The OrganismTagger System

- Corpora
- GATE Components
- Open Mutation Miner
- OrganismTagger
- Semantic Assistants
- Bioinformatics
- NLP
- Text Mining

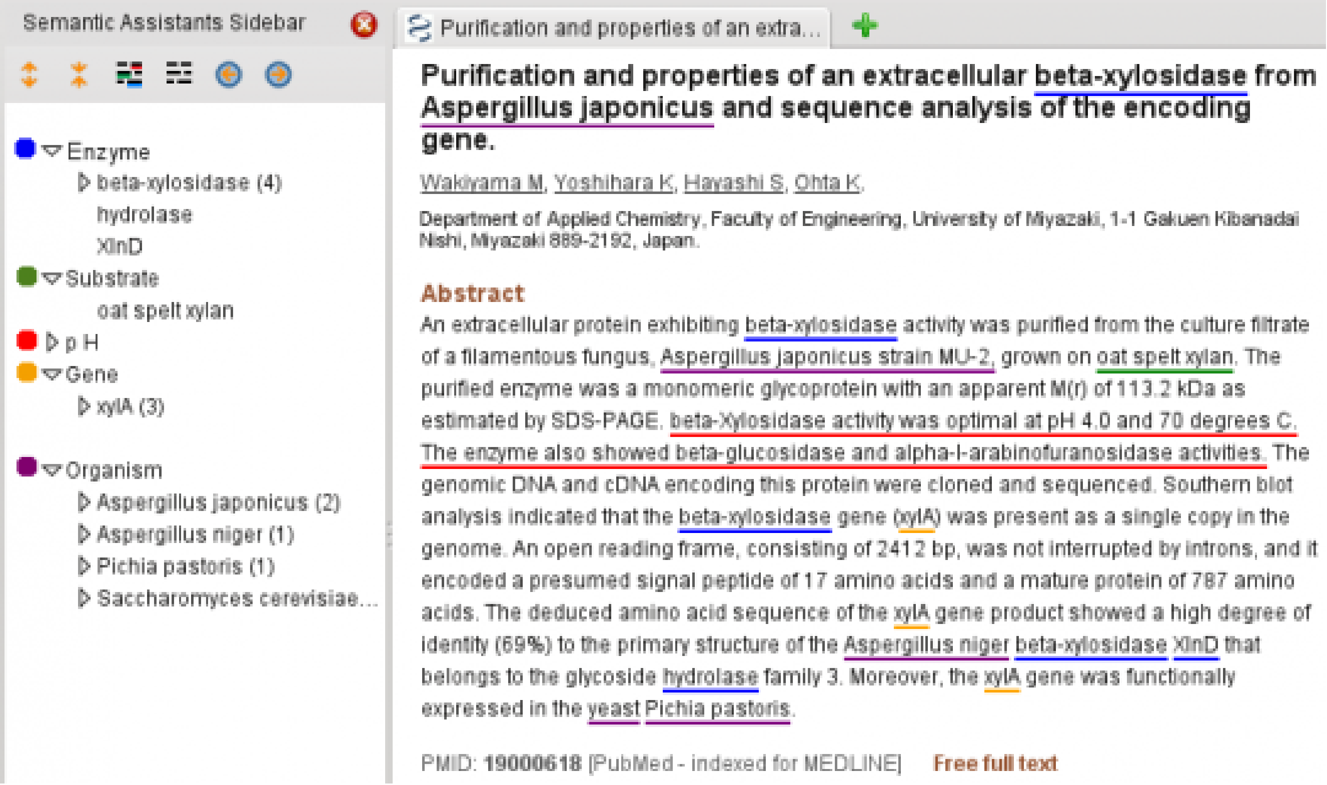
toc_collapse=0; Table of Contents

1. Introduction

The OrganismTagger is a hybrid rule-based/machine-learning system that extracts organism mentions from the biomedical literature, normalizes them to their scientific name, and provides grounding to the NCBI Taxonomy database. Our pipeline provides the flexibility of annotating the species of particular interest to bio-engineers on different corpora, by optionally including detection of common names, acronyms, and strains. The OrganismTagger performance has been evaluated on two manually annotated corpora, *OT* and Linneaus. On the *OT* corpus, the OrganismTagger achieves a precision and recall of 95% and 94% and a grounding accuracy of 97.5%. On the manually annotated corpus of Linneaus-100, the results show a precision and recall of 99% and 97% and grounding with an accuracy of 97.4%. It is published as open source software and described in detail in our publication, Naderi, N., T. Kappler, C. J. O. Baker, and R. Witte, "OrganismTagger: Detection, normalization, and grounding of organism entities in biomedical documents", *Bioinformatics*, vol. 27, no. 19 Oxford University Press, pp. 2721--2729, August 9, 2011.

