

# Semantic Benchmarking Infrastructure for Text Mining: Leverage of Corpora, Ontologies and SPARQL to Evaluate Mutation Text Mining Systems

[Artjom Klein](#), Alexandre Riazanov, Christopher J. O. Baker

Computer Science and Applied Statistics,  
University of New Brunswick, Saint John, Canada.

[Matthew M. Hindle](#)

Synthetic and Systems Biology,  
Edinburgh University,  
Edinburgh, UK.

CSHALS 2013

March 1<sup>st</sup>

Boston, MA

(Mutation) Text Mining  
and  
Motivation of our work

# Sample Mutation Text Mining Tasks

UniProt ID:P22643

“Haloalkane dehalogenase (DhlA) from Xanthobacter autotrophicus GJ10 hydrolyses terminally chlorinated and brominated n-alkanes to the corresponding alcohols.”

GO:0018786

“The A147F mutant showed only a slight reduction of the enzyme activity (Vmax) and while D157P resulted in a larger increase of Km with 1,2-dibromoethane.”

Mutation / Protein / Gene / Organism / Direction / Protein Property / Chemical

- Event: Mutation Impacts (Protein Property + Impact Direction)
- Impact Direction: Positive / Negative
- Linking / Grounding Impact mentions to Mutation mentions in text
- Linking / Grounding Mutation Mentions to Protein mentions in text
- Linking / Grounding Protein Property mentions to GO Molecular Function terms
- Normalizing Mutation mentions (e.g. to HGVS Nomenclature)

# Comparative Evaluation Issues

- **Availability** Often developers only publish their results and not their corpora and systems.
  - **Reproducibility** Developers do not provide instructions on how to reproduce their results, or the instructions require considerable effort from the user.
  - **Interoperability** Evaluation is hindered by the diversity and heterogeneity of formats and annotation schemas of corpora and systems.
  - **Comparability** Corpora vary in qualitative and quantitative characteristics which might significantly affect the performance evaluation results.
  - **Diversity of metrics** Different flavours of precision and recall are used by different system developers.
    - <sup>1</sup>15 different metrics for evaluation of protein mutation extraction systems.
    - Different granularities, e. g., the mutant protein property change: binary outcomes (*has effect* vs *no effect* ) vs. the direction of the effect – e. g., *positive effect* or *negative effect*.
1. Witte R and Baker CJO: **Towards a systematic evaluation of protein mutation extraction systems**. JBCB 2007

# Need: Benchmarking Infrastructure for mutation text-mining

- Mutation grounding system performance accuracy,
  - 0.73, on a homogeneous corpus of 76 documents,
  - only 0.13 on a heterogeneous corpus of larger size.

When the system was re-implemented, the authors encountered another challenge – the evaluation of the new system by comparing it to the state-of-the-art was practically unaffordable, despite the existence of similar systems, due to the lack of consensus benchmarking infrastructure.

- Therefore, we developed an **extensible** and **multi-purpose** benchmarking infrastructure, **based on a consensus corpora and utilities** for the community, in order to make such benchmarking and evaluation easy.

# Our requirement on infrastructure

- **Easy extendable** schema: new corpora integration, schema changes, schema adaption, schema alignment, etc.
- **Easy processing** format which does not require a lot of development of supporting tools
- **Easy implementation** and change of evaluation metrics
- **Search** and query the corpus and text-mining system output results

## 3 aspects of benchmark development

1. Format
2. Schema
3. Metrics

# Our Representation Format: RDF

- Why not XML? - XML is a widely used standard format for corpora annotation and is supported by a large number of tools.
- The processing of complex annotations in XML – parsing, storing, querying, evaluation – is usually virtually impossible with off-the-shelf XML tools.
- Developers need to develop schema-specific parsers and processing scripts and change them each time when the schema is changed or extended.



# RDF/OWL: Extensibility

- Different mutation text mining tasks and all requirements can not be foreseen
- Same data may be used for different tasks

-> The **reusability** and **extensibility** of data are among the main design goals of RDF.

OWL/RDF ontologies are highly extensible data schemas providing:

- easy integration of new corpora with annotation schemas that need not be identical, as long as they are compatible.
- easy merging of data defined modulo one ontology with data modulo another ontology.
- additional alignments between the ontologies can be provided by the annotation providers – corpus curators or text mining system developers.

