type
4 ! width="50" style="text-align: center;" | Start
5 ! width="50" style="text-align: center;" | End
6 ! Features
7 |– valign="top" |{{{content}}} | style="text-align: center;" | ((Property:{{{type}}}|{{{type}}}))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

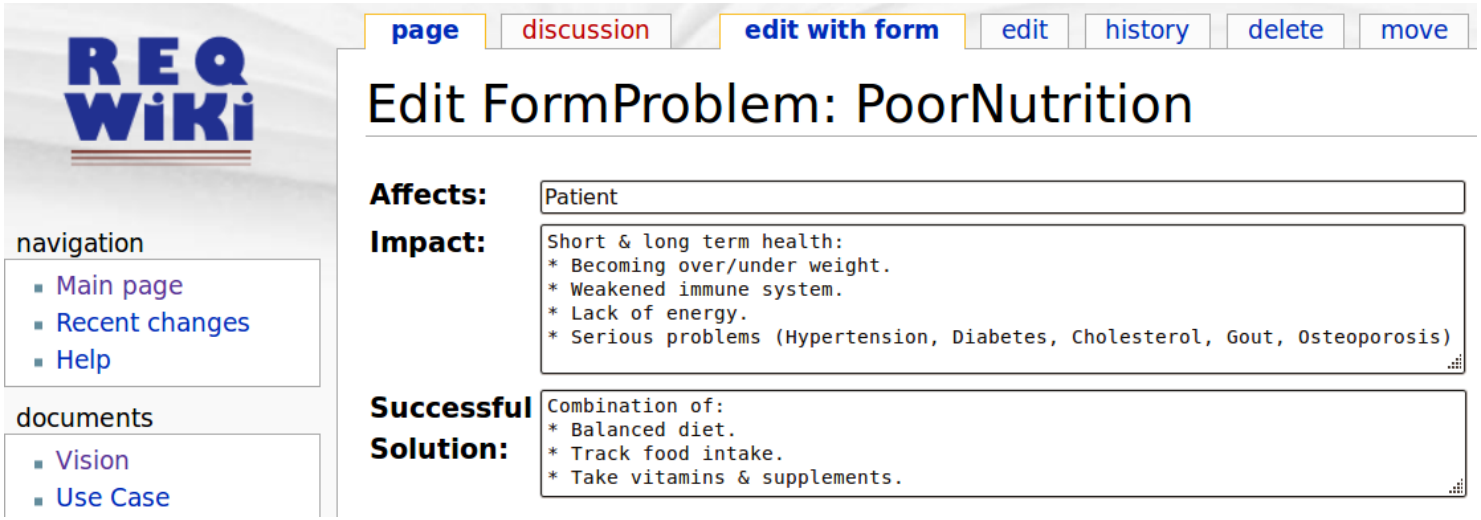
ReqWiki

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



ReqWiki

- Semantic Forms as data entry point

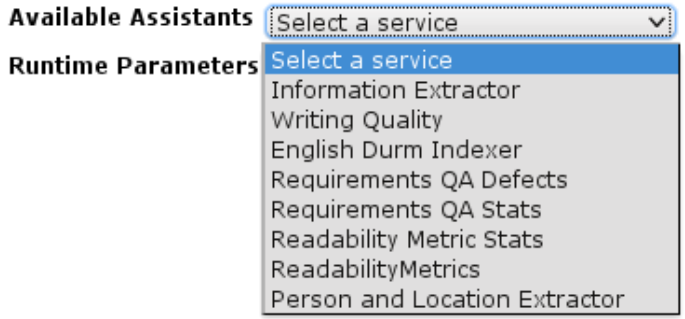


- Embedded traceability links with {{#ask}} inline queries

The problem of	StakeHolder	The Impact of Which is	A Successful Solution would be
PoorNutrition	Patient	Short & long term health: ▪ Becoming over/under weight. ▪ Weakened immune system. ▪ Lack of energy. ▪ Serious problems (Hypertension, Diabetes, Cholesterol, Gout, Osteoporosis)	Combination of: ▪ Balanced diet. ▪ Track food intake. ▪ Take vitamins & supplements.

