INRA
SCIENCE & IMPACT

Ontology-based text mining for microbiology research

Claire Nédellec, MaIAGE

DataIA-JIST International Symposium on Data Science and AI – 10 juillet 2018

How can we make sense of textual data

Over 60 million articles, 2.5 per year

160 million scientific documents indexed

50% of the papers are not read

90% of the papers are not cited

80% of the cited papers are not read

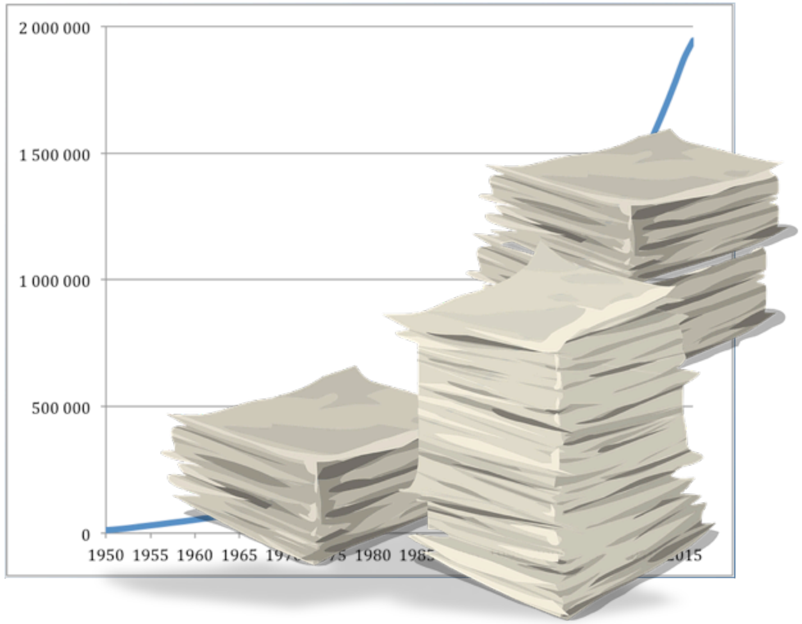
…

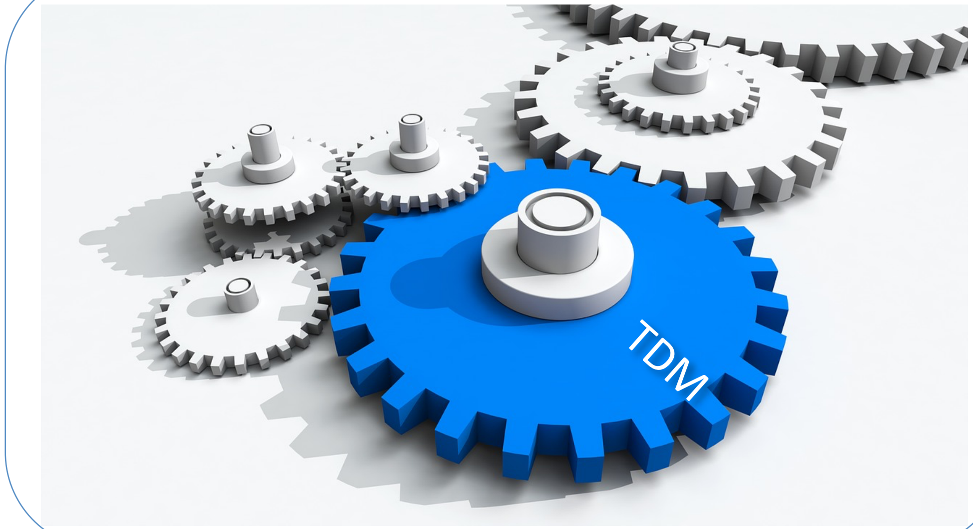
The STM report, 2015

Orduña-Malea et al. , 2014

Lokman I. Meho, the rise and rise of citation analysis, 2007.

Simkin & Roychowdhury. Read before you cite!



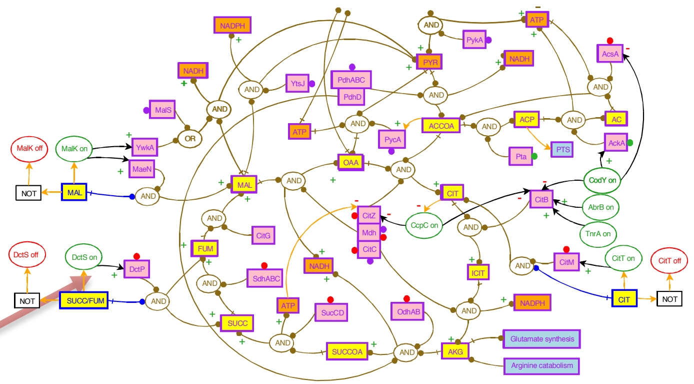


Text-mining (TDM)
Make sense of textual data
Transform unstructured data into structured, machine-readable data

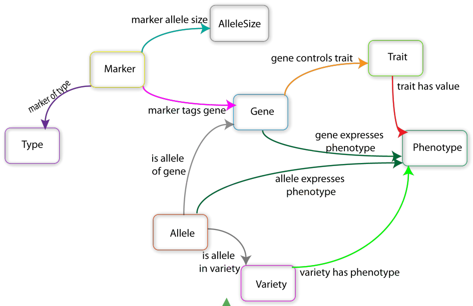
at the heart of the activity of non specialist researchers

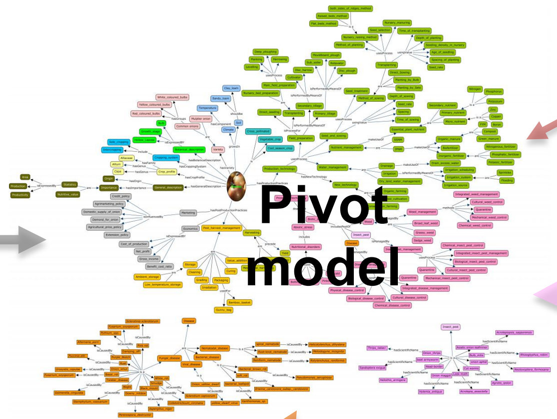
Research needs data integration, management, reasoning provided by *formal representation*

