FLORILEGE: an integrative database using text mining and ontologies
Estelle Chaix, Sandra Derozier, Louise Deleger, Hélène Falentin, Jean-Baptiste Bohuon, Mouhamadou Ba, Robert Bossy, Delphine Sicard, Valentin Loux, Claire Nédellec

To cite this version:
Estelle Chaix, Sandra Derozier, Louise Deleger, Hélène Falentin, Jean-Baptiste Bohuon, et al.. FLORILEGE: an integrative database using text mining and ontologies. JOBIM 2018, Jul 2018, Marseille, France. 2018. hal-01827946

HAL Id: hal-01827946
https://hal.archives-ouvertes.fr/hal-01827946
Submitted on 27 Sep 2019

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.
**FLORILEGE: an integrative database using text mining and ontologies**

Estelle Chaix¹, Sandra Derozier², Louise Deléger², Hélène Falentin¹, Jean-Baptiste Bohuon¹, Mouhamadou Ba³, Robert Bossy², Delphine Sicard², Valentin Loux¹, Claire Nédéléc²

¹ MalAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France ² STLO, INRA, Agrocampus Ouest, 35042, Rennes, France ³ SPO, INRA, Université Montpellier, 34000, Montpellier, France

---

**Biological question:** What microorganisms live in my food?

In recent years, developments in molecular technologies have led to an exponential growth of experimental data and publications spread over multiple sources. Therefore, researchers need applications that provide an unified access to both data and related scientific articles. The design of dedicated applications and services requires infrastructures and tools such that application developers and data managers can easily access to and process textual data, link them with other data and make the results available to scientists.

Florilege application dedicated to Food Microbiology is an example of application built on the top of the OpenMinTeD infrastructure.

### Knowledge resources

**Taxonomies and ontologies**

- Formal structured representations such as ontologies provide a shared reference representation (Kelso et al., 2010) for heterogeneous information from various sources. Ontologies also overcome the limitations of keyword-based search engines: semantic search engines extend simple string-matching with query facility on general terms that provide answers independently of how they are expressed in the searched text (Chaix et al., 2018).
- The Ontobiotope ontology is a formal machine-readable representation that enables indexing of information as well as normalization and reasoning.

**Ontobiotope ontology** is available on Agroportal: [http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE](http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE)

---

**How to process the text from:**

1. **Named Entity Recognition**
   - Detection of relevant biological entities to answer the biological question

2. **Relation extraction**
   - Extraction of links between the extracted entities

3. **Entity categorization**
   - Categorization allows to abstract and formalize the extracted entities from the form of the raw text to a generic class

---

**FLORILEGE database**

- Aggregation of heterogeneous data
  - A database with predicted relationships by text mining
    - Taxon ↔ Habitat (820,000 relations)
    - Phenotype ↔ Habitat (86,000 relations)
    - Query filtering on data source, QPS status…
  - Data export in tabulated format
  - Critical information stored in:
    - GenBank (NCBI)
    - GOLD (JGI)
    - BacDive (DSMZ)
    - CIRM database (INRA)
  - A key step for heterogeneous data integration

---

---

---