ReqWiki

- Various NLP services



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

UC/Manage Tasks

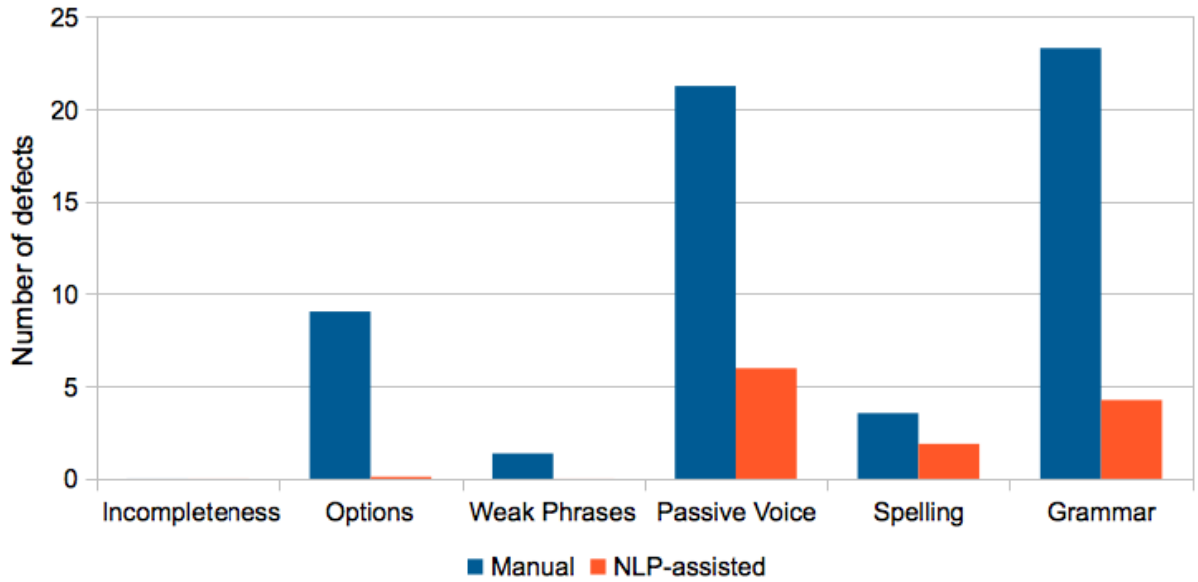
Description	The manager receives a customer service request. The manager directs the operation for creating, updating, deleting and querying a task. Some operations use either the automatic or manual task assignation functionality that were defined in the Supplementary Specification document.
Level	user-goal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

Writing Quality on UC/Manage_Tasks (View)

Content	Type	Start	End	Features
were defined	AtD	236	248	▪ problem: Passive voice ▪ suggestion: -
must be	AtD	434	441	▪ problem: Passive voice ▪ suggestion: -
is created	AtD	521	531	▪ problem: Passive voice ▪ suggestion: -
The tasks is	AtD	587	599	▪ problem: Subject Verb Agreement ▪ suggestion: The tasks are, The task is

User Study I

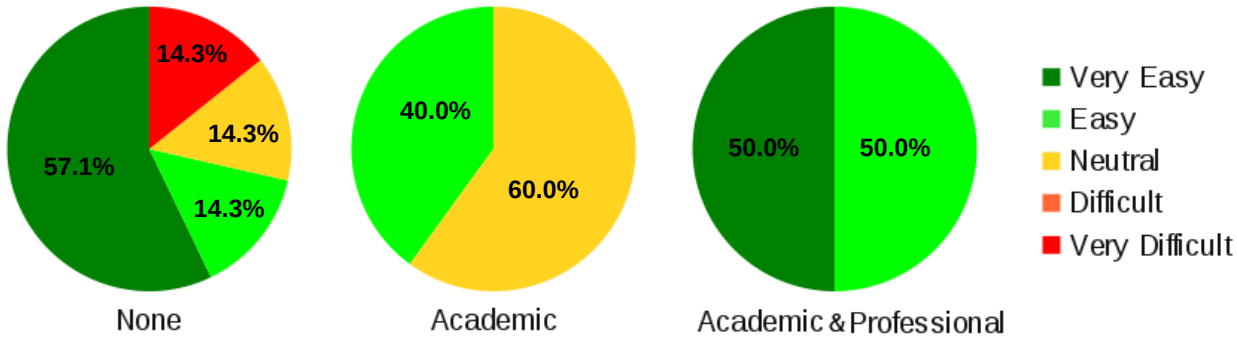
- Can text mining assistants help to improve requirements specifications? [1]
- 2 Software Engineering classes at Concordia University, Montréal
  - **Goal:** Identifying defects in manual vs. NLP-assisted requirements engineering
  - **NLP Services:** Spell checking, Readability Analysis, Passive Voice 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