2. Features

The OrganismTagger has a number of advanced features for species name recognition (for the details, please read the documentation and our paper):



Purification and properties of an extracellular beta-xylosidase from Aspergillus japonicus and sequence analysis of the encoding gene.

Wakiyama M, Yoshihara K, Hayashi S, Ohta K.

Department of Applied Chemistry, Faculty of Engineering, University of Miyazaki, 1-1 Gakuen Kibanadai Nishi, Miyazaki 889-2192, Japan.

Abstract

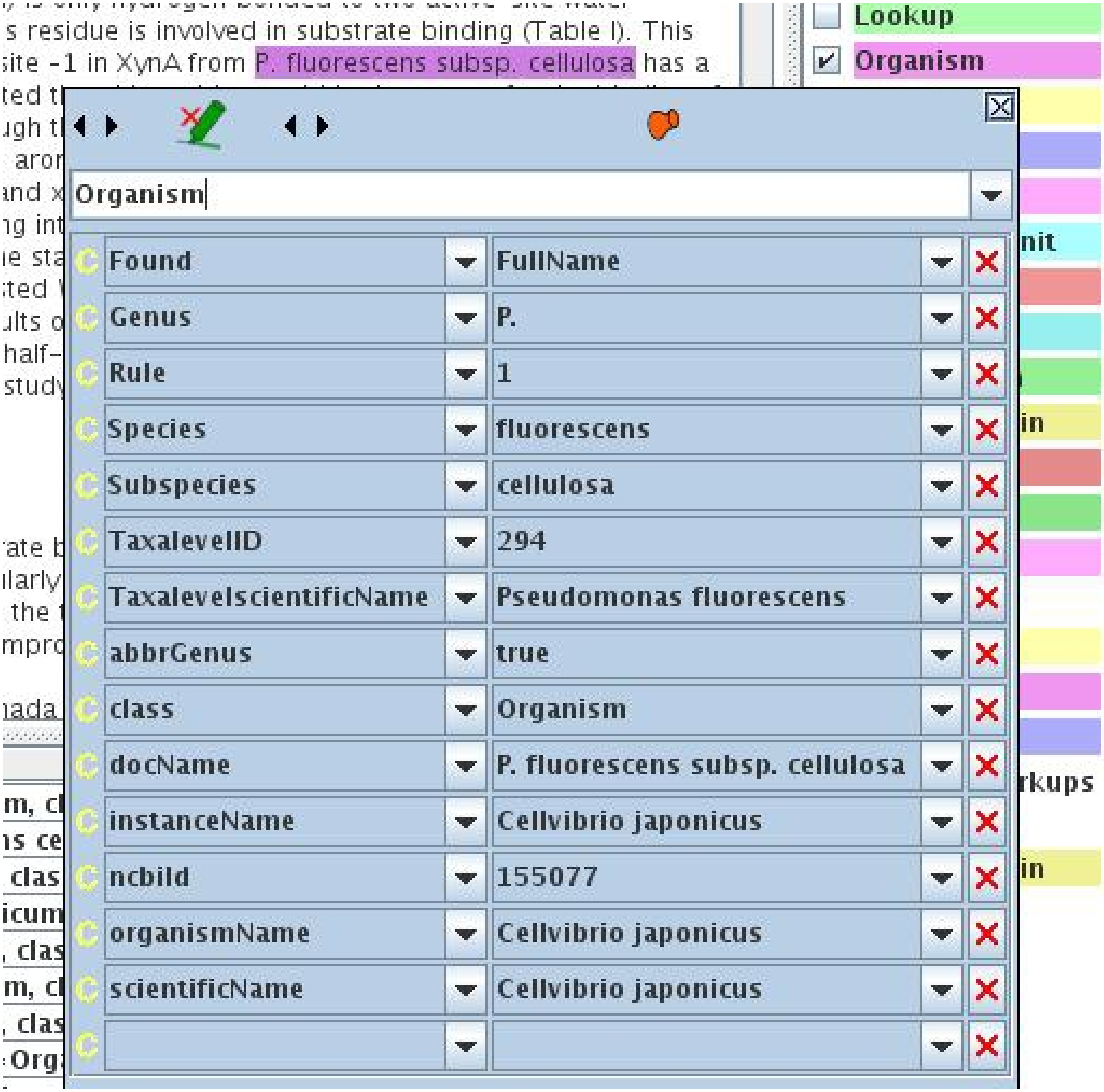
An extracellular protein exhibiting beta-xylosidase activity was purified from the culture filtrate of a filamentous fungus, Aspergillus japonicus strain MU-2, grown on oat spelt xylan. The purified enzyme was a monomeric glycoprotein with an apparent M(r) of 113.2 kDa as estimated by SDS-PAGE. beta-Xylosidase activity was optimal at pH 4.0 and 70 degrees C. The enzyme also showed beta-glucosidase and alpha-l-arabinofuranosidase activities. The genomic DNA and cDNA encoding this protein were cloned and sequenced. Southern blot analysis indicated that the beta-xylosidase gene (xylA) was present as a single copy in the genome. An open reading frame, consisting of 2412 bp, was not interrupted by introns, and it encoded a presumed signal peptide of 17 amino acids and a mature protein of 787 amino acids. The deduced amino acid sequence of the xylA gene product showed a high degree of identity (69%) to the primary structure of the Aspergillus niger beta-xylosidase XlnD that belongs to the glycoside hydrolase family 3. Moreover, the xylA gene was functionally expressed in the yeast Pichia pastoris.

