

# OpenMinted community-driven applications

Sophia Ananiadou  
National Centre for Text Mining  
University of Manchester



[twitter.com/openminted\\_eu](https://twitter.com/openminted_eu)

# Engaging with the communities

- Scholarly communications
  - Research performance, research publications recommendation system
  - Rock art mining; TM Leica microscopes
- Life Sciences
  - Metabolites, Curation of neuroscience, modeling chronic liver diseases
- Social Sciences
- Agriculture, Biodiversity

# Methodology: application design

- General description
- Resources
  - Document formats
  - Knowledge bases
  - Tools, components, services
- Deployment plan
- Data interfaces
- User interfaces
- Data processing scenarios
- Limitations
- Release Plan

# Scholarly Communications

## Funding Mining Services


## Rock art research

Frontiers

Ease the speed

XMI, JSON, PDF

openMIN7ED

 OpenAIRE mining service Beta

Home page **Project mining** Data citation mining Document Classification Software mining Interactive project mining Citation matching Document similarity

**Project mining details**

Provide your UTF-8 encoded text, on the current URL using the HTTP POST method. You may also choose among the available mining processes.

HTTP POST parameters:

- document: UTF-8 encoded text
- projects: Project processing (on/off)
- data citations: Data citation processing (on/off)
- classification: Classification processing (on/off)

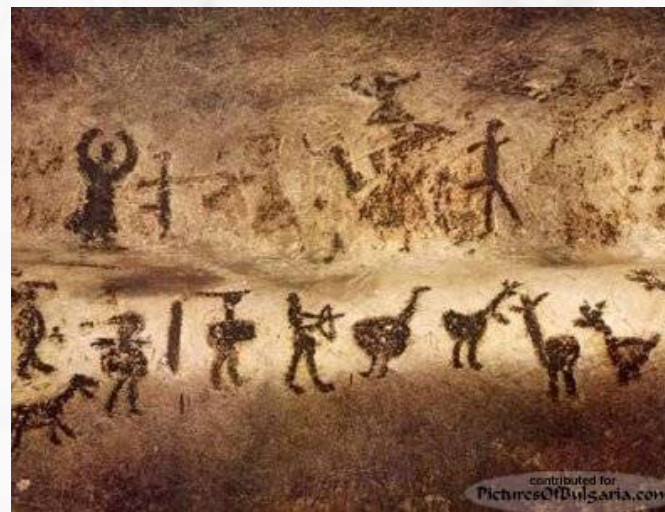
The service will return a JSON encoded result containing the following fields:

- funding\_info: Result category
- fund: FP7 or Wellcome Trust
- acronym: The project acronym (only for FP7 projects)
- grantid: The project grant identifier
- confidence: Confidence weight
- EGI-related: true/false

Paste your document here

☒ Projects ☐ Data citations ☐ Classification

To contact us, click on: <http://www.openaire.eu/en/supportthedesk> select "Submit new question" and then "subject: Technical - Mining service"



FORCE 2017



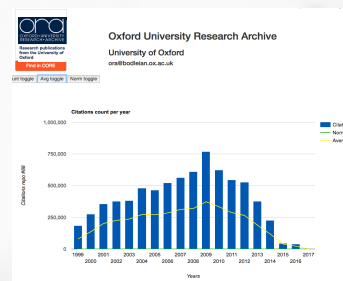


# Scholarly Communications

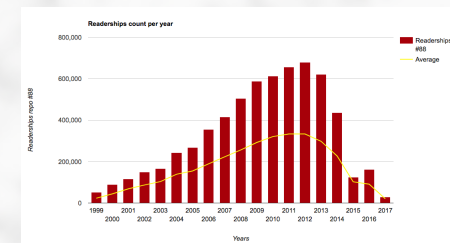
Research Publications  
Recommendation system

Research Excellence  
Trends Explorer

Citation  
counts



Readership  
counts



which can be drawn upon for informing SD project design and implementation  
**Year:** 2005  
**OAI Identifier:** oai:oro.open.ac.uk:112  
**Provided by:** Open Research Online  
**Downloaded from:** [http://oro.open.ac.uk/112/1/SDRC\\_Helsinki\\_05\\_v11.pdf](http://oro.open.ac.uk/112/1/SDRC_Helsinki_05_v11.pdf)

**Suggested articles**

**Economic analysis of World Bank education projects and project outcomes**  
**Provided by:** Research Papers in Economics  
**By:** Vawda Ayesha Yaqub, Mook Peter, Gittinger J. Price, Patrinos Harry...