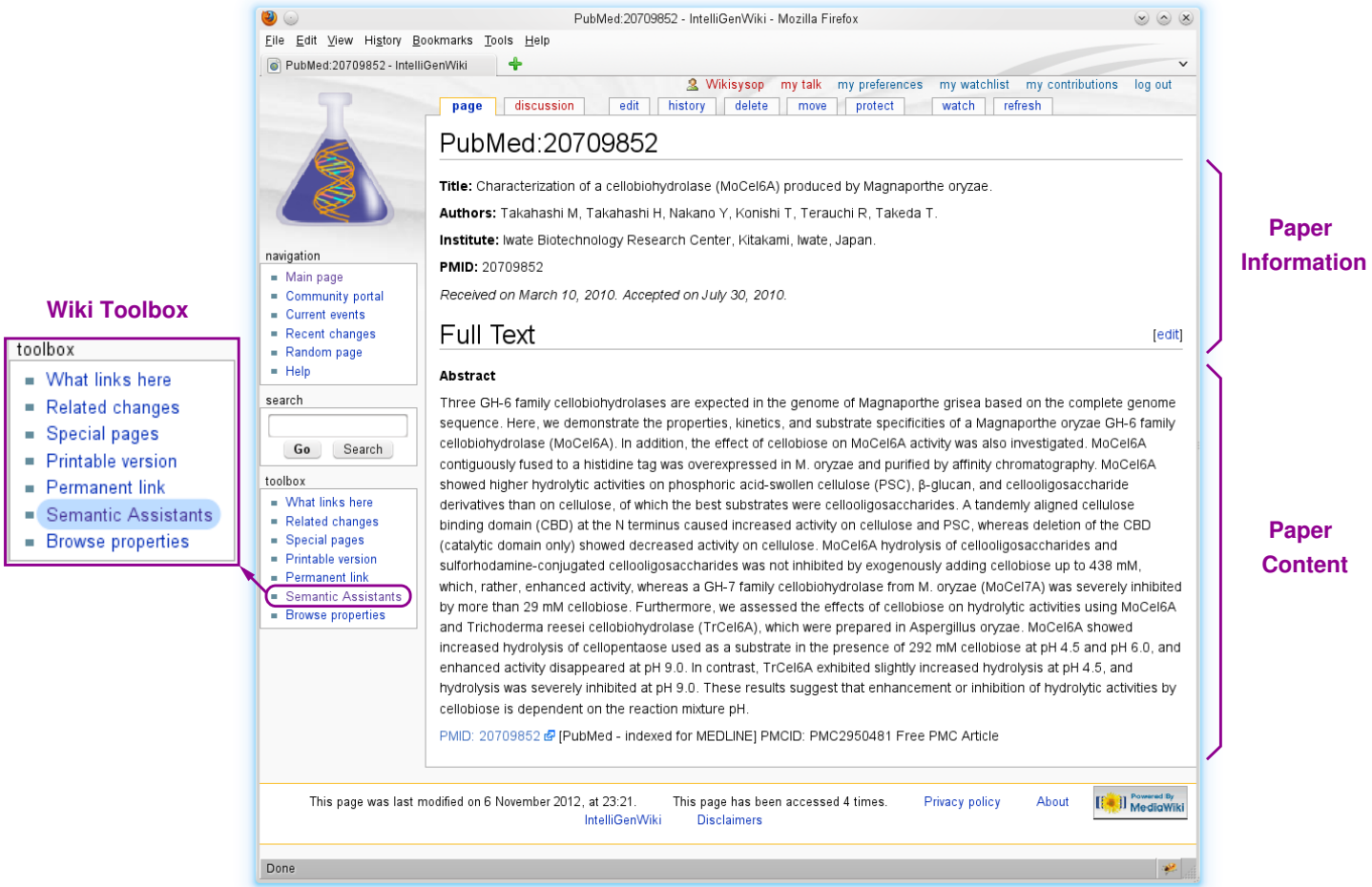
- How much NLP background do users need in order to use semantic capabilities?
- Same scenario as User Study I
- Anonymized questionnaire asking participants:
  - Their proficiency level in NLP
  - ReqWiki ease-of-use
- Results:



- 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



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	▪ 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?"*
- 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
	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
  - **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.

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[1]
  - 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

[1] 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`