# RDF/OWL: Tool Availability

- **SPARQL** for calculating system performance metrics as well as for various drill-down searches in the gold standard corpora
- Multiple implementations of **RDF triple stores** for efficient storing and querying of large volumes of annotations
- **RDF and OWL APIs** for multiple programming languages, including Java, C++, Perl and Python
- **OWL reasoners** for data integrity checking

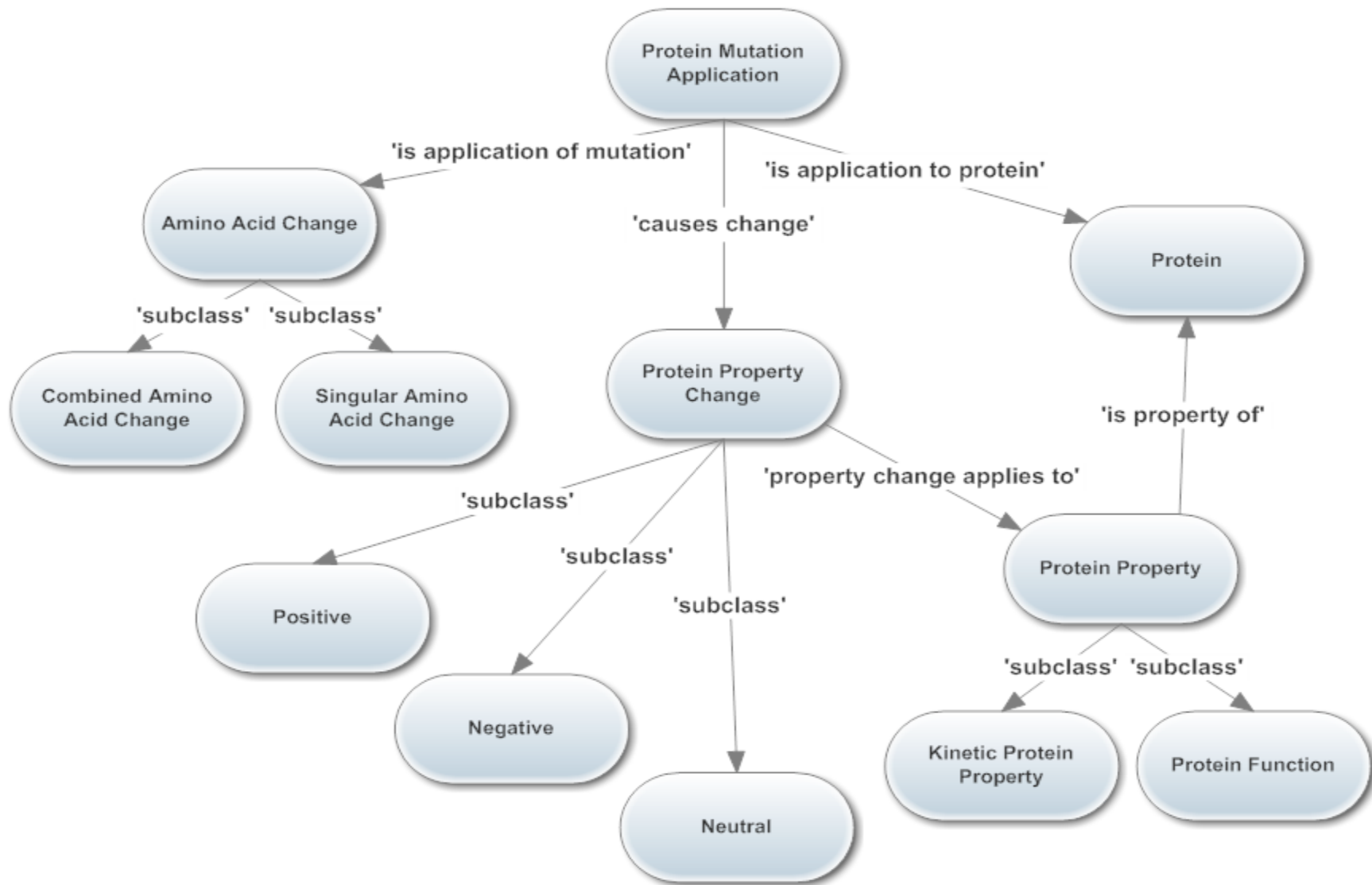
# OWL Ontologies: Schema

- Annotations (Structure): Annotation Ontology (AO)
- Extracted knowledge (Entities, Relations): Domain ontologies (e.g. Mutation Impact Extraction Ontology)