**The Use of System Dynamics Simulation Models in Project Management Education**  
**Provided by:** Sunderland University Institutional Repository  
**By:** Ahmed Heba Saleh

**Muslim Pupils, Children's Fiction and Personal Understanding**  
**Provided by:** University of Worcester Research and Publications | **Publisher:** Shah Abdul Latif University, Khairpur Sindh, Pakistan.  
**By:** Gilani-Williams F., Bigger Stephen

**The uptake and implementation of sustainable construction: Transforming policy into practice**

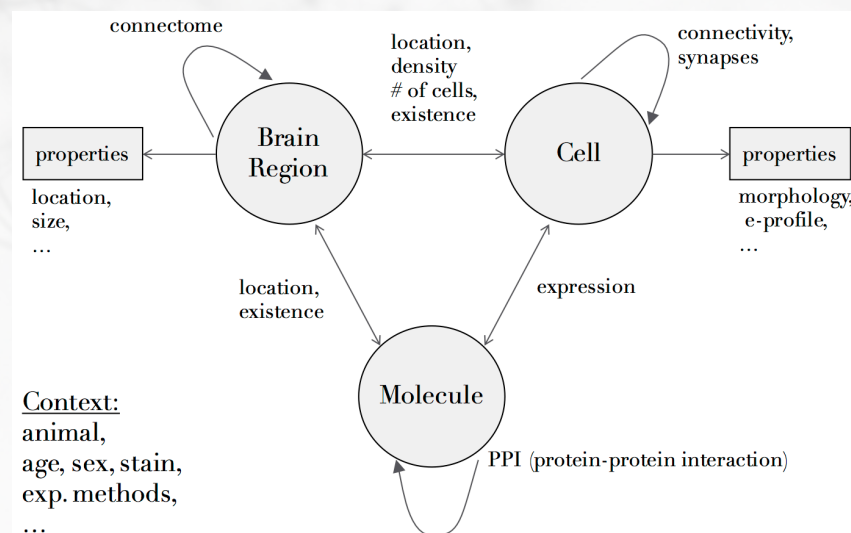
# Life Sciences

Curation  
Metabolites



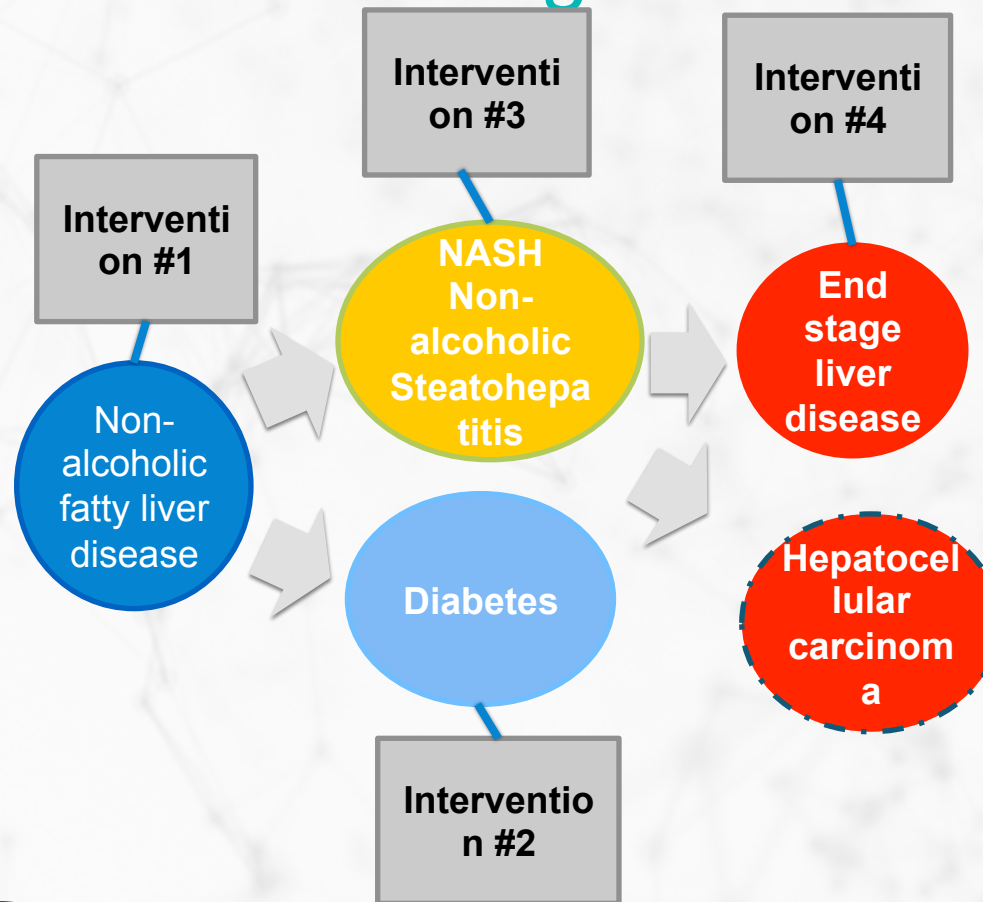
ChEBI

Neuroscience



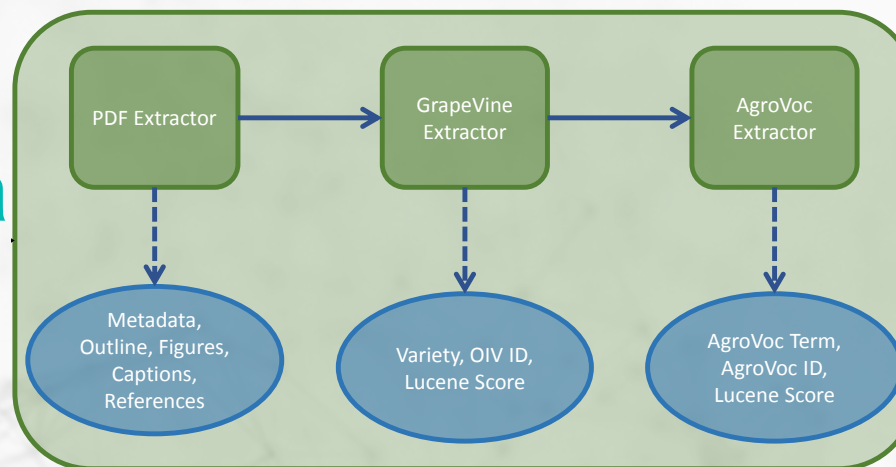
# Life Sciences

## Health State Modelling

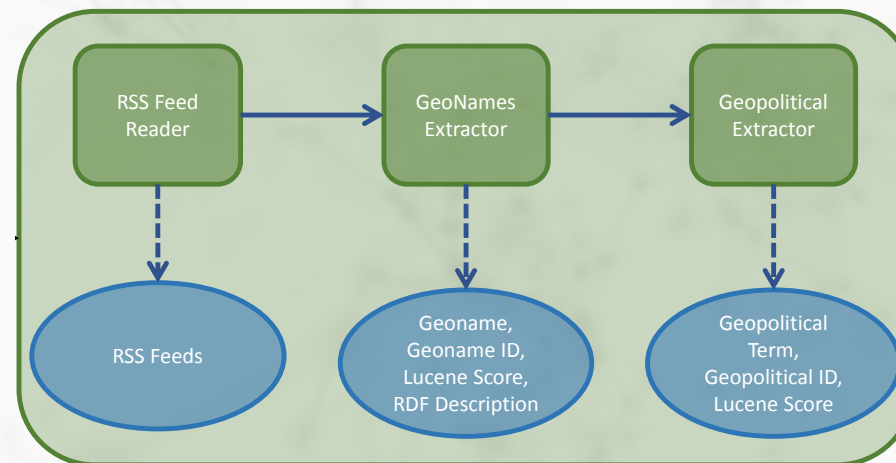


# Agriculture and biodiversity

Text mining over  
bibliographic data



Text Mining over  
RSS Feeds





# Agriculture and Biodiversity

## Microbial Biodiversity

The screenshot shows the Alvis Search Engine interface. The search bar contains the query "Psychrobacter aquimaris" "bacteria habitat". The results are displayed in a table format. On the left, there is a facet table for "Bacteria".

facet value	freq.	doc.
Psychrobacter aqu	16	11
Pseudorhodobacte	8	4
Photobacterium pi	8	4
Psychrobacter	14	3

The main search results show a list of documents. The first result is from the "International journal of systematic and evolutionary microbiology" (2005) and is titled "Psychrobacter aquimaris sp. nov. and Psychrobacter namhaensis sp. nov., isolated from sea water of the South Sea in Korea." The abstract mentions that two Gram-negative, non-motile, non-spore-forming, slightly halophilic bacterial strains, SW-210(T) and SW-242(T), were isolated from sea water of the South Sea in Korea, and were characterized taxonomically by means of a polyphasic approach. The two isolates grew optimally at 10-20°C, pH 6-8, and 0-10% NaCl. The strains were found to be closely related to Psychrobacter aquimaris.