PMID: 19000618 [PubMed - indexed for MEDLINE]   Free full text

OrganismTagger annotations displayed in Firefox

- **Named Entity Detection:** each occurrence of an organism (species name) is semantically annotated as an *Organism*, with additional features showing the *genus* and *species* parts.
- **Normalization:** organism names can also be abbreviated or appear as acronyms, both of which are recognized as well. Additionally, each detected mention is *normalized* by adding the full, scientific name as a feature. The OrganismTagger includes a novel normalization heuristic that can find the full form even if it does not appear in the document.
- **Grounding:** each recognized and normalized mention additionally receives an *NcbiId* feature that links to its entry in the NCBI Taxonomy Database — this is especially useful for disambiguating other biological entities in a document. An additional *url* feature directly links to the organism's web page on NCBI Taxonomy.
- **Strain Recognition:** using a novel machine learning approach, the OrganismTagger also detects and annotates strain-level information.
- **Updateability:** the OrganismTagger resources (gazetteering lists, RDF files, etc.) are automatically generated from a download of the NCBI database. All scripts are included to allow end users to update (and customize) their own installation.
- **Web Service:** you can run the OrganismTagger as a Web service, using standard SOAP requests; an OWL service description for the Semantic Assistants architecture is included with the distribution. We also provide a demo Web service in case you want to try it out before running your own installation.
- **Desktop Clients:** the Semantic Assistants also allow you to run the OrganismTagger interactively using a number of desktop plug-ins, like for Firefox or OpenOffice.
- **Training Data and Corpora:** the distribution includes our manual annotations for evaluation as well as the training data and configuration files for the machine learning algorithm in case you want to experiment with improving the performance.
- **Customization:** the OrganismTagger is highly customizable, for details on run-time parameters please refer to the user's guide. The default configuration, as included in the demo pipeline, provides generally excellent performance.
- **Scalability:** based on the General Architecture for Text Engineering (GATE), you can easily batch-process large amounts of documents by deploying the OrganismTagger with the GATECloud Paralleliser. On a standard quad-core desktop processor, OrganismTagger can process roughly 7 full-text papers/second (i.e., 100 full-text papers in ~14 seconds).

3. Result Annotations



s residue is involved in substrate binding (Table I). This
site -1 in XynA from P. fluorescens subsp. cellulosa has a
□ Lookup
☑ Organism
nit
in
rkups
in

Organism

Found	FullName
Genus	P.
Rule	1
Species	fluorescens
Subspecies	cellulosa
TaxalevelID	294
TaxalevelscientificName	Pseudomonas fluorescens
abbrGenus	true
class	Organism
docName	P. fluorescens subsp. cellulosa
instanceName	Cellvibrio japonicus
ncbiId	155077
organismName	Cellvibrio japonicus
scientificName	Cellvibrio japonicus