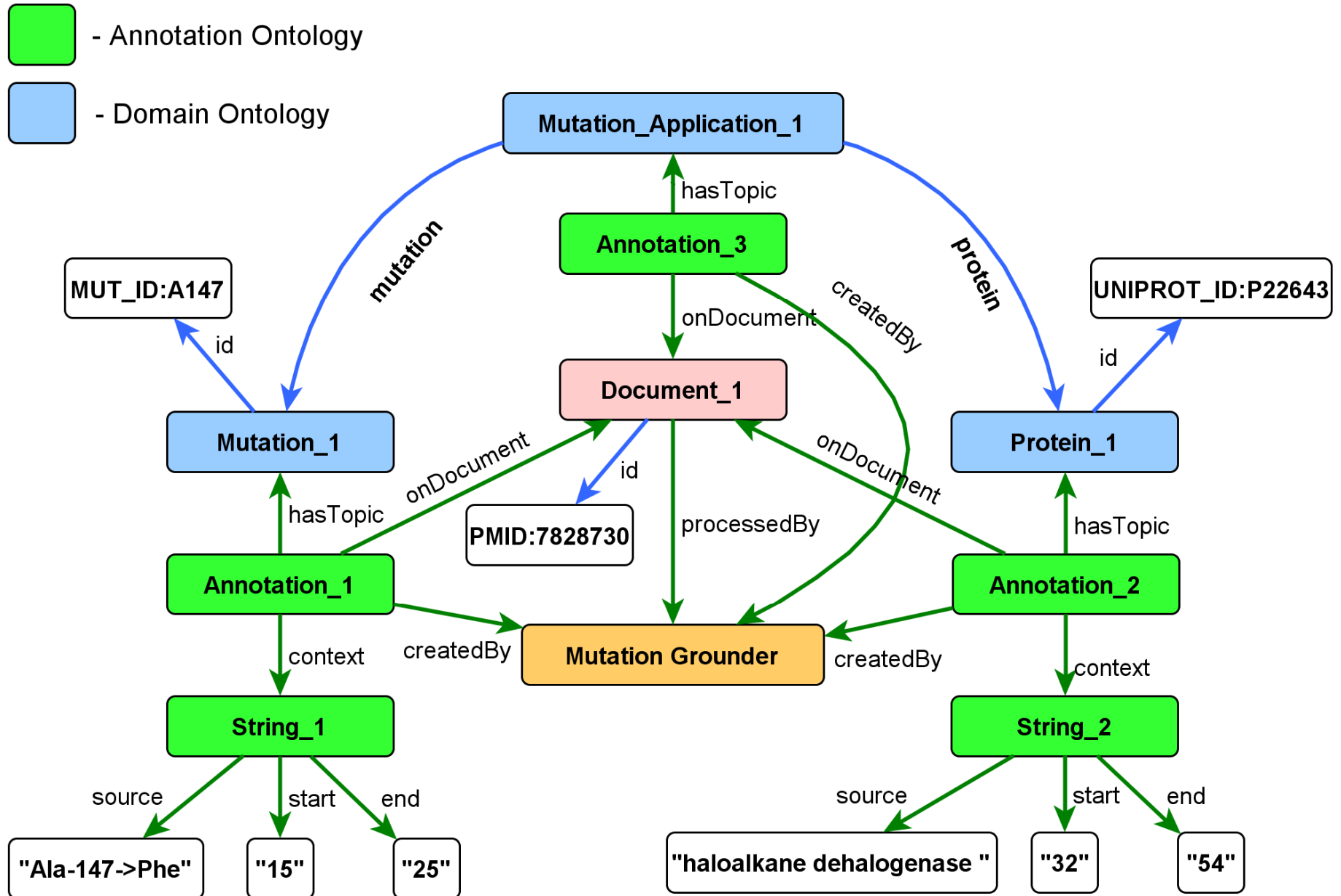
# Annotation Ontology v1.0

- The Annotation Ontology (AO) is an open-source ontology for annotating scientific documents.
- We use AO to model **corpora annotations** as well as **the relevant parts of the text-mining systems results**.
- Two types of annotations:
  - (1) **Document level annotations** are not anchored to specific fragments of text. They annotate a document as a whole, e. g. “mutation A is contained in document B”, “protein A is the topic of document B” .
  - (2) **Text level annotations** are anchored to specific fragments of text, e.g. “Mutation A appears in document B at position C”.

# Mutation Impact Extraction Ontology (MIEO)



# Modelling example



# Benchmarking with SPARQL

- SPARQL is query language for RDF data.
- Create a new SPARQL query or change an existing one is usually easier than create or rewrite some scripts.
- We use named graphs (identified subsets of RDF statements) to separate results coming from different systems or different experiments and gold-standard data: results from different experiments.
- **Precision** - the fraction of correct cases(true positives) over all retrieved cases (true positives + false positives)  
**Recall** - the fraction of correct cases over all cases in the gold standard (true positives + false negatives).
- There are 3 SPARQL queries to calculate:
  - The number of all **correct** cases (query1)
  - All **retrieved** cases (query2)
  - All cases in **gold-standard** (query 3).
- Basically, metrics are calculated by comparing 2 graphs and finding overlaps between them.