On the right side, there is a sidebar with taxonomic information for "Psychrobacter aquimaris (taxon) (10)". It lists synonyms (10) and sub-concepts (1). Below that, it shows "bacteria habitat (habitat) (889751)" with synonyms (1).

Where does *Psychrobacter aquimaris* usually live?

## Linking Wheat Data with Literature

The screenshot shows the Alvis Search Engine interface. The search bar contains the query "wheat" ("resistance to rust" and Ir34). The results are displayed in a table format. On the left, there is a facet table for "Phenotype".

facet value	freq.
resistance to Leaf Rust	19
resistance to Stripe Rust	12
resistance to noxious weed	8
lodging resistance	6
resistance to Stem Rust	6
resistance to rust	3
response to environmental c	1

The main search results show a list of documents. The first result is from the "High-resolution mapping and new marker development for adult plant stripe rust resistance QTL in the wheat cultivar Karioga" (2014). The abstract mentions that three major quantitative trait loci (QTL) contribute to the durable adult plant stripe rust resistance in the high-quality bread wheat cultivar Karioga; QYr.sgi-2B.1 and QYr.sgi-4A.1, and the pleiotropic resistance gene Lr34/Yr18/Sr57. While marker-assisted selection is currently being used to incorporate the Karioga stripe rust adult plant resistance into new South African wheat breeding lines, effective selection of the large QTL intervals remains a challenging task. In this study, we describe the development of expressed

Is *Ir34* gene related to wheat resistance to rust disease?



# Agriculture and Biodiversity

## Extracting gene regulation networks involved in seed development (SeeDev)



<http://bibliome.jouy.inra.fr/demo/seedev/alvisir/webapi/>

search? **Alvis** Search Engine

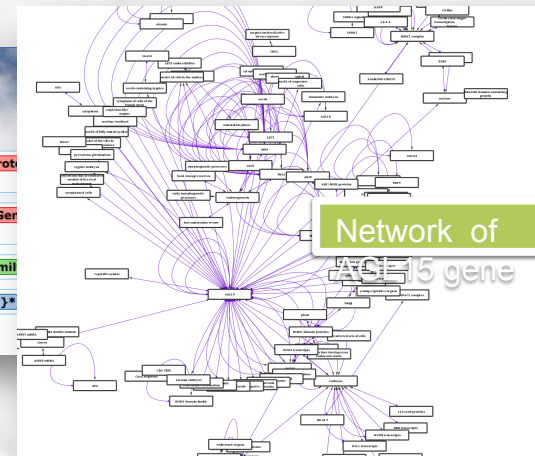
"TFL 1" and (gene\_family)\* {protein}\* Search

10 Page 1 of 1 [1 to 3 of 3]

**The Evolution of the FT/TFL1 Genes in Amaranthaceae and Their Expression Patterns in the Course of Vegetative Growth and Flowering in *Chenopodium rubrum***

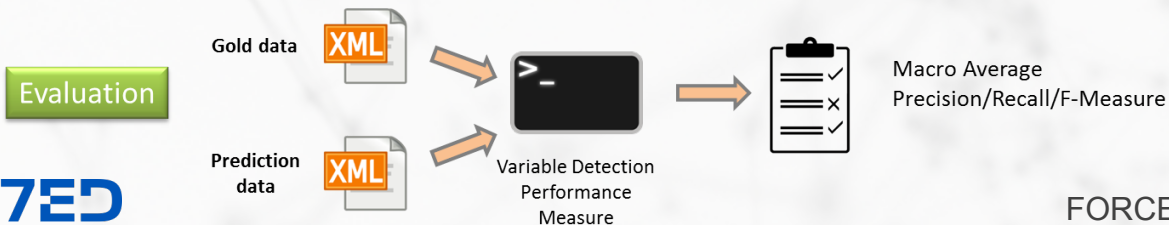
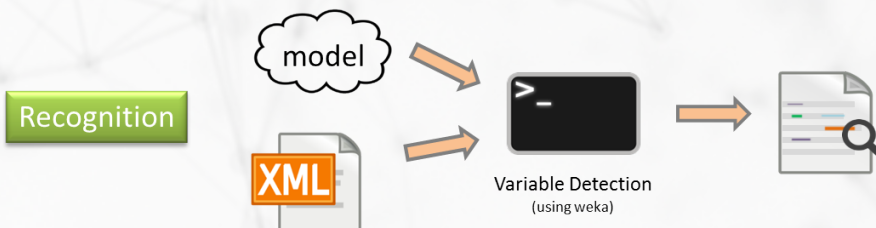
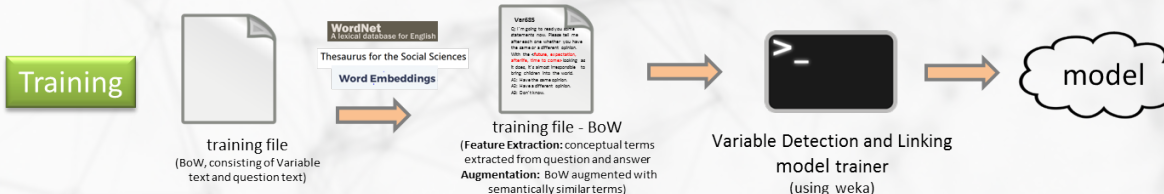
2016

