INRA SCIENCE & IMPACT

openMINTED Open Mining Infrastructure for Text & Data

What can semantic text-mining do for food quality improvement?

Claire Nédellec, Robert Bossy

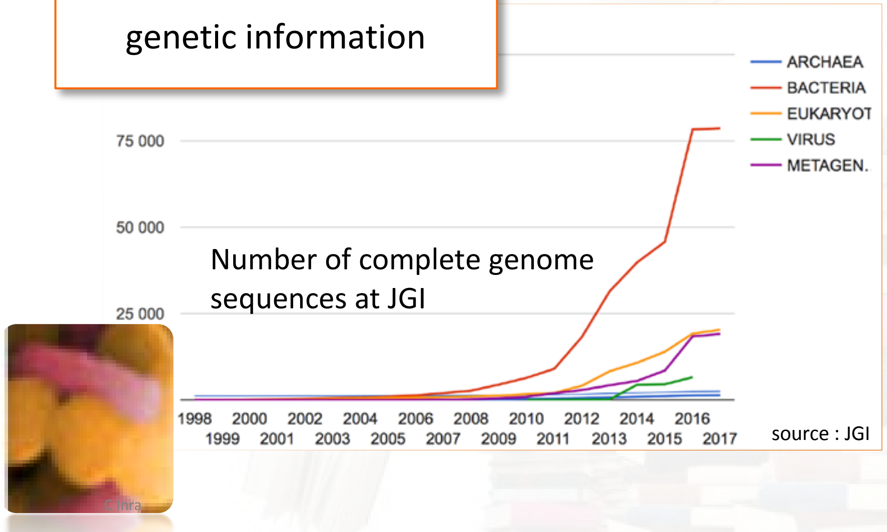
TDM: Unlocking a goldmine of information

Open Science Fair, Sept. 2017 Athens

Microorganisms, food and scientific literature

Billions of microorganims everywhere.
A critical role in all aspects of our life, *e.g.* food processing.
Researchers study their ecosystems and their genetics for better understanding, control, and use.

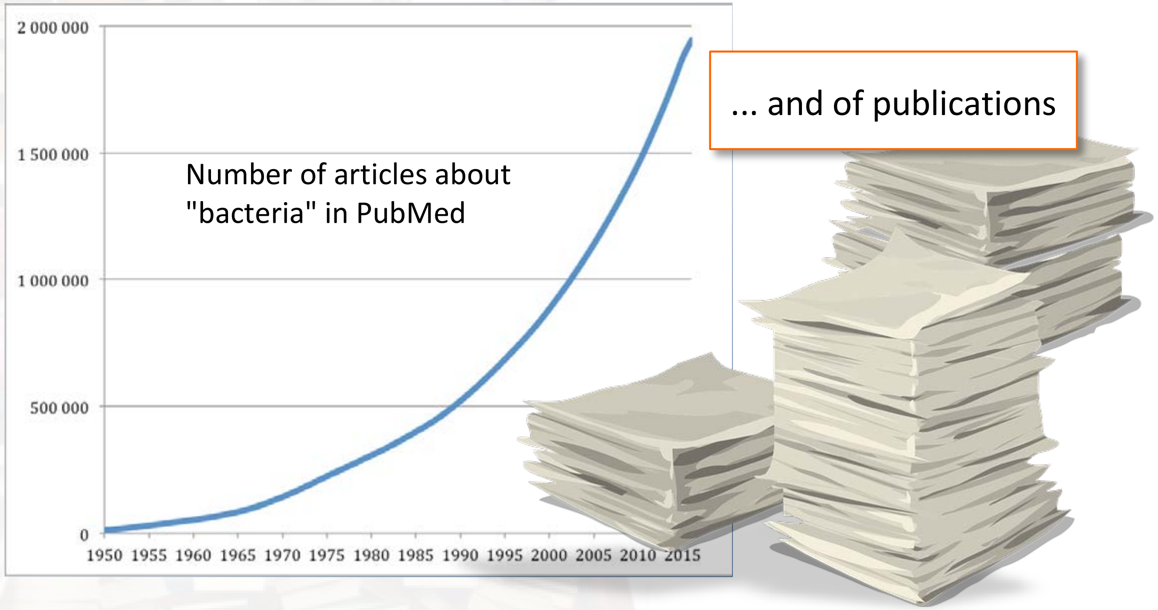
Exponential growing of genetic information



