FLORILEGE: an integrative database using text mining and ontologies

Estelle Chaix, Sandra Derozier, Louise Deleger, Hélène Falentin, Jean-Baptiste Bohuon, Mouhamadou Ba, Robert R. Bossy, Delphine Sicard, Valentin Loux, Claire Nédellec

FLORILEGE: an integrative database using text mining and ontologies

Estelle Chaix[1], Sandra Dérozier[1], Louise Deléger[1], Hélène Falentin[2], Jean-Baptiste Bohuon[1], Mouhamadou Ba[1], Robert Bossy[1], Delphine Sicard[3], Valentin Loux[1], Claire Nédellec[1]

[1] MaIAGE, 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.

What information?

★ **Entities** (terms with particular interest for microbiologists)

Microbe | Habitat | Phenotype

★ **Relationships** (links between entities)

Microbe —Lives in→ Habitat

Microbe —Exhibits→ Phenotype

Critical information stored in

★ **public databases**
- GenBank (NCBI)
- GOLD (JGI)
- BacDive (DSMZ)
- CIRM database (INRA)

★ **scientific publications**
- Abstract (*e.g.* PubMed)
- Open access full texts (*e.g.* PMC)
- Non-open access publications

Text mining

How to process the text from:

This workflow is available on the OpenMinTeD platform : Alvis Habitat-Phenotype Relation Extractor for Microbes
https://frama.link/FlorilegeWorkflow

➢ **scientific publications?**

The effect of high hydrostatic pressure on the survival of the psychrotrophic organisms *Listeria monocytogenes*, *Bacillus cereus*, and *Pseudomonas fluorescens* was investigated in ultrahigh-temperature milk.

➢ **databases?**

GenBank | BacDive | Inra CIRM

Named Entity Recognition

Detection of relevant biological entities to answer the biological question

The effect of high hydrostatic pressure on the survival of the psychrotrophic organisms *Listeria monocytogenes*, *Bacillus cereus*, and *Pseudomonas fluorescens* was investigated in ultrahigh-temperature milk.

TaxID	Taxon	Isolation source
610251	*Lactococcus fujiensis*	outer leaves of Chinese cabbages
???	*L. fujiensis*	Three fermented soybean food samples

Relation extraction

Extraction of links between the extracted entities

Listeria monocytogenes —Exhibits→ psychrotrophic
Bacillus cereus —Exhibits→ psychrotrophic
Pseudomonas fluorescens —Exhibits→ psychrotrophic
Listeria monocytogenes —Lives in→ ultrahigh-temperature milk
Bacillus cereus —Lives in→ ultrahigh-temperature milk
Pseudomonas fluorescens —Lives in→ ultrahigh-temperature milk

+

Lactococcus fujiensis —Lives in→ outer leaves of Chinese cabbages
L. fujiensis —Lives in→ fermented soybean food samples

A key step for heterogeneous data integration

Entity categorization

Categorization allows to abstract and formalize the extracted entities from the form of the raw text to a generic class

Lactococcus fujiensis —Lives in→ outer leaves of Chinese cabbages
NCBI taxon 610251 *Lactococcus fujiensis* — Ontobiotope OBT:0000094 *Chinese cabbage*
L. fujiensis —Lives in→ fermented soybean food samples
NCBI taxon 610251 *Lactococcus fujiensis* — Ontobiotope OBT:0000433 *soybean and related product* + Ontobiotope OBT:0001668 *fermented soybean*
Listeria monocytogenes —Exhibits→ psychrotrophic
NCBI taxon 1639 *Listeria monocytogenes* — Ontobiotope OBT: 0000328 *psychrotrophic*
…

Knowledge resources

NCBI taxonomy | **Ontobiotope Habitat** (Papazian *et al.*, 2012; Bossy *et al.*, 2015)

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 in a formal machine-readable representation that enables indexing of information as well as conceptualization and reasoning.

Ontobiotope ontology is available on Agroportal :
http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE

Detection and normalization of:

Microbe

Class of entities represents: Genetic proximity (phylogeny)

Example:
- Lactobacillaceae
  - Lactobacillus
    - Lactobacillus acetotolerans
      - Lactobacillus acetotolerans
    - Lactobacillus acidifarinae
      - Lactobacillus acidifarinae DSM 19394
    - Lactobacillus acidipiscis
      - Lactobacillus acidipiscis DSM 15836
      - Lactobacillus acidipiscis KCTC 13900
    - Lactobacillus acidophilus
      - Lactobacillus acidophilus 30SC
      - Lactobacillus acidophilus ATCC 4796
      - Lactobacillus acidophilus CFH
      - Lactobacillus acidophilus CIP 76.13
      - Lactobacillus acidophilus CIRM-BIA 442
      - Lactobacillus acidophilus CIRM-BIA 445
      - Lactobacillus acidophilus CRBIP 24179
      - Lactobacillus acidophilus DSM 20079
      - Lactobacillus acidophilus DSM 20242
      - Lactobacillus acidophilus DSM 9126
      - Lactobacillus acidophilus JV3179
      - Lactobacillus acidophilus La-14
      - Lactobacillus acidophilus NCFM

NCBI taxon 1579
Lactobacillus acidophilus
synonym: Thermobacterium intestinale, Bacillus acidophilus