**Abstract** The FT/**TFL1** gene family controls important aspects of plant development: **MFT-like** genes affect germination, **TFL1-like** genes act as floral inhibitors, and FT-like genes are floral activators. Gene duplications produced paralogs with modified functions required by the specific



# Social Sciences

## Extracting Named Entities from survey Data



---

# Focus: Text Mining for ChEBI

---

# Text Mining for ChEBI

- Identifying metabolites for curation in ChEBI
- Linking metabolites to species, chemical information

# Text Mining for ChEBI

EMBL-EBI




## ChEBI

[Home](#) | [Advanced Search](#) | [Browse](#) | [Documentation](#) | [Download](#) | [Tools](#) | [About ChEBI](#)

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds.



Search for  only ☐ All in ChEBI ☒

Example: [iron\\*](#), [InChI=1S/H2O/h1H2](#), [water](#)

[Advanced Search](#) | [About ChEBI](#)



# Text Mining for ChEBI




ChEBI Name **water**

ChEBI ID **CHEBI:15377**

Definition An oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.

Stars ★★ This entity has been manually annotated by the ChEBI Team.

Secondary ChEBI IDs CHEBI:5585, CHEBI:42857, CHEBI:42043, CHEBI:44292, CHEBI:44819, CHEBI:43228, CHEBI:44701, CHEBI:10743, CHEBI:13352, CHEBI:27313

Supplier Information 

 [Download Molfile](#)

- [Find compounds which contain this structure](#)
- [Find compounds which resemble this structure](#)
- [Take structure to the Advanced Search](#)

[more structures >>](#)

# Text Mining for ChEBI

- Majority of entries are manually curated
- Time consuming
- Annotator fatigue
- Lack of completeness

# Text Mining for ChEBI

Screenshot of the brat text mining interface showing a document with chemical entities and relationships.

Browser: nactem10.mib.man.ac.uk/brat-v1.3/#/OpenMinted-Chebi-2/Abstracts

Buttons: Collection, Data, Search, Options, Logout matt, Help

Document Text:

1 Aurasperone F- a new member of the naphtho-gamma-pyrone class isolated from a cultured microfungus, *Aspergillus niger* C-433.

3 A novel dimeric naphtho-gamma-pyrone, named aurasperone F (1), was isolated from the fermentation broth of the culture extracts of *Aspergillus niger* C-433, isolated from grapes, along with the known compounds fonsecin (2), aurasperone B (3), aurasperone C (4), aurasperone D (5) and aurasperone E (6).

Entities and Relationships:

- Entities: *Aspergillus niger* (Species), aurasperone F (1), fonsecin (2), aurasperone B (3), aurasperone C (4), aurasperone D (5), aurasperone E (6).
- Relationships: Isolated From (multiple instances connecting the novel compounds to the source organism and other compounds).