Number of complete genome sequences at JGI

source : JGI

Ecosystem, habitats, properties described in millions of free text



Number of articles about "bacteria" in PubMed

... and of publications

Pixabay

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

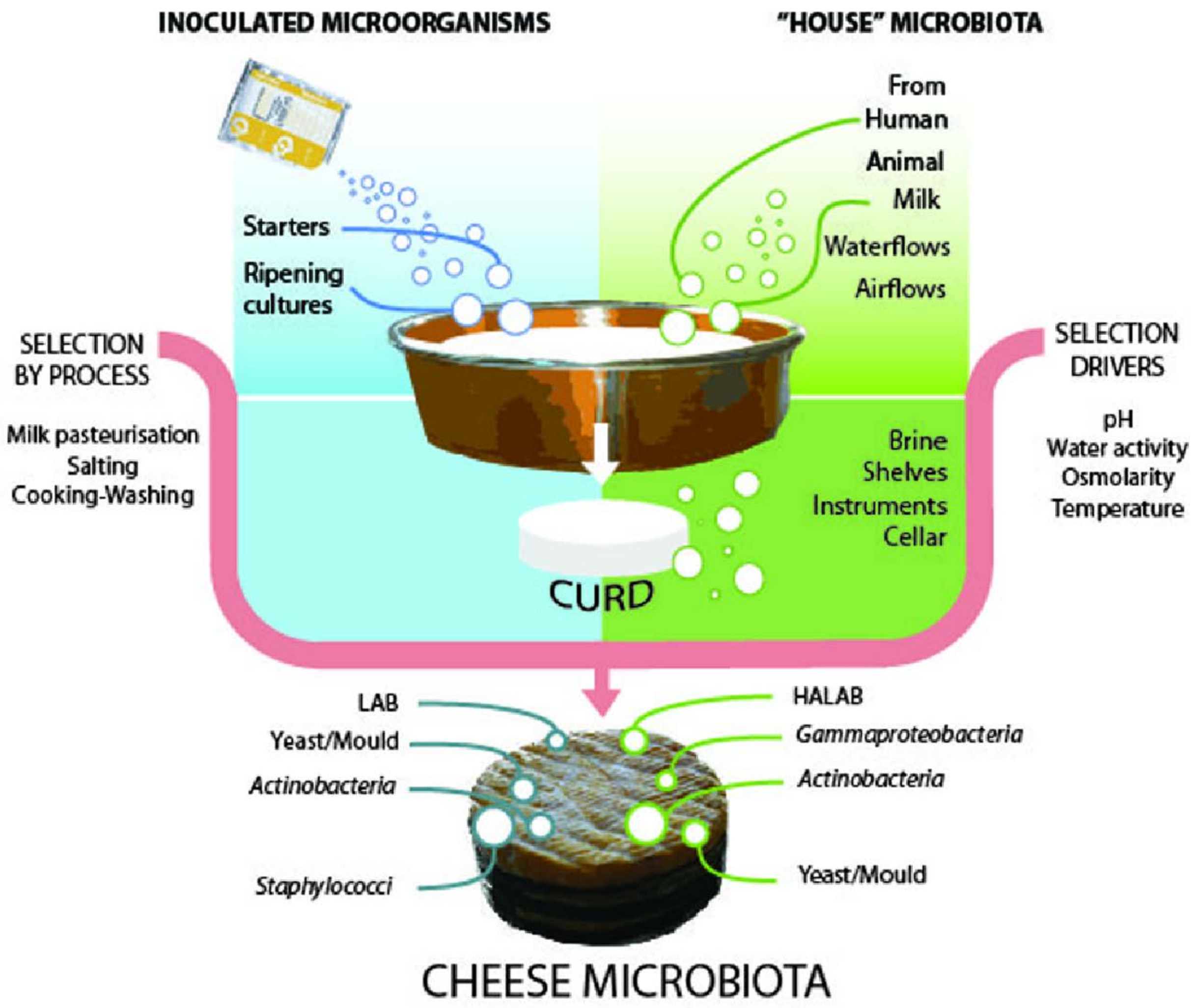
OntoBiotope, a semantic text-mining service