Habitat

Class of entities represents: Habitats with similar physico-chemical characteristics

- milk and milk product
  - butter
  - buttermilk
  - cheese
    - fermented cheese
      - brined cheese
      - fermented fresh cheese
      - ripened cheese
      - streched curd cheese
      - whey cheese
    - fresh cheese
      - chhena
      - cottage cheese
        - fermented cottage cheese
      - fermented fresh cheese
      - quark
      - queso blanco
      - queso fresco
    - processed cheese

Ontobiotope OBT:0000194
fermented cheese
is_a: cheese
is_a: fermented fairy product

Phenotype

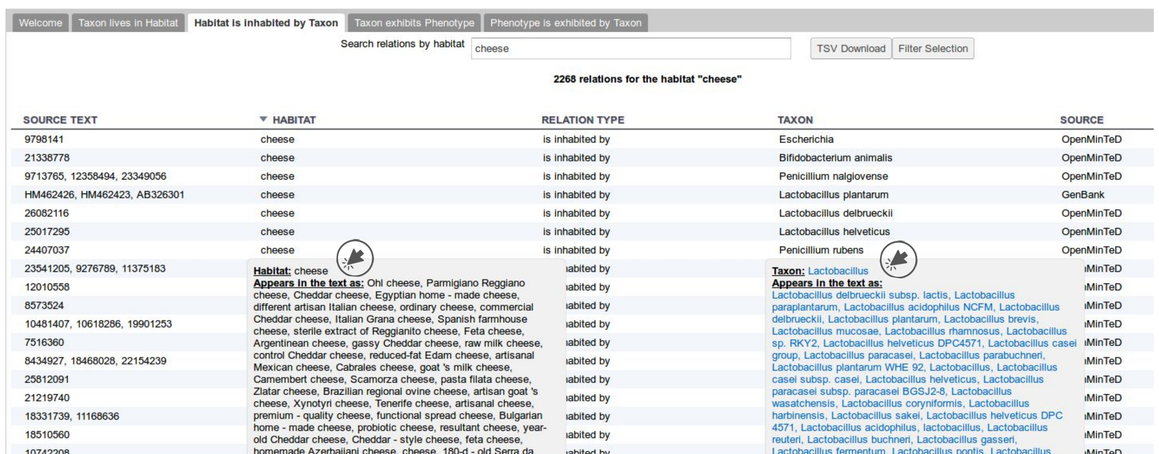
Class of entities represents: Phenotypes describing the impact of a factor on microbial behaviour

- microbial phenotype
  - phenotype wrt adhesion
  - phenotype wrt community behaviour
  - phenotype wrt growth
  - phenotype wrt metabolic activity
  - phenotype wrt microbial-host interaction
  - phenotype wrt morphology
  - phenotype wrt motility
    - motile
    - non motile
    - taxis phenotype
  - phenotype wrt ploidy
  - phenotype wrt stress
  - phenotype wrt chemical composition
    - acido resistant
    - acido sensitive
    - acido tolerant
    - alkali resistant

Ontobiotope OBT: 0000372
acido resistant
synonym: acidoresistant,acidresistant
is_a : phenotype wrt chemical composition
is_a: stress resistant

FLORILEGE database

Aggregation of heterogeneous data



- ★ Web application available at http://migale.jouy.inra.fr/Florilege/
- ★ A database with predicted relationships by text mining
  - Taxon ↔ Habitat (820,000 relations)
  - Taxon ↔ Phenotype (86,000 relations)
- ★ A direct link to external public data (DSMZ, GenBank, CIRM)
- ★ Hierarchical and synonym search
- ★ Query filtering on data source, QPS status…
- ★ Data export in tabulated format

PostgreSQL

GWT

Perspectives

- ➢ Update Ontobiotope on Agroportal
- ➢ Adding BioSample data & Updating GenBank data
- ➢ Taxonomy and ontology navigation on Florilege database
- ➢ Improvement of Phenotype detection by a text mining process

Acknowledgements

This work was supported by the OpenMinTeD project (EC/H2020-EINFRA 654021). We would like to thank the biologists of the Florilège working group of the metaprogramme MEM - Meta-omics and microbial ecosystems- of the French National Institute for Agricultural Research (Inra) and the Food Microbiome project, for their participation in the enrichment of the OntoBiotope Habitat ontology.

References:
Bossy, R., Golik, W., Ratkovic, Z., Valsamou, D., Bessieres, P., & Nédellec, C. (2015). Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task. *BMC bioinformatics*, *16*(10), S1.
Chaix, E., Déleger, L., Bossy, R. Nédellec, C. (2018). Text-mining tools for extracting information about microbial biodiversity in food. *Food Microbiology* (In Press).
Kelso, J., Hoehndorf, R., & Prüfer, K. (2010). Ontologies in biology. In *Theory and applications of ontology: Computer applications* (pp. 347-371). Springer, Dordrecht.
Papazian, F., Bossy, R., & Nédellec, C. (2012, July). AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop* (pp. 149-152). ACL
Nédellec, C., Bossy, R., Chaix, E., & Deléger, L. (2018). Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. *arXiv preprint arXiv:1805.04107*.