Dynamic model

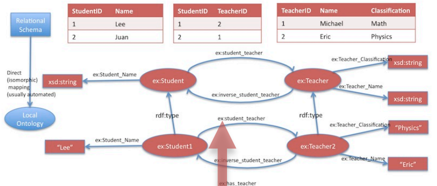


Knowledge model of the text





Data model

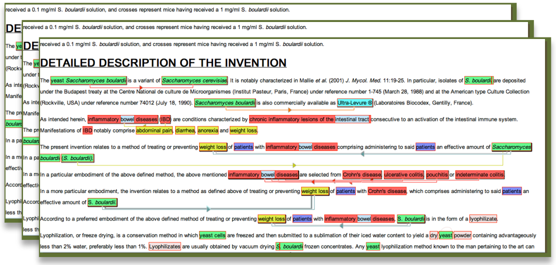


StudentID	Name
1	Lee
2	Juan

StudentID	TeacherID
1	2
2	1

TeacherID	Name	Classification
1	Michael	Math
2	Eric	Physics

Text



received a 0.1 mg/ml S. *boulardii* solution, and crosses represent mice having received a 1 mg/ml S. *boulardii* solution.

DETAILED DESCRIPTION OF THE INVENTION

The yeast *Saccharomyces boulardii* is a variant of *Saccharomyces cerevisiae*. It is notably characterized in Mallie *et al.* (2001) *J. Mycol. Med.* 11:19-25. In particular, isolates of *S. boulardii* are deposited under the Budapest treaty at the Centre National de culture de Microorganismes (Institut Pasteur, Paris, France) under reference number 1-745 (March 28, 1988) and at the American type Culture Collection (Rockville, USA) under reference number 74012 (July 18, 1990). *Saccharomyces boulardii* is also commercially available as Ultra-Levure ® (Laboratoires Biocodex, Gentilly, France).

As intended herein, inflammatory bowel diseases (IBD) are conditions characterized by chronic inflammatory lesions of the intestinal tract consecutive to an activation of the intestinal immune system. Manifestations of IBD notably comprise abdominal pain, diarrhea, anorexia and weight loss.

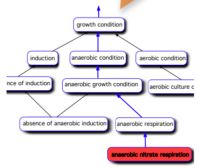
The present invention relates to a method of treating or preventing weight loss of patients with inflammatory bowel diseases comprising administering to said patients an effective amount of *Saccharomyces boulardii* (*S. boulardii*).

In a particular embodiment of the above defined method, the above mentioned inflammatory bowel diseases are selected from Crohn's disease, ulcerative colitis, pouchitis or indeterminate colitis.

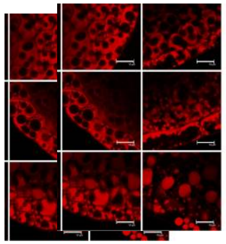
In a more particular embodiment, the invention relates to a method as defined above of treating or preventing weight loss of patients with Crohn's disease, which comprises administering to said patients an effective amount of *S. boulardii*.

According to a preferred embodiment of the above defined method of treating or preventing weight loss of patients with inflammatory bowel diseases, *S. boulardii* is in the form of a lyophilizate.

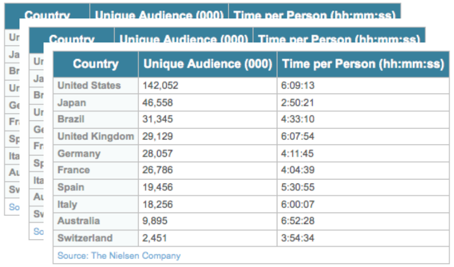
Lyophilization, or freeze drying, is a conservation method in which yeast cells are freezed and then submitted to a sublimation of their iced water content to yield a dry yeast powder containing advantageously less than 2% water, preferably less than 1%. Lyophilizates are usually obtained by vacuum drying *S. boulardii* frozen concentrates. Any yeast lyophilization method known to the man pertaining to the art can



Image



Data



Country	Unique Audience (000)	Time per Person (hh:mm:ss)
United States	142,052	6:09:13
Japan	46,558	2:50:21
Brazil	31,345	4:33:10
United Kingdom	29,129	6:07:54
Germany	28,057	4:11:45
France	26,786	4:04:39
Spain	19,456	5:30:55
Italy	18,256	6:00:07
Australia	9,895	6:52:28
Switzerland	2,451	3:54:34

Source: The Nielsen Company

Text: a source of information that requires specific treatments

Massive, diverse and under-exploited scientific information
Raising specific questions of access, analysis and interpretation

Handled at INRA by automatic *text-mining* methods
deployed on the new European infrastructure **OpenMinTeD**

An example

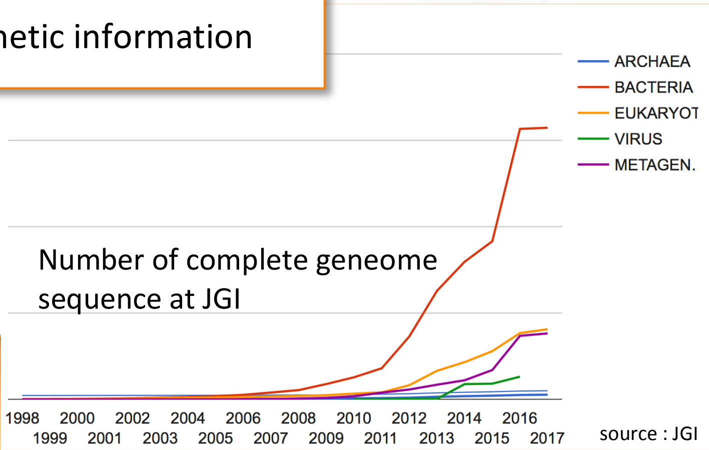
- microbial biodiversity

Microorganisms, food and scientific litterature

Billions of microorganisms everywhere, mostly unknown.
Play a critical role in food quality and transformation and its effect on health.
Microbiology research study their ecosystem and genetics for a better understanding, control and use.