# SPARQL: Evaluation of Mutation Grounding

```
SELECT DISTINCT ?pubmed_id ?mut_id ?uniprot_id
WHERE {
  GRAPH <http://example.com/gold-standard.rdf> {
    ?document1 a sio:'article' .
    ?document1 has_identifier ?pubmed_id . # "PMID7750556"
    ?document1 sio:'refers to' ?mutation1 .
    ?mutation1 a mieo:Mutation .
    ?mutation1 has_identifier ?mut_id . # "N30A"
    ?document1 sio:'refers to' ?mutation_application1 .
    ?mutation_application1 a mieo:ProteinMutationApplication .
    ?mutation_application1 mieo:isApplicationOfMutation ?mutation1 .
    ?mutation_application1 mieo:isApplicationOfMutationToProtein ?protein1 .
    ?document1 sio:'refers to' ?protein1 .
    ?protein1 has_identifier ?uniprot_id . # "P22643"
  }
  GRAPH <http://example.com/experiment.rdf> {
    ?document2 a sio:'article' .
    ?document2 has_identifier ?pubmed_id .
    ?document2 sio:'refers to' ?mutation2 .
    ?mutation2 a mieo:Mutation .
    ?mutation2 has_identifier ?mut_id .
    ?document2 sio:'refers to' ?mutation_application2 .
    ?mutation_application2 a mieo:ProteinMutationApplication .
    ?mutation_application2 mieo:isApplicationOfMutation ?mutation2 .
    ?mutation_application2 mieo:isApplicationOfMutationToProtein ?protein2 .
    ?document2 sio:'refers to' ?protein2 .
    ?protein2 has_identifier ?uniprot_id .
  }
}
```



## Results

- Integrated corpora (7 corpora = 282 documents) to support several mutation text-mining tasks
- Implemented several benchmarking queries
  - 2 case studies for concept validation

# Corpora

## Total:

- 282 documents
- 58 protein molecular functions normalized to Gene Ontology (GO)
- 24 host organisms
- >60 UniProt IDs

	Corpus Size	UniProt IDs	Mutations
EnzyMiner	38	49	176
KinMutBase	255	42	624
DHLA	13	4	52
PIK3CA	30	1	169
FGFR3	26	1	175
MEN1	7	1	22

Corpora for mutation grounding

	Corpus Size	Impact mutations	Impacts (mutation, protein property, impact direction)	Impact sentences	Impact sentences grounded to mutations
Omm Impact	40	223	-	2045	1997
Enzyminer	38	172	282	440	440
DHLA	13	52	73	-	-

Corpora for mutation impact extraction tasks

# Implemented metrics

- **Mutation grounding to proteins<sup>1</sup>.** Mutation and protein with corresponding UniProtID that identifies protein sequence.
- **Extraction of mutation impacts on molecular functions of proteins, - *mutation-impact relations*<sup>1</sup>.** Extracted mutation-impact relation to be considered correct, if all the parts had to be correct i.e. the affected protein property, the direction of the impact and the cause mutation. If the protein property was a molecular function, it had to be normalized by grounding to Gene Ontology.
- **Impact sentence recognition<sup>2</sup>.** Sentence or text fragment containing fact about mutation impact on protein property.
- **Grounding impact sentences to mutations<sup>2</sup>.** Accuracy for this task was defined as the fraction of correctly identified impact sentences, grounded to correct mutations, over all correctly identified impact sentences.

1. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJO: **Algorithms and semantic infrastructure for mutation impact extraction and grounding**. BMC Genomics 2010, 11(Suppl 4):S24.

2. Nona Naderi and René Witte, **Automated extraction and semantic analysis of mutation impacts from the biomedical literature**, BMC Genomics 2012, 13(Suppl 4):S10

# Case study 1: Mutation Grounder improvement

	Original Prototype		New System	
	Precision	Recall	Precision	Recall
EnzyMiner (Dev.)	0.31	0.12	0.75	0.72
KinMutBase	0.36	0.14	0.92	0.92
DHLA	0.83	0.73	0.96	0.94
PIK3CA	0.86	0.70	0.98	0.81
FGFR3	0.89	0.66	0.27	0.25
MEN1	0.54	0.32	0.54	0.52
<b>Total w/o EnzyMiner</b>	<b>0.64</b>	<b>0.35</b>	<b>0.82</b>	<b>0.77</b>

On almost all corpora the new system outperforms the original prototype ....