OrganismTagger annotation displayed in GATE Developer

The output of the OrganismTagger is a semantic annotation of the textual entities representing organisms, including the detected information from normalization and grounding. The screen shot shows the organism *P. fluorescens subsp. cellulosa* with its annotations displayed in the GATE Developer GUI. The genus, species and subspecies parts, as well as the scientific name are based on the NCBI Taxonomy database. You can also see the *normalization* of the abbreviation and the *grounding* of the detected organism to the NCBI Taxonomy database. NCBI IDs and the scientific names for both the taxa level and full form of the organism are provided. In detail, the features of an organism annotation are:

- **Found:** This feature is added through normalization: If the search heuristic based on the non-abbreviated form in the document is successful, this feature is set to *FullName*, indicating that the non-abbreviated form is available in the document. If the first heuristic fails, and the abbreviated mention is resolved through the second heuristic, it is set to *Genus*.
- **Genus:** Gives the *genus* part of the organism as it appears in the text (either the abbreviated form or the non-abbreviated

one).

- **Rule:** Specifies which particular rule detected the organism mention.
- **Species:** Denotes the *species* name of the organism.
- **Subspecies:** In case the subspecies part of the organism is mentioned, it is detected and added to the feature list.
- **Strain:** Indicates the detected strain part of the organism.
- **TaxalevelID:** For the organism mentions with a more precise *subspecies* or *strain* level identification, the NCBI ID for the taxa level is also provided by the *TaxalevelID* feature during the grounding step.
- **TaxalevelscientificName:** The scientific name of the taxa level, as provided by the NCBI Taxonomy database.
- **abbrGenus:** In case the abbreviated form of the *genus* is used, this feature is set to true.
- **class:** The ontology class of the detected entity, here *Organism*.
- **docName:** The matched text in the document.
- **instanceName:** The standardized name that can be used for ontology population.
- **ncbiId:** The corresponding NCBI ID for the organism mention.
- **organismName:** The standard name as given by the NCBI Taxonomy database.
- **scientificName:** This feature provides the scientific name of the organism.
- **url:** A hyperlink connecting uniquely grounded organisms to their entry in the NCBI Taxonomy database.