# Text Mining for CHEBI

## Corpus Stats:

- 200 abstracts
- 100 full papers

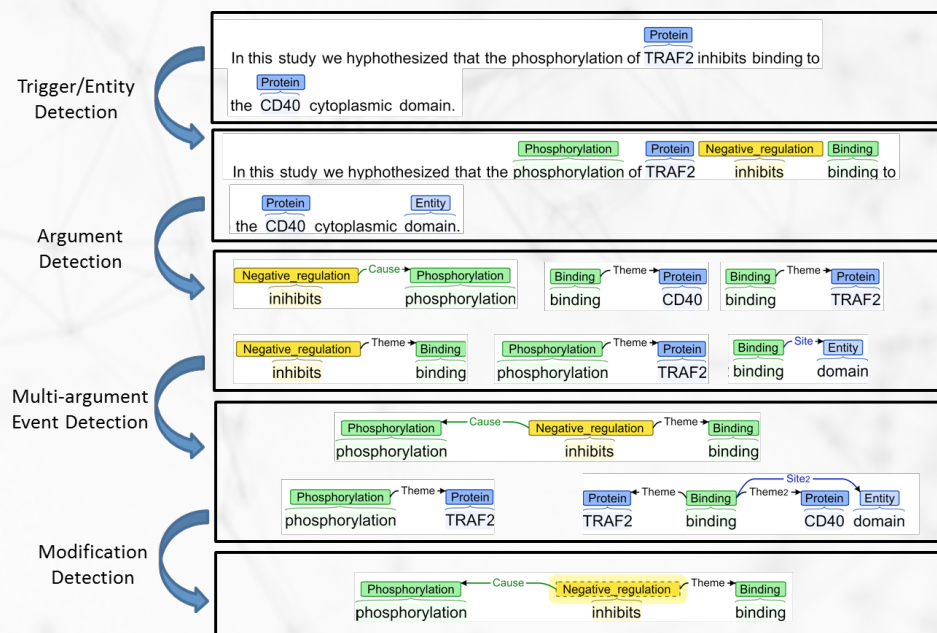
## Agreement:

- 0.934 (Entities)
- 0.779 (Relations)

# Text Mining for ChEBI

Identification of Entities + Events Models  
trained using corpora

<http://www.nactem.ac.uk/EventMine/>



Miwa, M., S. Ananiadou  
(2015)  
BMC Bioinformatics, 16  
(Supl. 10)



---

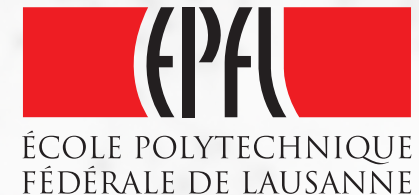
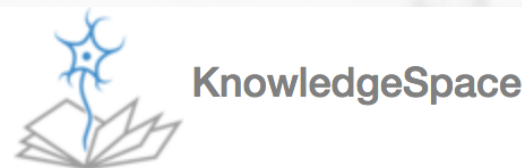
# Focus: Text mining for Neuroscience

---

# Text Mining for neuroscience

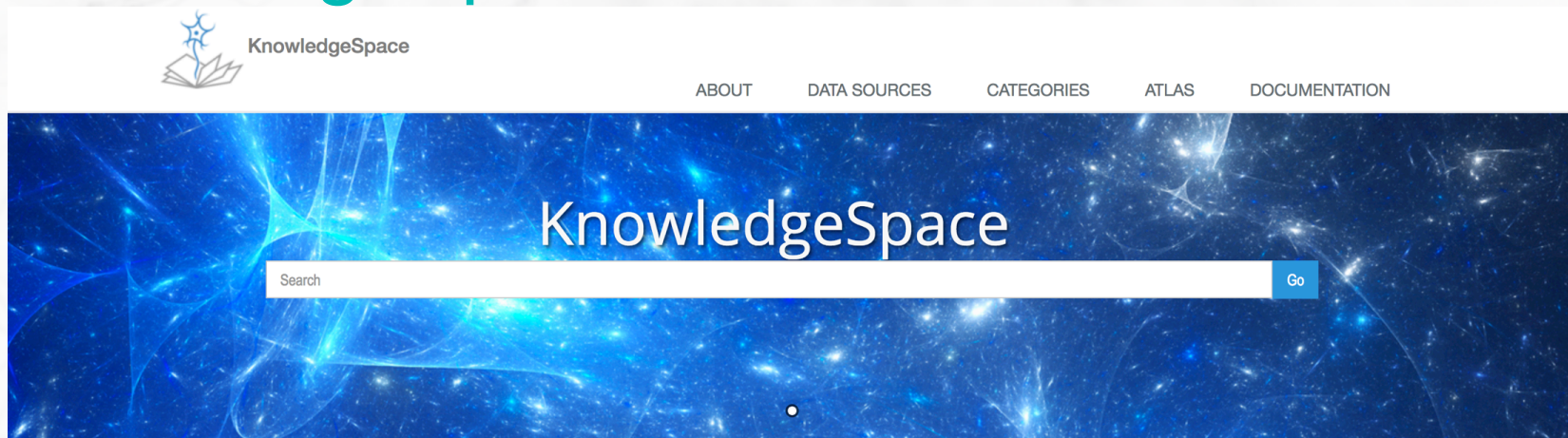
## Background

- Use these to aid curation in KnowledgeSpace
- In collaboration with Blue Brain Project at EPFL
- Curation for Neurolex