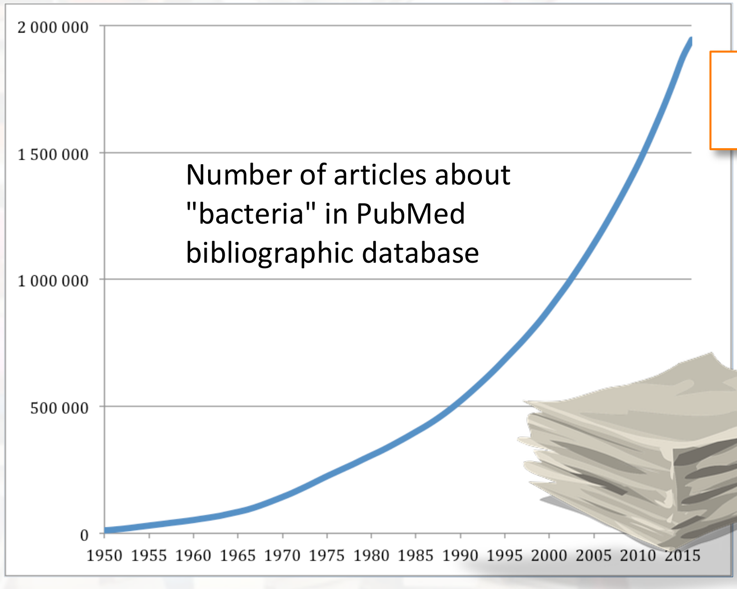
Ecosystems, habitats, properties
in millions of documents

Exponential growth of
genetic information



Number of complete geneome
sequence at JGI

source : JGI



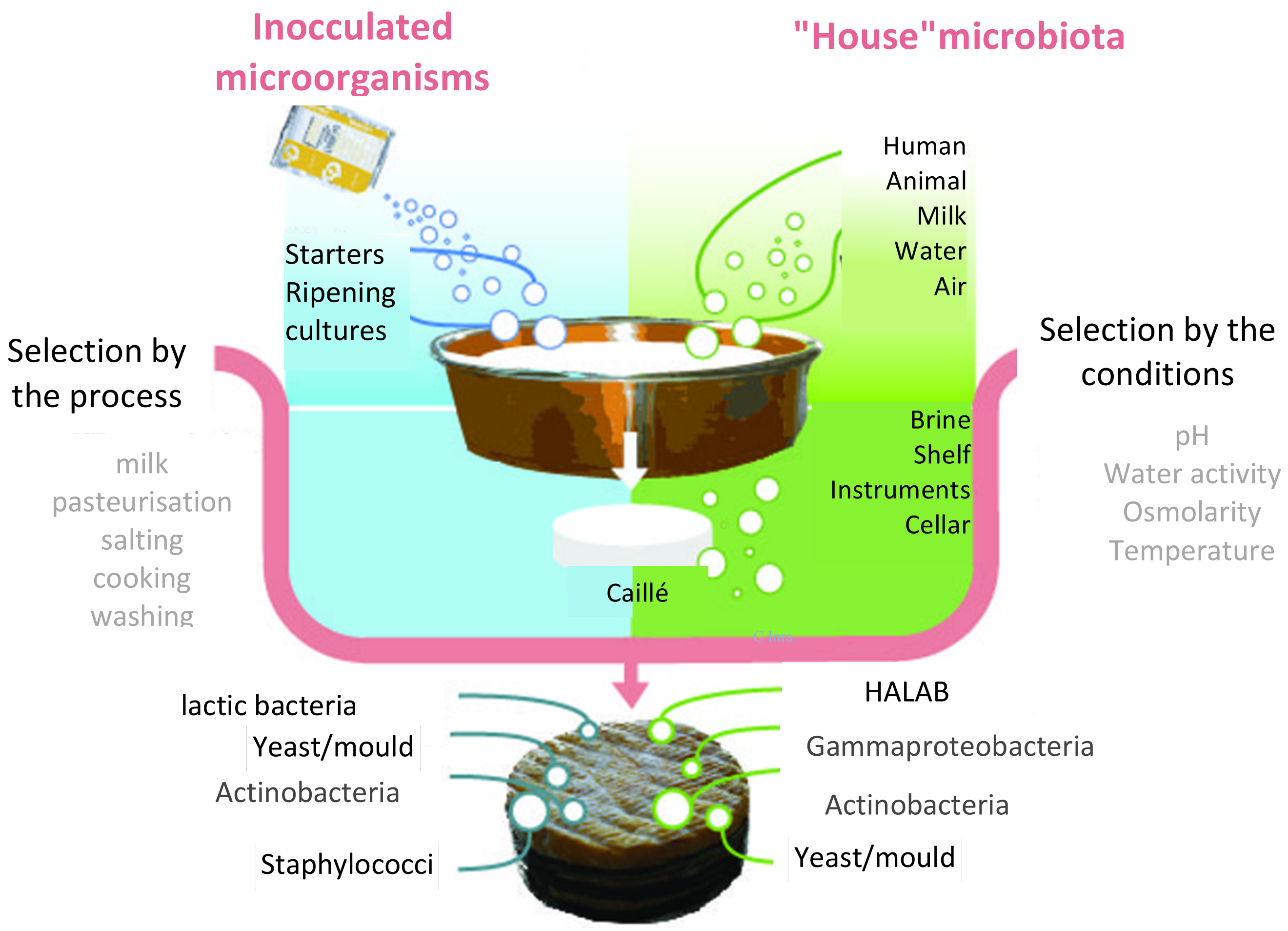
Number of articles about
"bacteria" in PubMed
bibliographic database

... and publications

Pixabav

http://cm1douzant.blogspot.fr/2014/11/conte-des-droits-des-enfants-2.html

What microbes in my cheese?