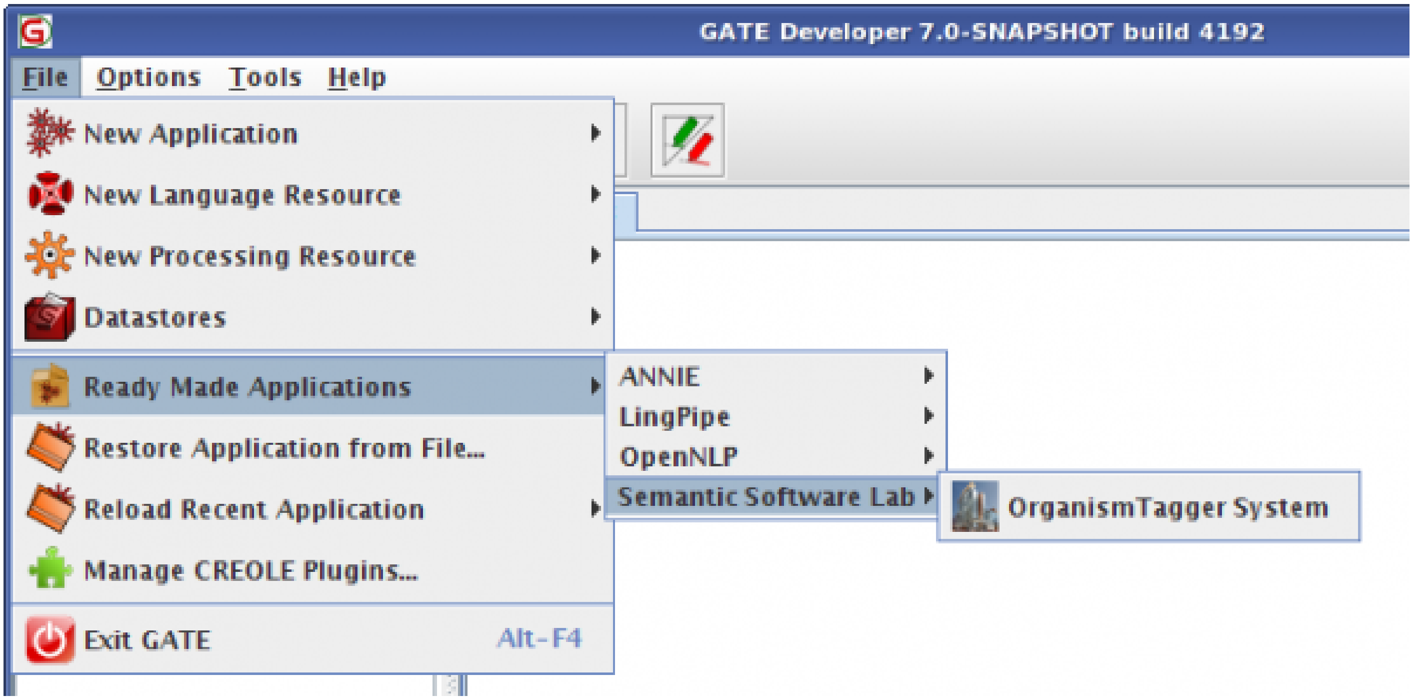
4. Corpora

Tags from two manually annotated corpora, (i) a corpus of documents on protein engineering and fungi (ii) a corpus of open access biomedical documents from the PMC from the Linnaeus project are also provided. The organism mention tags are in tab-delimited text files:

NCBI ID	Document	Start	End	Text
51453	15260499	754	772	Trichoderma reesei
51453	15260499	9295	9304	T. reesei
51453	15260499	9867	9891	T. reesei Rut-C30 strain
562	15260499	9999	10015	Escherichia coli

5. Quick Start Guide

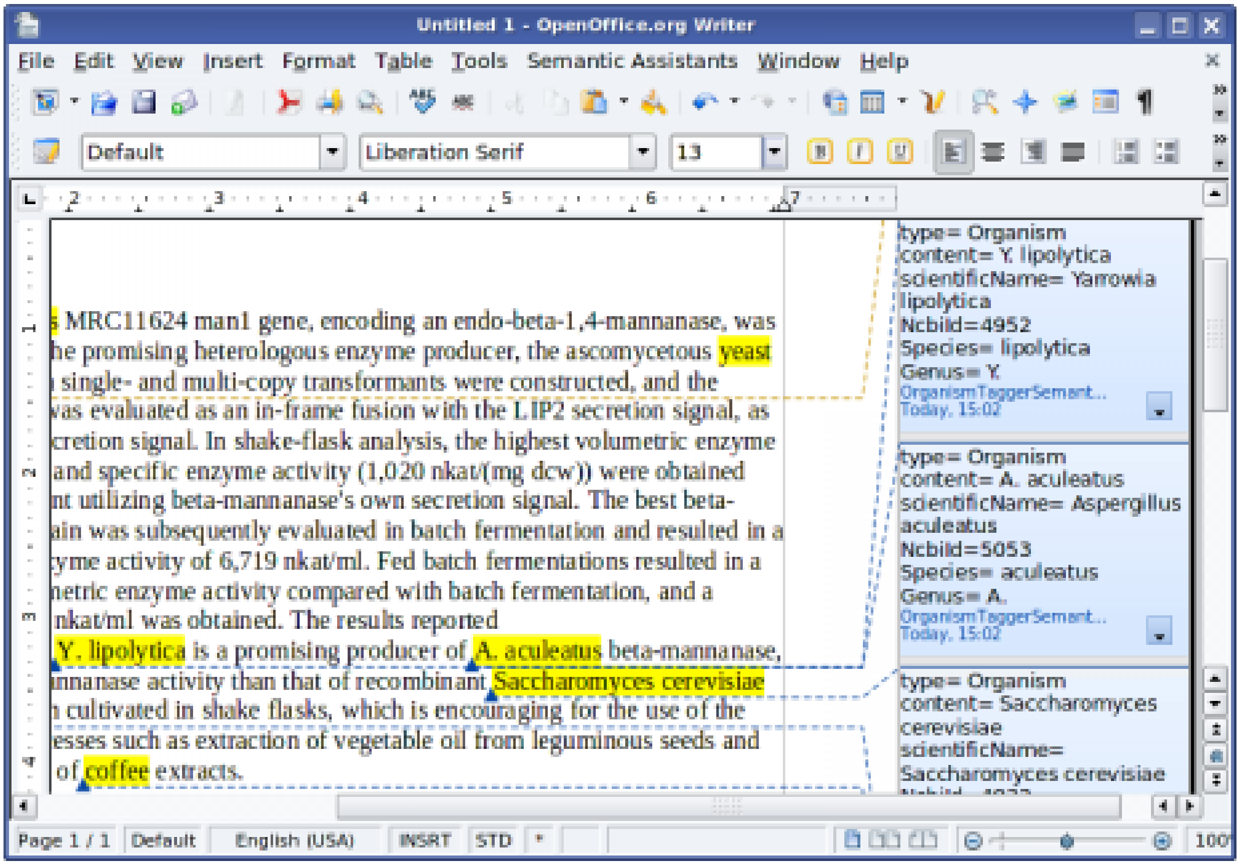
As our OrganismTagger comes in form of a GATE pipeline, you will need GATE itself. Note that you will need GATE version 7.0 or better to run the OrganismTagger. You can install the system directly from within GATE using the CREOLE Plugin Manager by selecting our *Semantic Software Lab* repository. Please note that you will need a larger than default memory setting to run the OrganismTagger, we recommend 1.5GB RAM per thread (2GB RAM per thread when using 64bit Java). Under Linux, you can start GATE in the following way: `bin/gate.sh -Xmx1500M`.