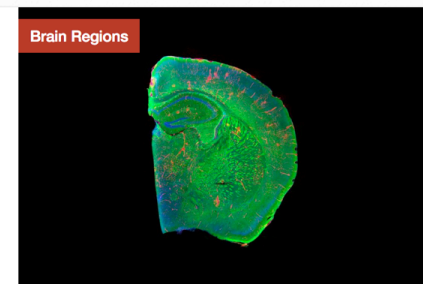
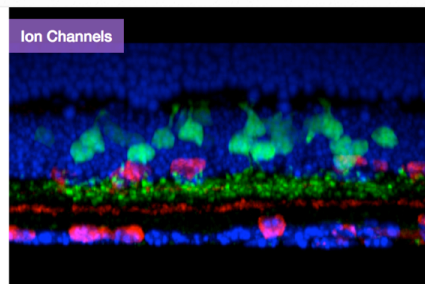


# Text Mining for neuroscience

## KnowledgeSpace



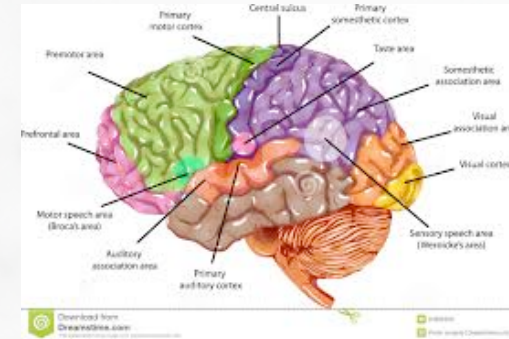
A community encyclopedia linking brain research concepts to data, models, and literature.



# Text Mining for neuroscience

Entities of Interest

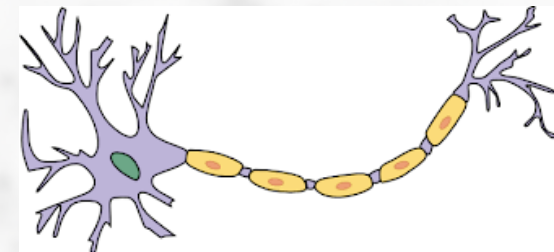
Brain Region



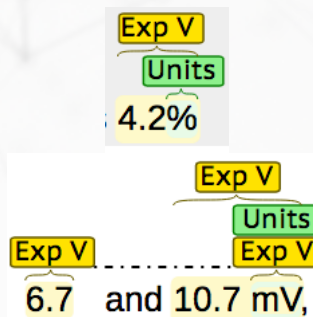
Electric Current/Channel

Model Organism

Neuron



Scientific Units/Values

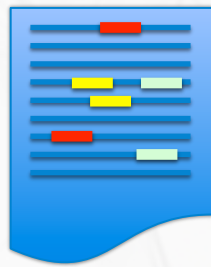




# Text Mining For Neuroscience

## Active Learning

1. Annotator labels (or corrects) examples



2. Examples are used to create new models



3. New models are used to automatically label new documents



4. Most informative sentences are selected





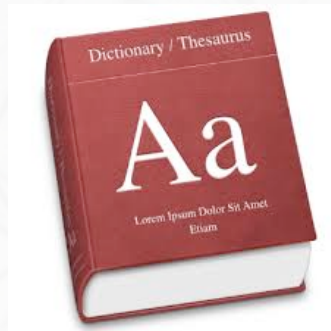
# Text Mining for Neuroscience

Entity	Agreement	Total in corpus
Brain Region	0.891	1055
Neuron	0.825	767
Model Organism	0.846	299
Ionic Channel	0.639	201
Ionic Current	0.904	339
Ionic Conductance	0.810	76
Value	0.784	594
Unit	0.902	507

# Text Mining for Neuroscience

## Methods

- Dictionary Fuzzy Matching



Entry	Match	Type
Brown Rat	Brown Rat	Exact Match
c elegans	C.Elegans	Fuzzy Match
Drosophilia	Young Drosophilia	Fuzzy Match