Inocculated microorganisms

"House"microbiota

Starters
Ripening
cultures

Human
Animal
Milk
Water
Air

Selection by the process

milk
pasteurisation
salting
cooking
washing

Selection by the conditions

pH
Water activity
Osmolarity
Temperature

Brine
Shelf
Instruments
Cellar

Caillé

lactic bacteria
Yeast/mould
Actinobacteria
Staphylococci

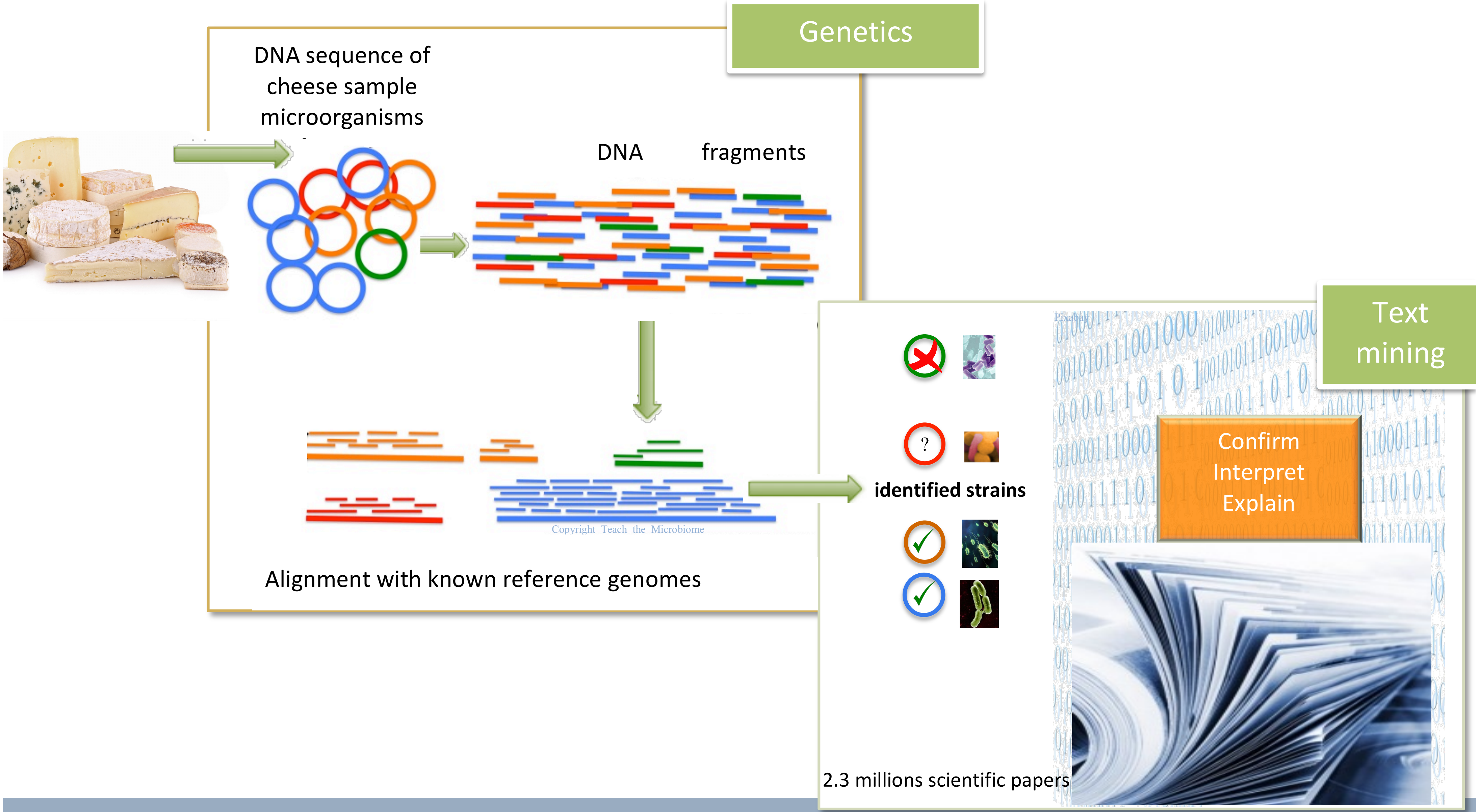
HALAB
Gammaproteobacteria
Actinobacteria
Yeast/mould

Cheese microbiota

Irlinger et al., FEMS Microbiol Lett (2015) 362 (2).

FoodMicrobiome Transfert project (CNIEL, Micalis, MaIAGe)

Identification of microorganisms by their DNA



Genetics

DNA sequence of cheese sample microorganisms

DNA fragments

Copyright Teach the Microbiome

Alignment with known reference genomes

?

