



HAL
open science

FLORILEGE: an integrative database using text mining and ontologies

Estelle Chaix, Sandra Derozier, Louise Deleger, H el ene Falentin,
Jean-Baptiste Bohuon, Mouhamadou Ba, Robert R. Bossy, Delphine Sicard,
Valentin Loux, Claire N edellec

► To cite this version:

Estelle Chaix, Sandra Derozier, Louise Deleger, H el ene Falentin, Jean-Baptiste Bohuon, et al.. FLO-
RILEGE: an integrative database using text mining and ontologies. JOBIM 2018, Jul 2018, Marseille,
France. , 2018, ABSTRACTS JOBIM 2018. hal-01827946

HAL Id: hal-01827946

<https://hal.science/hal-01827946v1>

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 destin ee au d ep ot et  a la diffusion de documents scientifiques de niveau recherche, publi es ou non,  emanant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv es.

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 a 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

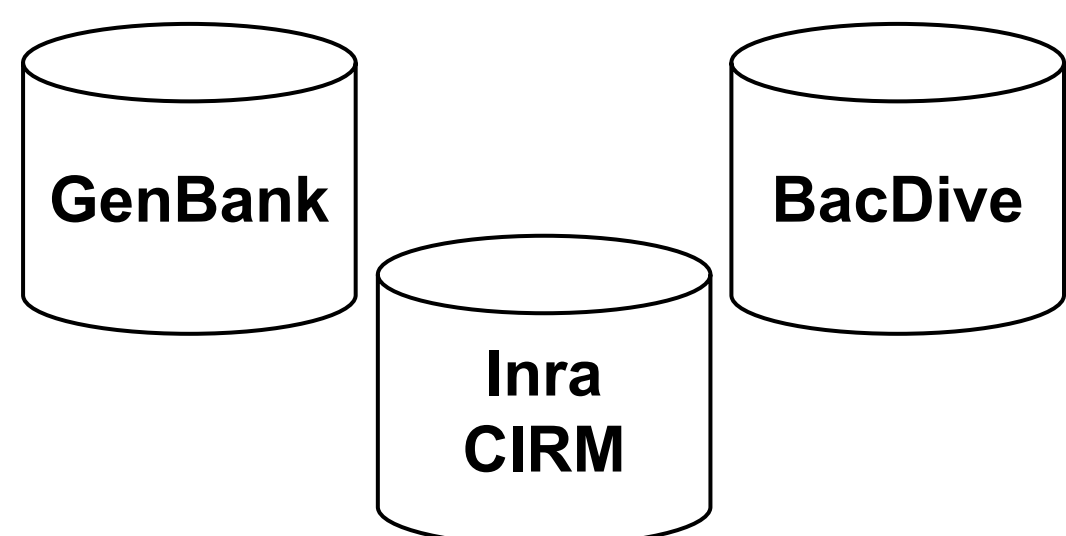
Text mining

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?



How to process the text from:

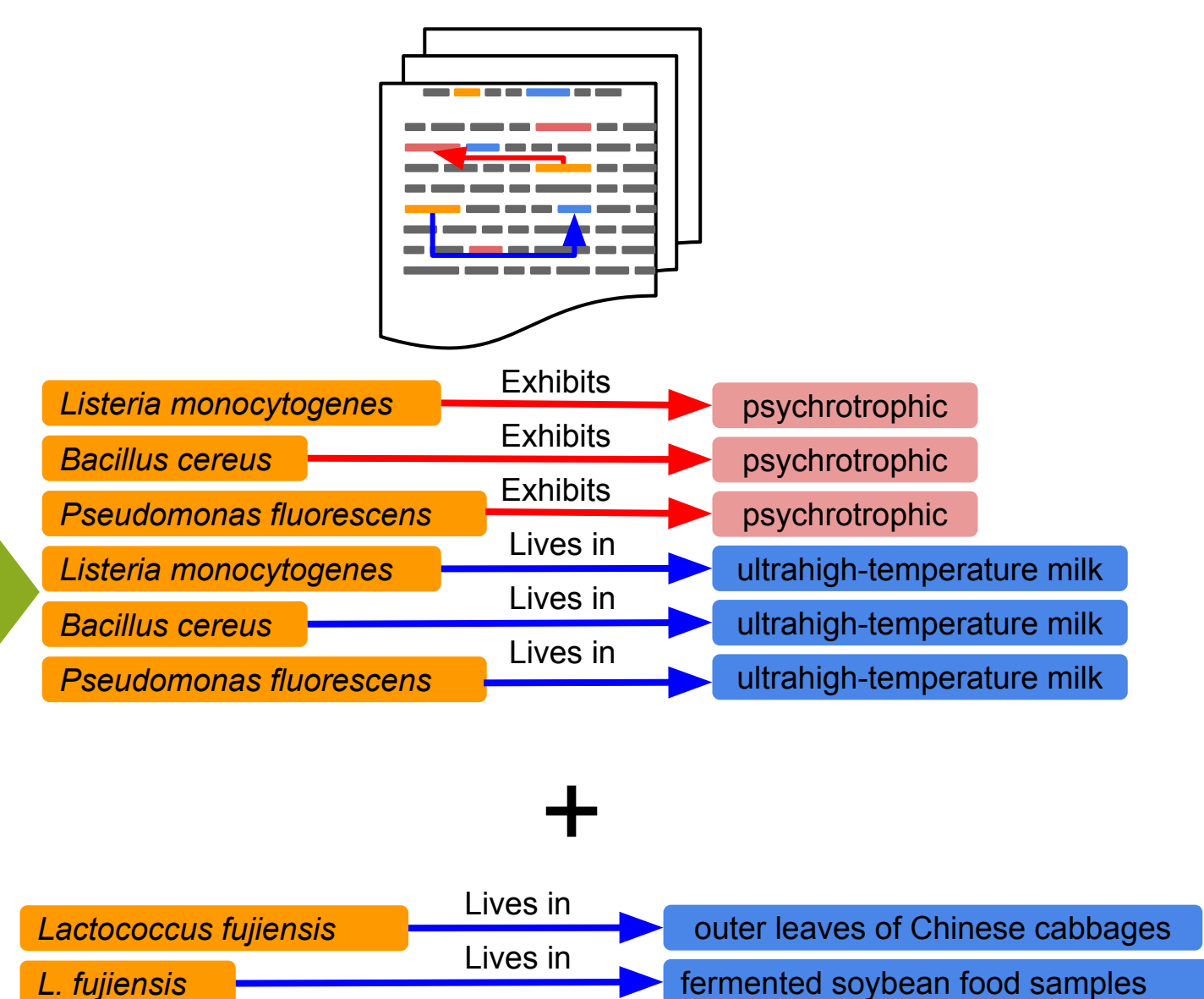


The effect of high hydrostatic pressure on the survival of the psychrotrophic organisms *Listeria monocytogenes*, *Bacillus cereus*, and *Pseudomonas fluorescens* was investigated in ultrahigh-temperature milk.

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



This workflow is available on the OpenMinTeD platform : Alvis Habitat-Phenotype Relation Extractor for Microbes