- Regular Expression

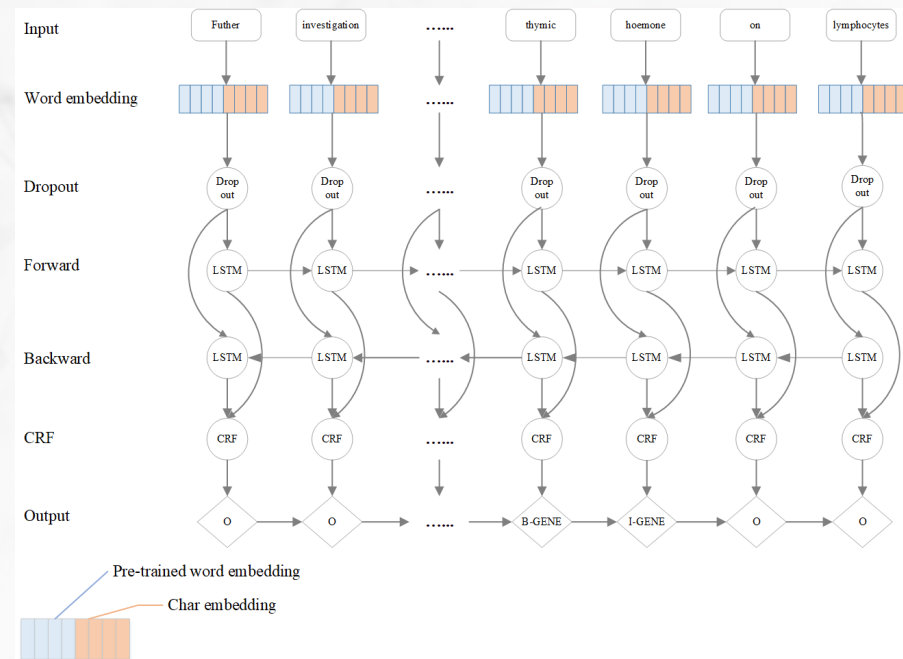
`^.*(neuron(e?s)?)|(cells?)(.*)?$`

Match any phrase with the strings 'Neuron, Neurone, Neurons, Neurones, or cells.

# Text Mining for neuroscience

## Methods

- Conditional Random Field
  - Dictionary Features
  - NER Suite – Generic Model
- Deep Learning NER
  - Neural Architecture
  - Data Driven



# Text Mining for Neuroscience

Entity	Rules / Dictionaries	CRF	Deep Learning
Brain Region	0.314	0.822	<b>0.844</b>
Neuron	0.269	0.757	<b>0.814</b>
Model Organism	0.435	0.844	<b>0.869</b>
Ionic Channel	0.278	0.600	<b>0.800</b>
Ionic Current	0.118	0.690	<b>0.764</b>
Ionic Conductanc e	0.070	0.364	<b>0.813</b>
Value	0.289	<b>0.867</b>	0.860
Unit	0.348	0.929	<b>0.930</b>

# Text Mining for Neuroscience

## Examples

- 2 Neuron GABAergic interneurons differ from Neuron glutamatergic principal neurones in their ability to discharge high-frequency trains of action potentials without adaptation.
- 3 To examine whether Na<sup>+</sup> channel gating contributed to these differences,
- Ionic Current Na<sup>+</sup> currents were recorded in nucleated patches from Neuron interneurons
- Neuron Brain Region (dentate gyrus basket cells, Neuron BCs) and Neuron principal neurones
- Br Rgn Neuron (CA1 pyramidal cells, Neuron PCs) of Species rat Brain Region hippocampal slices.



# Text Mining for Neuroscience

## Examples

- 1 Low-threshold  $\text{Ca}^{2+}$  spikes (LTS) are an indispensable signaling mechanism for Neuron Br Rgn Brain Region Brain Region Br Rgn neurons in areas including the cortex, cerebellum, basal ganglia, and thalamus.
- 2 They have critical physiological roles and have been strongly associated with disorders including epilepsy, Parkinson's disease, and schizophrenia.
- 3 However, although dendritic Ionic Current T-type  $\text{Ca}^{2+}$  channels have been implicated in LTS generation, because the properties of low-threshold spiking neuron dendrites are unknown, the precise mechanism has remained elusive.

# Text Mining for Neuroscience

## Examples

- 1 The signaling properties of thalamocortical (TC) neurons depend on the diversity of ion conductance mechanisms that underlie their rich membrane behavior at subthreshold potentials.  
Neuron
- 2 Using patch-clamp recordings of TC neurons in brain slices from mice and a realistic conductance-based computational model, we characterized seven subthreshold ion currents of TC neurons and quantified their individual contributions to the total steady-state conductance at levels below tonic firing threshold.  
Neuron Species  
Ionic Current Neuron

openMIN7ED

# Thank You

WP9 – Use case Scenarios and applications

Sophia Ananiadou



sophia.ananiadou@manchester.ac.uk



[twitter.com/openminded\\_eu](https://twitter.com/openminded_eu)



[www.openminded.eu](http://www.openminded.eu)