A shining example in cheese processing

An European Open Science perspective

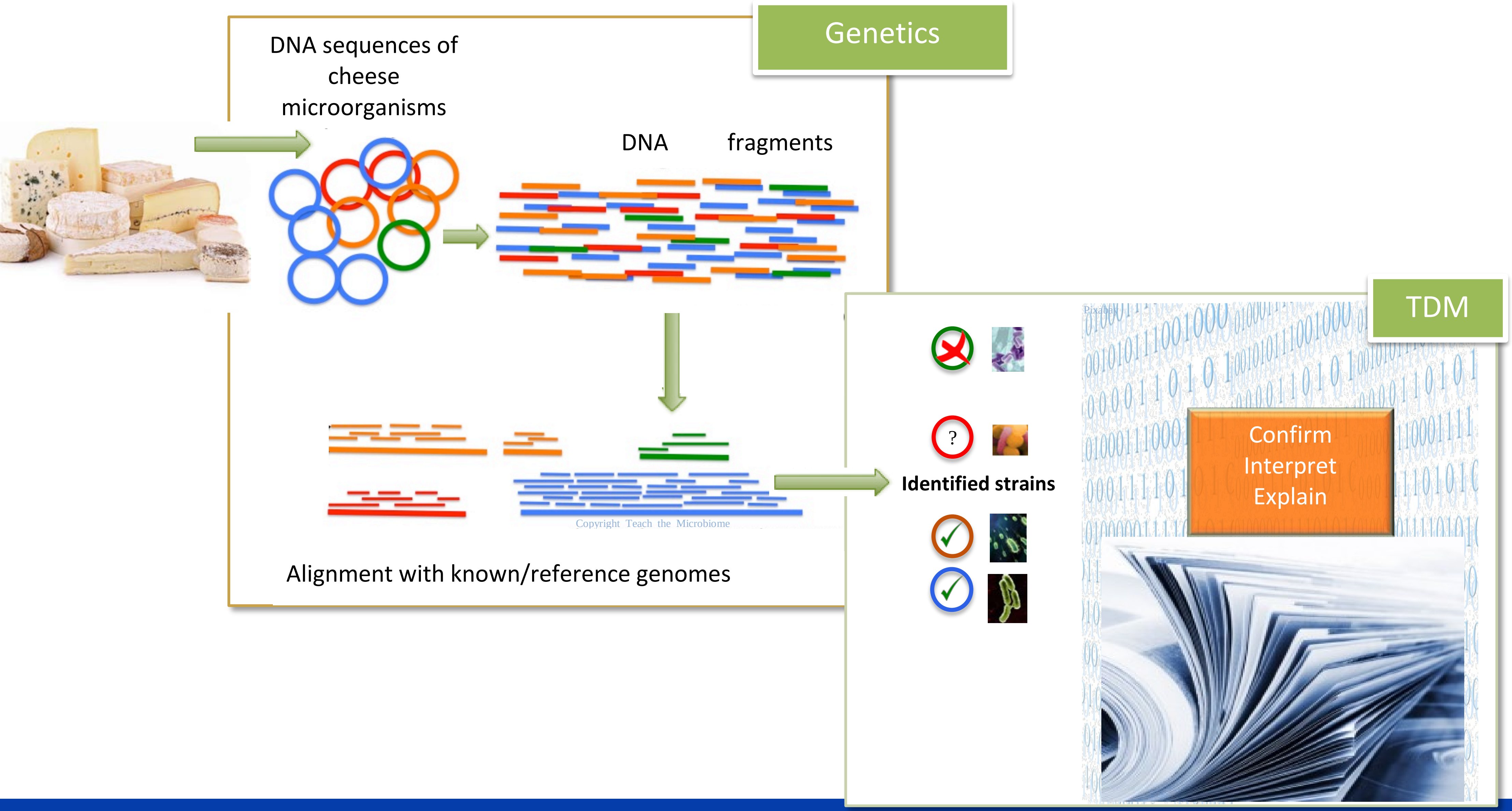
Opportunities and barriers

Microorganisms in my cheese?