## Case study 2: Comparative evaluation

	Open Mutation Miner <sup>1</sup>	Mutation Impact Extraction System <sup>2</sup>
Mutation recognition	+	+*
Mutation series recognition	+	-
Mutation-protein grounding	-	+
<b>Impact sentence recognition</b>	+	+
<b>Impact sentence grounding to mutation</b>	+	+
<b>Protein property recognition and normalization</b>	+	+
<b>Impact direction recognition</b>	+	+
Physical quantity recognition	+	-
Protein property- Physical quantity grounding	+	-

1. Nona Naderi and René Witte, **Automated extraction and semantic analysis of mutation impacts from the biomedical literature**, *BMC Genomics* 2012, **13**(Suppl 4):S10

2. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJO: **Algorithms and semantic infrastructure for mutation impact extraction and grounding**. *BMC Genomics* 2010, **11**(Suppl 4):S24.

(\*) – uses Mutation Finder

# Comparative evaluation: Results

## Task 1: Mutation Impact Extraction (P/R)

	Enzyminer Corpus	DHLA Corpus
OMM	0.03/0.02	0.34/0.29
MIES	0.21/0.04	0.78/0.44

## Task 2: Impact Sentence Recognition (P/R)

	Enzyminer Corpus	OMM Impact Corpus
OMM	0.59/0.32	0.62/0.51
MIES	0.15/0.10	0.23/0.04

## Task 3: Impact Sentence Grounding (A)

	Enzyminer Corpus	OMM Impact Corpus
OMM	0.59	0.72
MIES	0.84	0.68



# mutation-text-mining

Benchmarking infrastructure for mutation text mining

[Project Home](#)

[Downloads](#)

[Wiki](#)

[Issues](#)


[Source](#)

[Administer](#)

**Summary** [People](#)

## Project Information

 Recommend this on Google

 Starred by 0 users

[Project feeds](#)

### Code license

[Other Open Source](#)

See source for details

### Labels

TextMining, Mutation,  
MutationImpact,  
InformationExtraction,  
Biology, Benchmark,  
SPARQL, Semantics,  
Bioinformatics, Biocuration

## Members

[artiom.unb](#),  
[alexandre.riazanov](#)

### Your role

[Owner](#)

## Centralized Resource for Development, Testing, Evaluation and Comparison of BioNLP Text Mining Systems in Domain of Mutations.

The project is currently under active development. Updates are coming.

## Corpora

- [Available corpora: Overview](#)
- [Demo corpus](#)
- [Easy registration form to get the full version \(242 docs\) corpus](#)

## Experiments

- [Example experiment results](#)

## Evaluation tasks

- [Mutation Grounding Evaluation](#)
- [DEMO Mutation Grounding SPARQL Queries](#)
- [Mutation Impact Extraction Evaluation](#)
- [DEMO Mutation Impact Extraction SPARQL Queries](#)

## Ontologies and modelling

- [Mutation Impact Extraction Ontology](#)

## Utilities

- [Evaluator](#)
- [Sesame Tools](#)

<http://code.google.com/p/mutation-text-mining/>

# Outlook: Future work

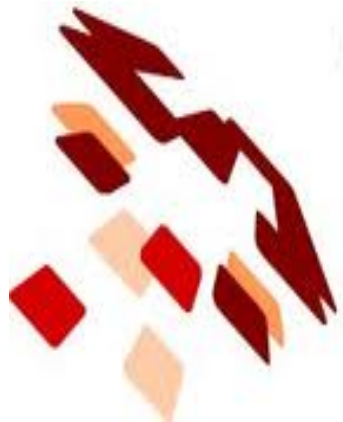
- Further stress-test the infrastructure with **text mining tasks other than mutation grounding and mutation-impact extraction**, and more third-party mutation text mining systems
- Extend the infrastructure to include protein properties other than molecular functions, such as enzyme kinetics, and DNA-level mutations (SNPs)
- Integrate with **BioNLP-SADI Semantic Web Services**
- Implement an interface to load annotated data into one of the following **graphical annotation toolkits** – GATE, UIMA, BART or **DOMEO**



# Acknowledgements



**NSERC**  
**CRSNG**



**C-BRASS**

Canadian Bioinformatics Resources  
As Semantic Services

Powered by SADI

This research was funded in part by the New Brunswick Innovation Foundation, New Brunswick, Canada; the NSERC, Discovery Grant Program, Canada and the Quebec - New Brunswick University Co-operation in Advanced Education – Research Program, Government of New Brunswick, Canada.