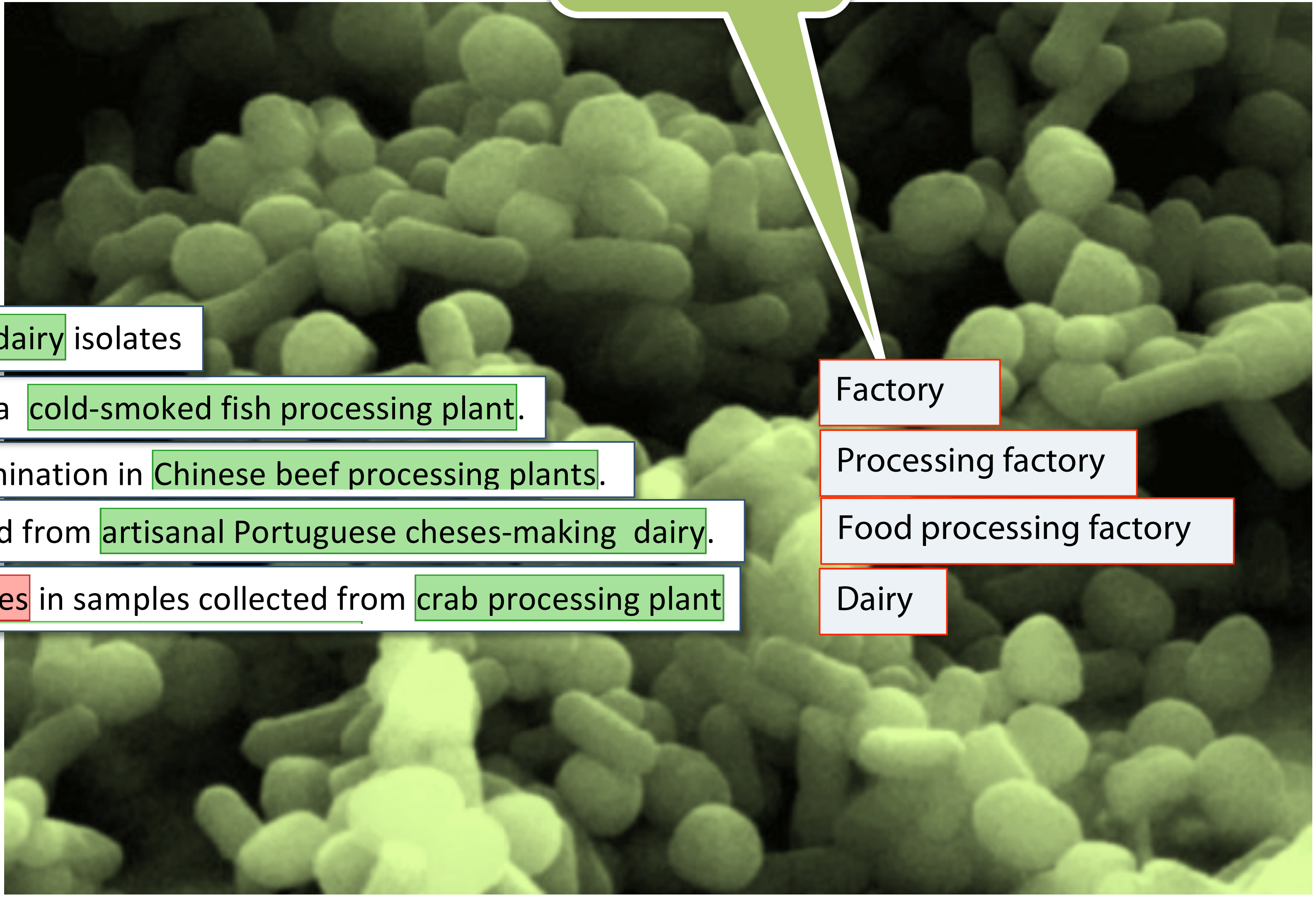
identified strains

Text mining

Pixabay

Confirm
Interpret
Explain

2.3 millions scientific papers



10,000 habitats of *Listeria monocytogenes* described in PubMed

Extracts of scientific articles

two L. monocytogenes cheese dairy isolates

L. monocytogenes persisting in a cold-smoked fish processing plant.

Listeria monocytogenes contamination in Chinese beef processing plants.

Listeria monocytogenes isolated from artisanal Portuguese cheses-making dairy.

the presence of L. monocytogenes in samples collected from crab processing plant

Habitat classes from ontology

Factory

Processing factory

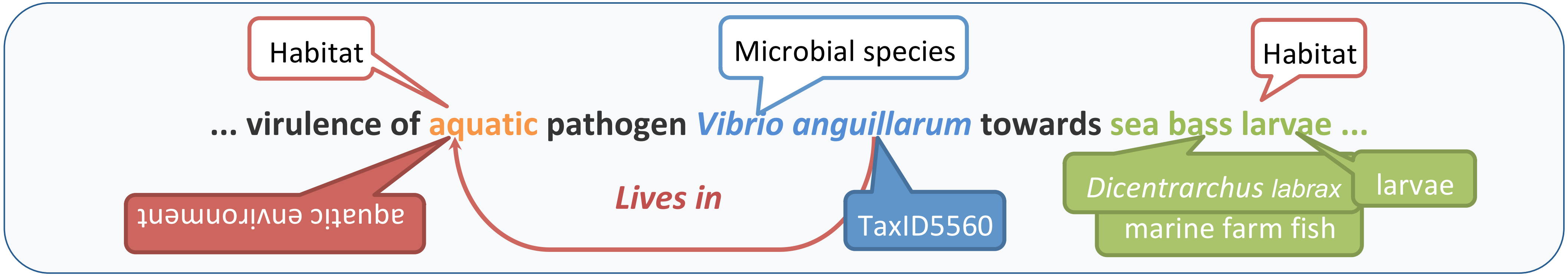
Food processing factory

Dairy

TDM challenge:

transform millions of text extracts into formal information despite the high variability

Information extraction



... virulence of aquatic pathogen *Vibrio anguillarum* towards sea bass larvae ...

Habitat

Microbial species

Habitat

aquatic environment

Lives in

TaxID5560

Dicentrarchus labrax

larvae

marine farm fish

1. Entity recognition = detection (text boundaries) and broad type assignment

2. Entity normalization = assignment to a category from a large set, >2,500 in OntoBiotope ontology

3. Relationship prediction = links entities together over sentences, microorganism to their properties

Artificial Intelligence methods

Natural language processing for semantics analysis

Machine learning for generalization from examples

Knowledge-intensive approach to deal with sparse small data

http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE
Ratkovic et al., *BMC Bioinformatics*, 2012
Nédellec et al., *Microbial diversity, EGC* 2018

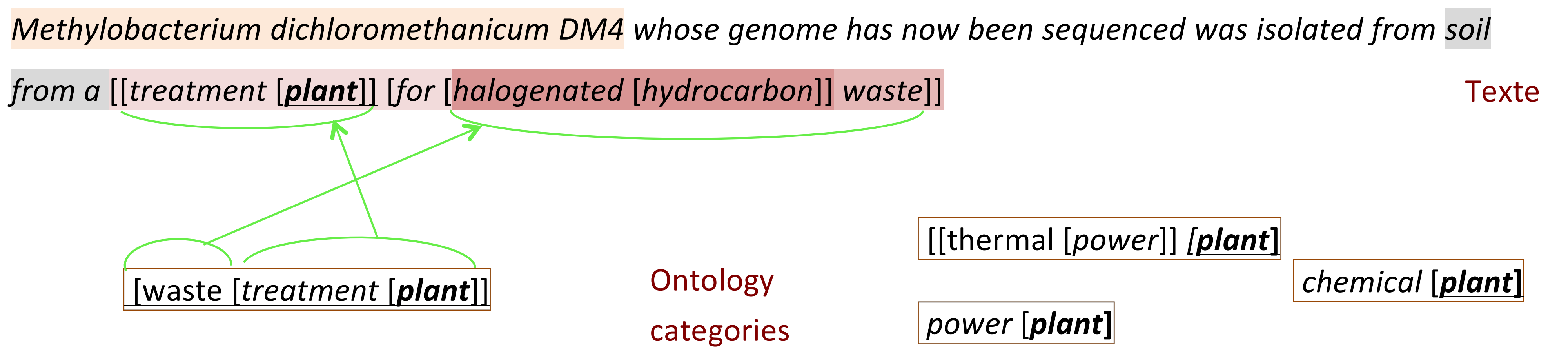
Entity recognition and normalisation

Entity recognition by term extraction by BioYateA

Entity normalization

by *Honor*, a 2-steps method to map the text terms to the labels of the ontology categories