OrganismTagger System in GATE Developer

6. Running the OrganismTagger as a Web Service

The OrganismTagger can also be used as a Web service, either by itself or when integrated into a more complex pipeline. Towards this end, an OWL service description for the Semantic Assistants framework is included in the release distribution. Simply follow these two steps, and after re-loading the SA server, it will now offer an "OrganismTagger" service:


Untitled 1 - OpenOffice.org Writer
File Edit View Insert Format Table Tools Semantic Assistants Window Help
Default
Liberation Serif
13
Page 1 / 1
Default
English (USA)
INSRT
STD
100

s MRC11624 man1 gene, encoding an endo-beta-1,4-mannanase, was
he promising heterologous enzyme producer, the ascomycetous yeast
single- and multi-copy transformants were constructed, and the
vas evaluated as an in-frame fusion with the LIP2 secretion signal, as
cretion signal. In shake-flask analysis, the highest volumetric enzyme
and specific enzyme activity (1,020 nkat/(mg dcw)) were obtained
nt utilizing beta-mannanase's own secretion signal. The best beta-
ain was subsequently evaluated in batch fermentation and resulted in a
zyme activity of 6,719 nkat/ml. Fed batch fermentations resulted in a
netric enzyme activity compared with batch fermentation, and a
nkat/ml was obtained. The results reported
Y. lipolytica is a promising producer of A. aculeatus beta-mannanase,
nnanase activity than that of recombinant Saccharomyces cerevisiae
n cultivated in shake flasks, which is encouraging for the use of the
esses such as extraction of vegetable oil from leguminous seeds and
of coffee extracts.

type= Organism
content= Y. lipolytica
scientificName= Yarrowia lipolytica
NcbiId=4952
Species= lipolytica
Genus= Y.
OrganismTaggerSemant...
Today, 15:02

type= Organism
content= A. aculeatus
scientificName= Aspergillus aculeatus
NcbiId=5053
Species= aculeatus
Genus= A.
OrganismTaggerSemant...
Today, 15:02

type= Organism
content= Saccharomyces cerevisiae
scientificName=
Saccharomyces cerevisiae

Text abstract shown from *Heterologous expression and optimized production of an Aspergillus aculeatus endo-1,4-beta-mannanase in Yarrowia lipolytica*, PMID: 19507068, processed by the OrganismTagger running in OpenOffice

1. Copy the organismtagger.owl service description into the directory `Resources/OwlServiceDescriptions/` inside your semantic-assist installation
2. Copy the directory "gate" to Resources/GatePipelines and rename it to "OrganismTagger"

Either by using the command line client that ships with the Semantic Assistants architecture or any other Semantic Assistants-enabled client, like OpenOffice, you can execute the service and process input documents. Of course, the server also accepts plain SOAP requests. Our server minion.cs.concordia.ca (port 8879) usually offers the "OrganismTagger" service, if you want to test it.

7. Download

Latest version is v1.5 from 24.09.2014. You can install this version directly from within GATE through the CREOLE Plugin Manager. The download includes documentation, the example pipeline, and our annotated corpora. You can download the install package manually (but the recommended way of installation is to use the GATE plugin manager through the GATE Developer GUI).

You can also look at the documentation (this is the same file as included in the distribution in the `doc/` folder).