Knowledge resources

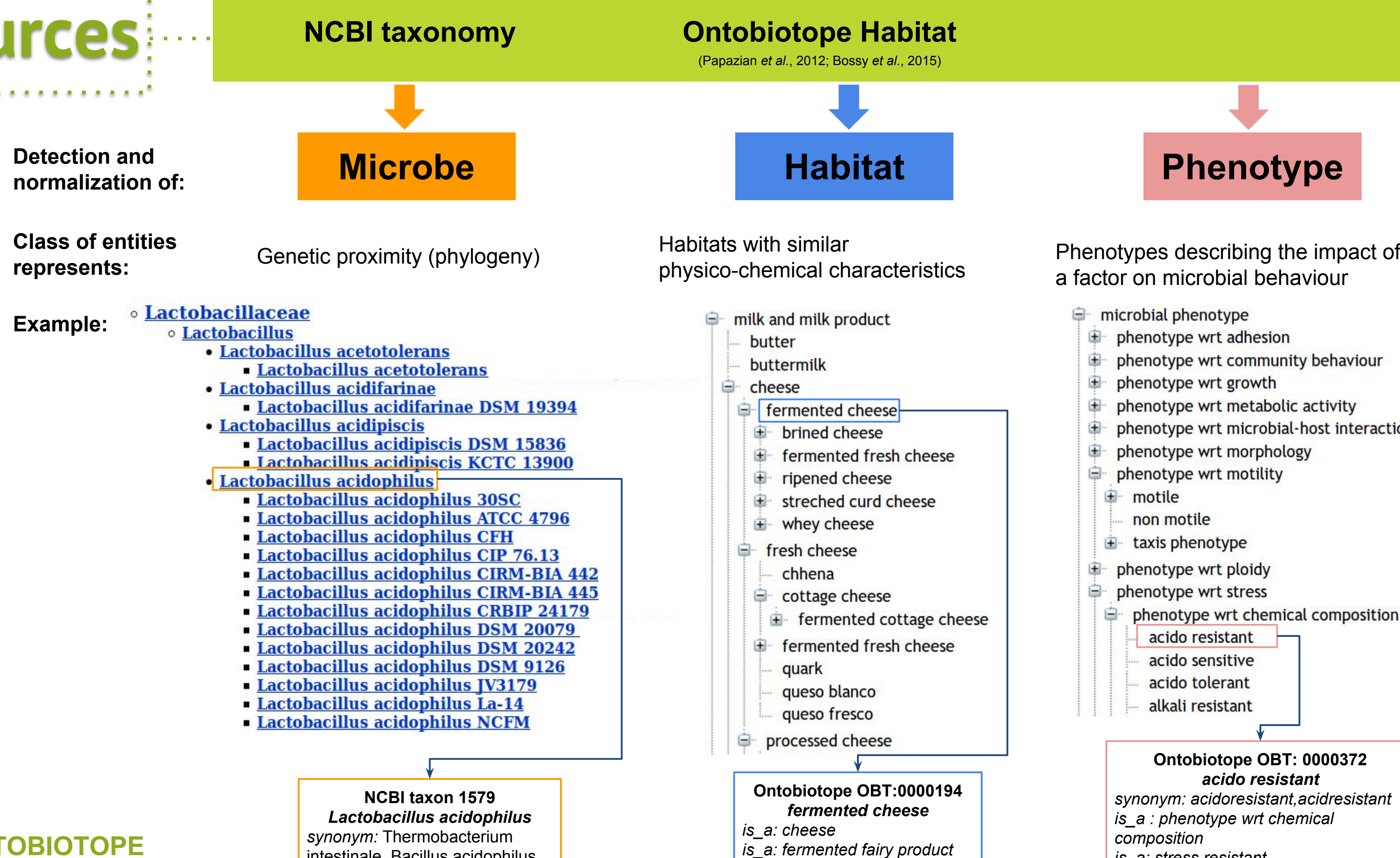
Taxonomies and ontologies

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 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>