Step 1. *ToMap* computes term similarity using syntactic structures (syntactic heads and subterms) and word similarities



Methylobacterium dichloromethanicum DM4 whose genome has now been sequenced was isolated from soil from a [[*treatment* [***plant***]] [*for* [*halogenated* [*hydrocarbon*]] *waste*]]

Texte

[waste [*treatment* [***plant***]]

Ontology categories

[[thermal [*power*]] [***plant***]

chemical [***plant***]

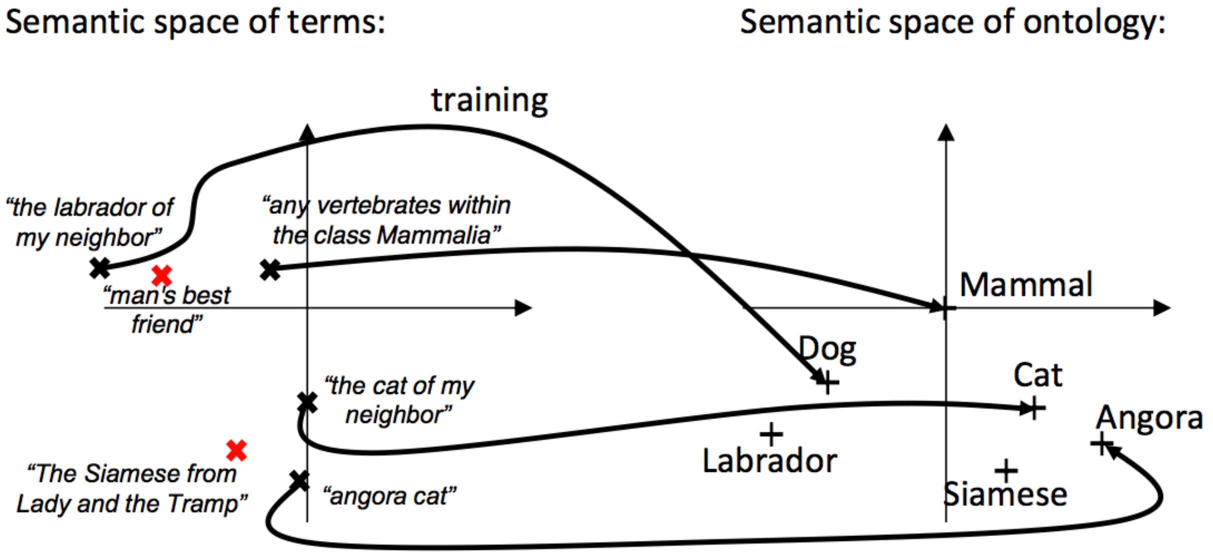
power [***plant***]

Ratkovic et al., BMC Bioinformatics, 2012

Golik et al., CiCLING, 2013

Entity recognition and normalisation

Step 2. *Contes* learns the projection of the term vector space to the ontology category vector space
by linear regression
from training examples
using the ontology structure


Semantic space of terms:
Semantic space of ontology:
training
"the labrador of my neighbor"
"man's best friend"
"any vertebrates within the class Mammalia"
"the cat of my neighbor"
"The Siamese from Lady and the Tramp"
"angora cat"
Mammal
Dog
Cat
Labrador
Angora
Siamese

System	Score
HONOR with domain specific heuristics	**0.73**
Distant supervised HONOR	**0.72**
ToMap with domain specific heuristics	0,66
Turku (2017)	0.63
BOUN (2016)	0.62
ToMap	0,61
CONTES	0.61

Honor system over performs other state of the art systems, as measured on Bacteria Biotope BioNLP-Shared Task 2016

Deléger et al., *BioNLP*, 2016

Ferré et al., *BioNLP* 2017 ; Ferré et al., *LREC* 2018

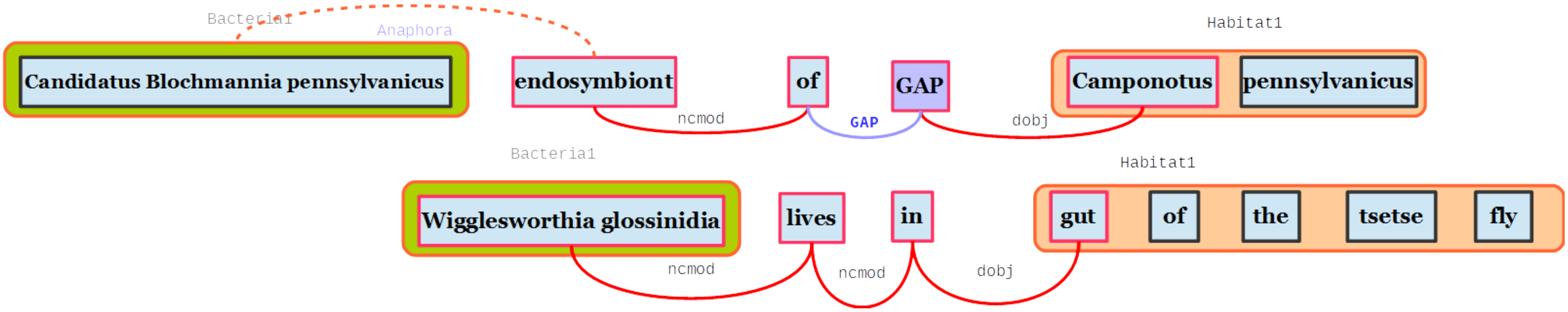
Automatic extraction of binary relationships, AlvisRE

