



**HAL**  
open science

## Implementing a Text Mining Service Offer on the Migale Bioinformatics Platform

Mouhamadou Ba, Véronique Martin, Olivier Rué, Sophie Schbath, Valérie Vidal, Valentin Loux

### ► To cite this version:

Mouhamadou Ba, Véronique Martin, Olivier Rué, Sophie Schbath, Valérie Vidal, et al.. Implementing a Text Mining Service Offer on the Migale Bioinformatics Platform. Journées Ouvertes de Biologie, Informatique et Mathématique (JOBIM), Jul 2022, Rennes, France. 2022. hal-04176740

**HAL Id: hal-04176740**

**<https://hal.inrae.fr/hal-04176740>**

Submitted on 3 Aug 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

## Implementing a Text Mining Service Offer on the Migale Bioinformatics Platform

Migale, as a collective scientific infrastructure of INRAE, is building a text mining service offer to enable the bioinformatics community to more easily exploit and extract the information contained in the scientific literature.

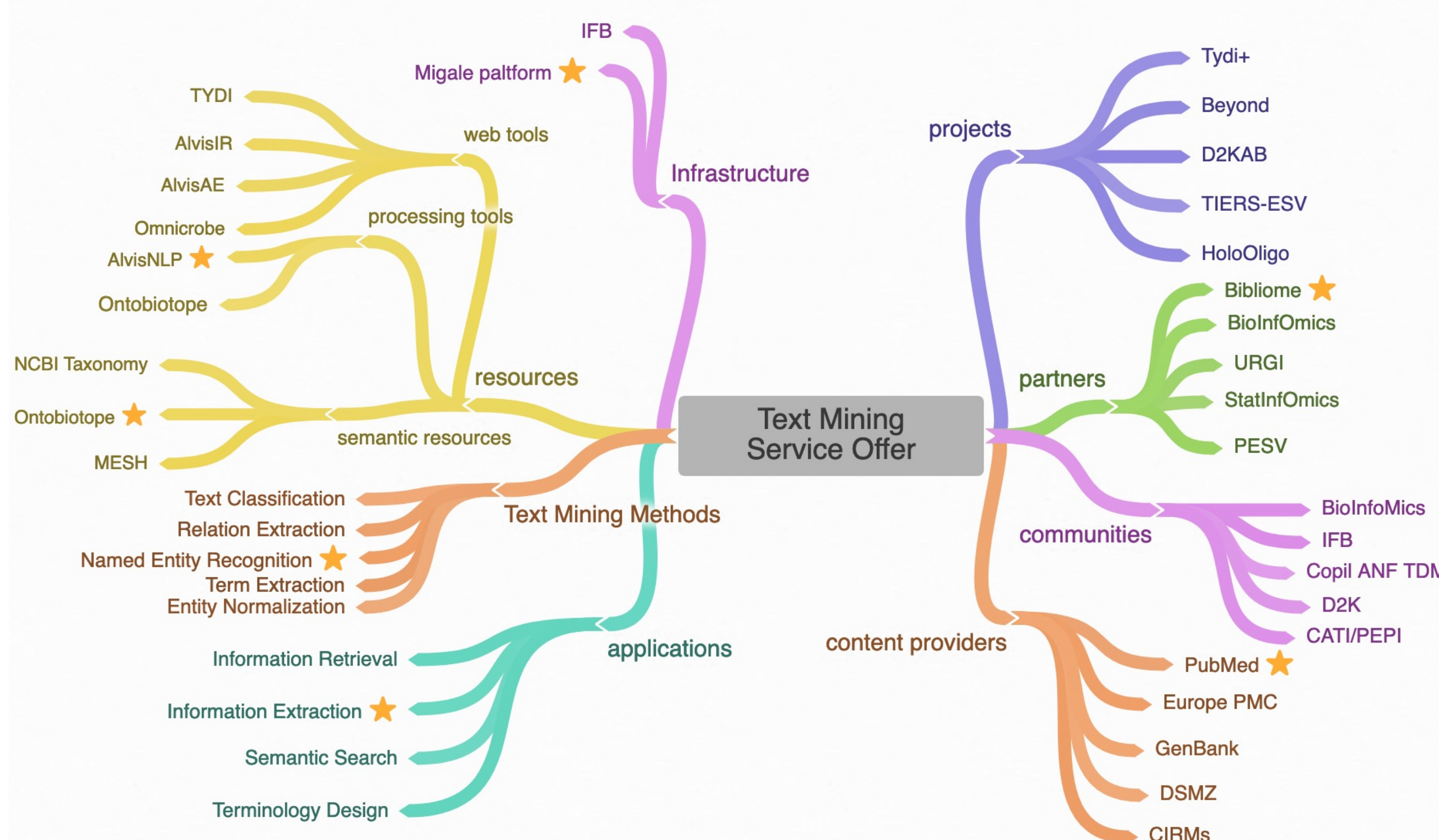


### As part of the platform missions

- Provision on the platform of **data** and **software** for text mining
- Textual data **analysis** (text mining)
- Design and **development** of applications
- Training of users in text mining **methods** and **tools**

➔ <https://migale.inrae.fr/>

### Reinforce the integration and use of text mining in bioinformatics



### Data Analysis

#### Expertise Area:

- Text and data mining

#### Applications:

- Creation of thematic corpora
- Extraction of named entities
- Classification of texts

#### Application Domain:

- Microbiologie

#### Example of projet « TDM4AnimalPhysiology »

- Extraction of genes, pesticides, polyphenols using thematic corpora collected from PubMed
- Study of male and female fertility and energy metabolism of different organisms (mammals, fish, c elegans, drosophila, plants, etc.)

### Design and Development

- **Text Mining APIs:** development of APIs wrapping text mining processes based on specialized uses
- **Tydi+:** deployment and management of instances of a new application for terminology edition
- **Omnicrobe:** management and evolution of the information extraction workflow that implements the text mining process of the Omnicrobe Information System

### Data and Tools

- **Provision** of ~ 15 instances of text mining tools developed by partners
  - AlvisIR: generic semantic search engine
  - Tydi: terminology editor
  - AlvisAE: text annotation editor
- **Regular updating of PubMed local databank**
- **Packaging, deployment and management** of text mining tools



### Trainings

- **A text mining module** in the « Bioinformatics by practicing » cycle of Migale
- Entitled: « **Introduction au text mining avec AlvisNLP** »
- Target audience: (Bio-)informaticians
- Contents:
  - Techniques for Named **Entity Recognition** (NER)
  - Use cases in biology (recognition of genes, proteins, habitats of bacteria, etc.)
  - Practice with **AlvisNLP**, a corpus processing engine developed by the Bibliome team at INRAE

Centre  
Île-de-France – Jouy-en-Josas-Antony

➔ Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France  
Université Paris-Saclay, INRAE, Bioinformatics, Migale, 78350, Jouy-en-Josas, France