FLORILEGE database

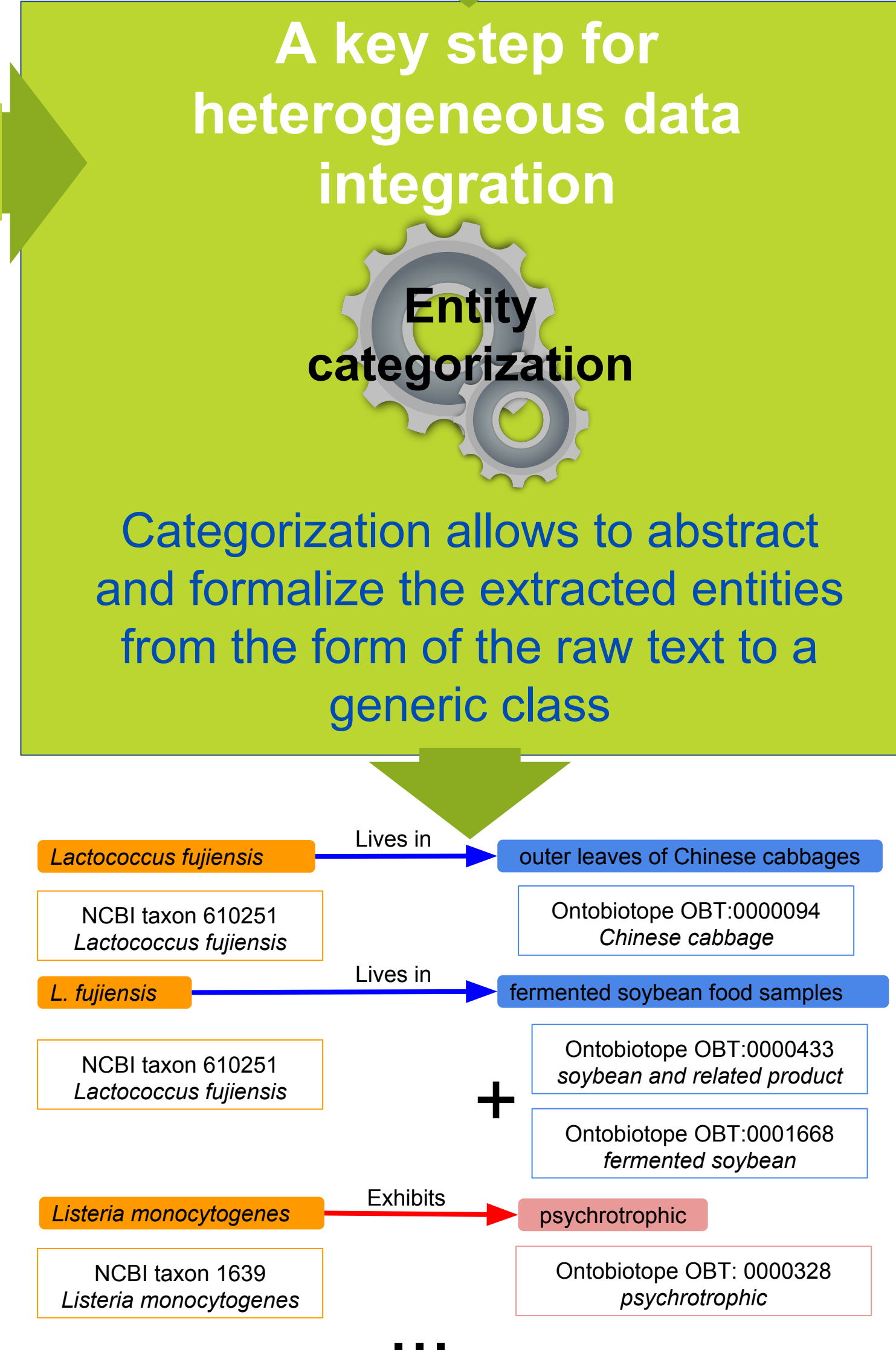


Detection and normalization of:

Class of entities represents:

Example:

- Genetic proximity (phylogeny)
 - NCBI taxon 1579 *Lactobacillus acidophilus* synonym: Thermobacterium intestinale, Bacillus acidophilus
- Habitats with similar physico-chemical characteristics
 - Ontobiotope OBT:0000194 fermented cheese is_a: cheese is_a: fermented dairy product
- Phenotypes describing the impact of a factor on microbial behaviour
 - Ontobiotope OBT: 0000372 acido resistant synonym: acidoresistant, acidresistant is_a: phenotype wrt chemical composition is_a: stress resistant



Aggregation of heterogeneous data

2268 relations for the habitat "cheese"

SOURCE TEXT	HABITAT	RELATION TYPE	TAXON	SOURCE
9798141	cheese	is inhabited by	Escherichia	OpenMinTeD
21338778	cheese	is inhabited by	Bifidobacterium animalis	OpenMinTeD
9713765, 12355494, 23349056	cheese	is inhabited by	Penicillium natigovense	OpenMinTeD
HM462426, HM462423, AB326301	cheese	is inhabited by	Lactobacillus plantarum	GenBank
2602116	cheese	is inhabited by	Lactobacillus delbrueckii	OpenMinTeD
25017295	cheese	is inhabited by	Lactobacillus helveticus	OpenMinTeD
24407037	cheese	is inhabited by	Penicillium rubens	OpenMinTeD
23541205, 9276789, 11375183	Habitat: cheese	is inhabited by	Taxon: Lactobacillus	MinTeD
12010508	Appears in the text as: OH cheese, Parmigiano Reggiano cheese, Cheddar cheese, Egyptian home-made cheese, different artisan Italian cheese, ordinary cheese, commercial Cheddar cheese, Italian Grana cheese, Spanish farmhouse cheese, starter extract of Reggiano cheese, Feta cheese, Argentinian cheese, gassy Cheddar cheese, raw milk cheese, control Cheddar cheese, reduced-fat Edam cheese, artisanal Mexican cheese, Cabrales cheese, goat's milk cheese,	is inhabited by	Lactobacillus plantarum	MinTeD
10481407, 10618286, 19901253	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
7516360	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
8434927, 15466028, 22154239	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
25612091	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
21219740	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
15331739, 11166536	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
18510590	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
10742208	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
440405, 26320771, 25998659	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD
21740724	is inhabited by	is inhabited by	Lactobacillus plantarum	MinTeD

★ 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...

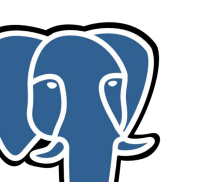
★ Data export in tabulated format

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.



PostgreSQL



GWT