System	PRE	REC	F-M
AlvisRE	0.51	**0.70**	**0.59**
Boun revised	0.52	0.53	0.53
LIMSI revised	0.42	0.60	0.49
TEES-2.1-2	**0.82**	0.28	0.42
IRISA-TexMex	0.46	0.36	0.40
Boun	0.38	0.21	0.27
LIMSI	0.19	0.04	0.06

BioNLP ’13: Bacteria Biotopes - Task 2

Machine learning method based on *shortest dependency path kernel*

- Dependency path computed by CCG and abstracted by Alvis Grammar.
- Anaphora resolution
- Word distance based on word embeddings
- Global alignments computed between dependency pathes (edit distance allowing gaps + Needleman-Wunsch dynamic programming algorithm)
- Empirical Kernel Map transformation

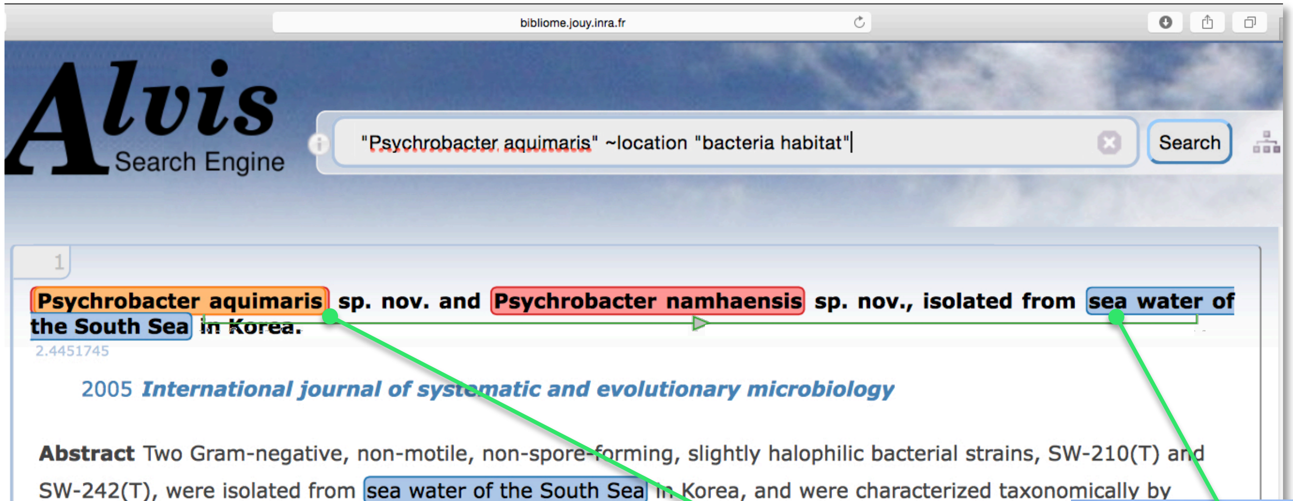


Bossy et al., BMC Bioinformatics, 2015
Ratkovic, PhD thesis, 2014
Valsamou, PhD thesis, 2017

Implementation and use

- All methods implemented in interoperable tools,
- combined into TDM workflows on **OpenMinTeD platform**
- Services on microbiology available at **IFP Migale Bioinformatics** plaform
    - Semantic search engine **AlvisIR**
    - **Florilege** database
- *OntoBiotope* ontology available on **AgroPortal**

Back to microbiology: text-mining explains *Psychrobacter aquimaris* presence in cheese samples


bibliome.jouy.inra.fr

Alvis Search Engine

"Psychrobacter aquimaris" ~location "bacteria habitat"

Search

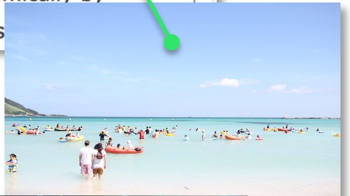
1

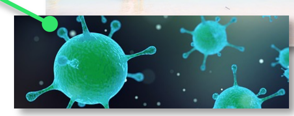
Psychrobacter aquimaris sp. nov. and Psychrobacter namhaensis sp. nov., isolated from sea water of the South Sea in Korea.

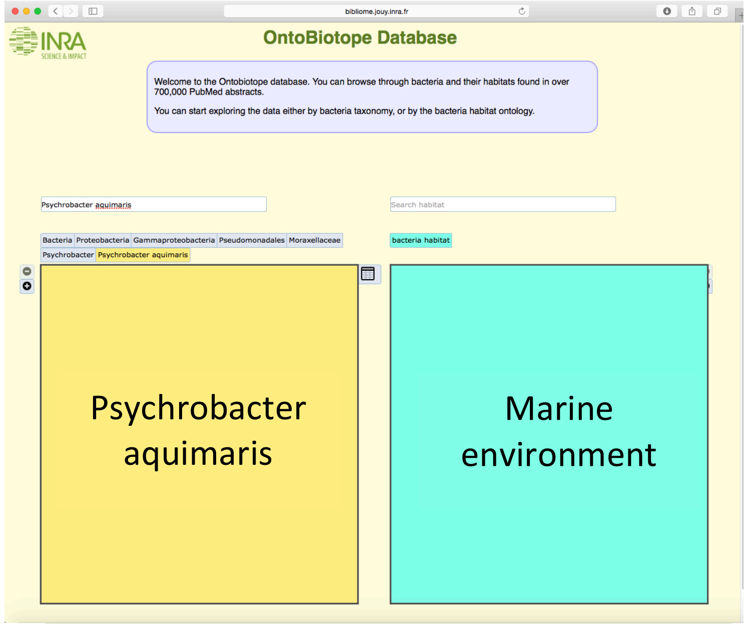
2.4451745

2005 ***International journal of systematic and evolutionary microbiology***

Abstract 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 25-30 degrees C and in the pres






bibliome.jouy.inra.fr

INRA SCIENCE & IMPACT

OntoBiotope Database

Welcome to the Ontobiotope database. You can browse through bacteria and their habitats found in over 700,000 PubMed abstracts.

You can start exploring the data either by bacteria taxonomy, or by the bacteria habitat ontology.