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

DNA identification of microorganisms



Genetics

DNA sequences of cheese microorganisms

DNA fragments

Alignment with known/reference genomes

TDM

Identified strains

Confirm
Interpret
Explain

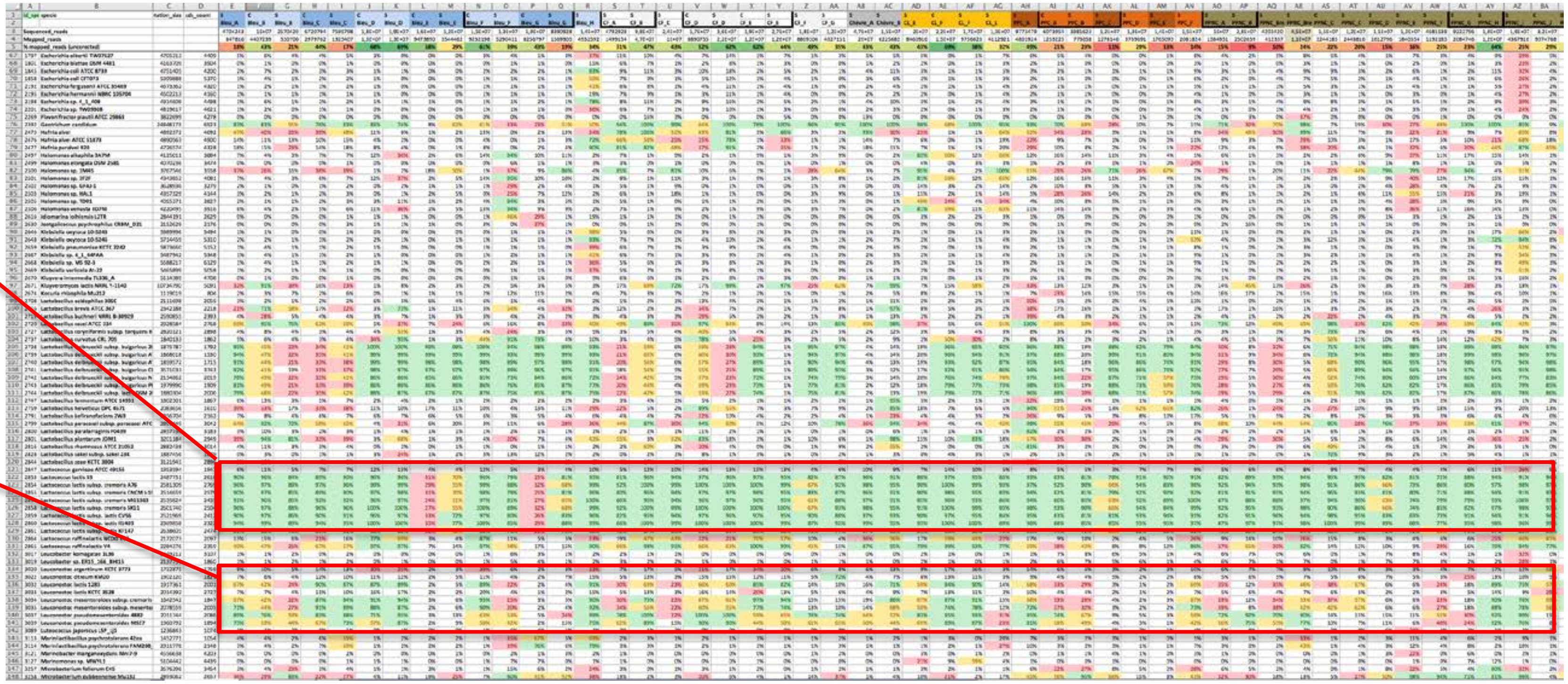
Metagenomics analysis of hundreds of French and Italian cheese samples