Psychrobacter aquimaris

Search habitat

Bacteria Proteobacteria Gammaproteobacteria Pseudomonadales Moraxellaceae Psychrobacter Psychrobacter aquimaris

bacteria habitat

Psychrobacter aquimaris

Marine environment

Querying and parsing PubMed yields

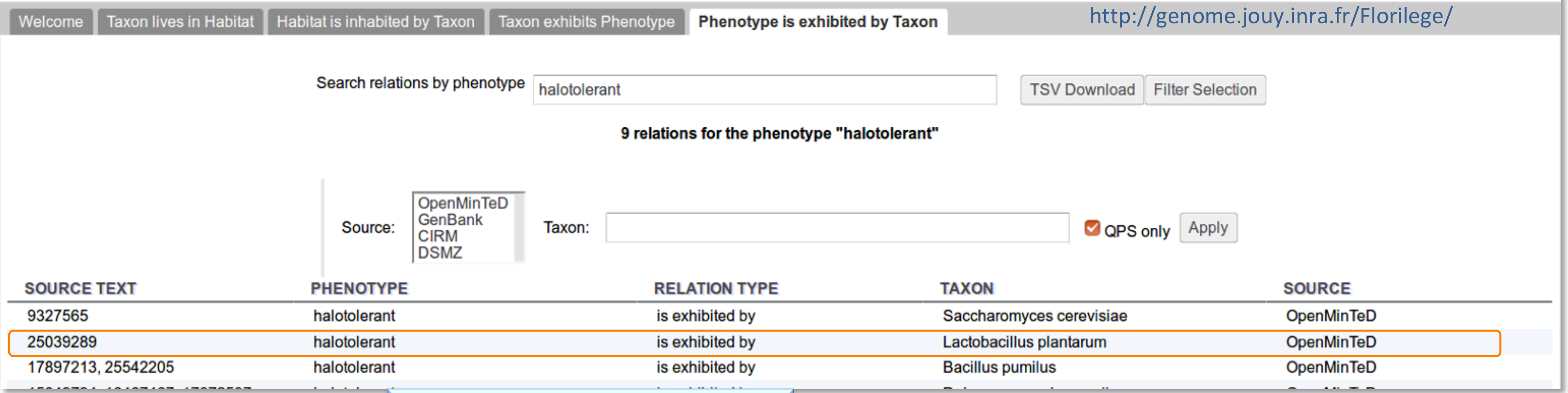
- 2,3 millions documents
- 8,3 microorganisms
- 18,5 millions habitats and phenotypes assigned to more than 2500 hierarchical classes
- 7,4 millions relationships

The researcher understands : added salt brings *P. aquimaris* to cheese

http://bibliome.jouy.inra.fr/demo/ontobiotope/alvisir2/webapi/search
http://genome.jouy.inra.fr/Florilege
Chaix et al., *Food Microbiology*, 2018

Using Florilege database for food research and innovation

what bacteria for a new salted cheese?



http://genome.jouy.inra.fr/Florilege/

Search relations by phenotype: halotolerant

9 relations for the phenotype "halotolerant"

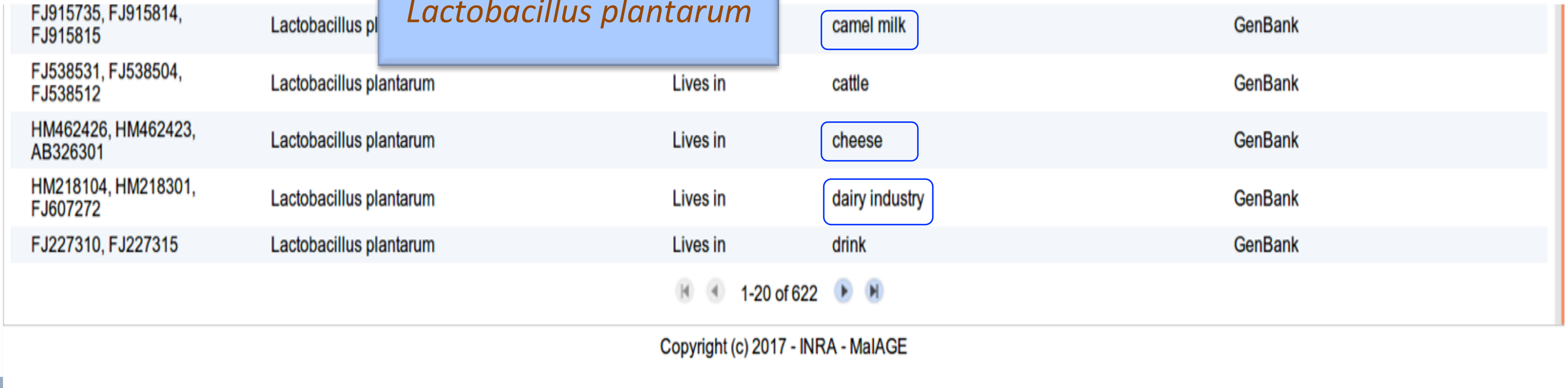
Source: OpenMinTeD / GenBank / CIRM / DSMZ

Taxon:

☑ QPS only

SOURCE TEXT	PHENOTYPE	RELATION TYPE	TAXON	SOURCE
9327565	halotolerant	is exhibited by	Saccharomyces cerevisiae	OpenMinTeD
25039289	halotolerant	is exhibited by	Lactobacillus plantarum	OpenMinTeD
17897213, 25542205	halotolerant	is exhibited by	Bacillus pumilus	OpenMinTeD

what known habitat for *Lactobacillus plantarum*



FJ915735, FJ915814, FJ915815	Lactobacillus pl		camel milk	GenBank
FJ538531, FJ538504, FJ538512	Lactobacillus plantarum	Lives in	cattle	GenBank
HM462426, HM462423, AB326301	Lactobacillus plantarum	Lives in	cheese	GenBank
HM218104, HM218301, FJ607272	Lactobacillus plantarum	Lives in	dairy industry	GenBank
FJ227310, FJ227315	Lactobacillus plantarum	Lives in	drink	GenBank

1-20 of 622

http://genome.jouy.inra.fr/Florilege

Chaix et al., *Food Microbiology*, 2018