Identify microorganisms to understand and control their presence,

Improve quality of food product and design new ones

Inra - Cniel project

FoodMicrobiome Transfert



among 400 strains

🟩 **300** very frequent and well-known strains

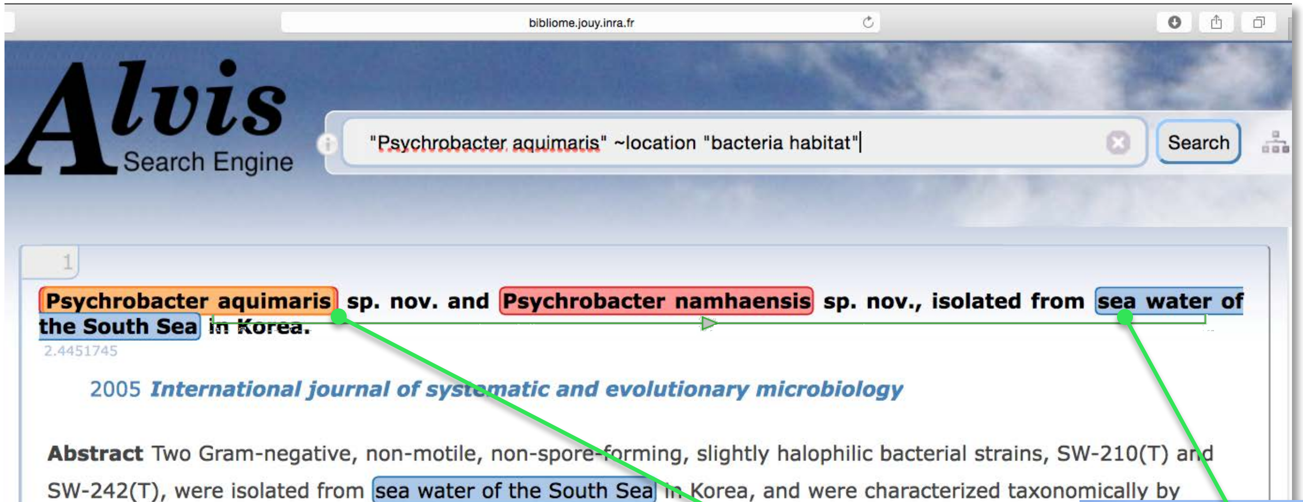
🟧 🟥 ⬜ **100** strains, little known and varyingly present

Psychrobacter aquimaris

ER15_174_BHI7

gorgonzola, roquefort, époisse, toscanello, st nectaire (very frequent), tomme, bleu

TDM explanation of Psychrobacter presence



bibliome.jouy.inra.fr

Alvis Search Engine

"Psychrobacter aquimaris" ~location "bacteria habitat"

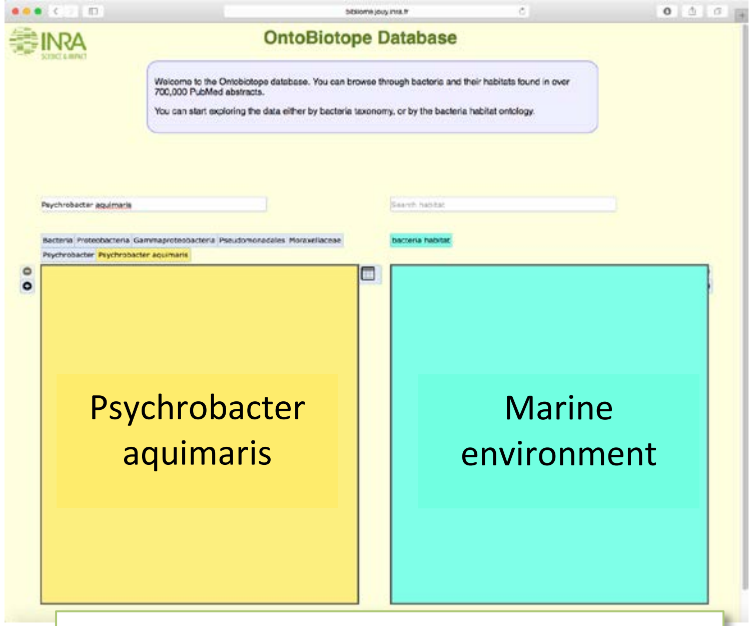
1

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

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

TDM OntoBiotope analyzed all PubMed information

- 1,16 millions de documents
- 3,63 millions of relations bacteria-habitat assigned to 2000 categories !
- prediction performances

Psychrobacter aquimaris comes from sea environment

The researcher understands :
the added salt brings the bacteria in the cheese

Many other good reasons to study food microbiome

Industrial interest

Better understanding and control of microbiome role in food process

Food innovation : transformation, preservation, antibiotic or nutrient production, flavour/taste

Public health

Better control of food spoilage and safety by explaining the microbial source and adaptation.

Improve intestinal microbiota by food intake.

Fundamental research

Better understanding of microorganism life and adaptation

Ecosystem dynamics

In food and related environments

A TDM based-service for food microbiome study

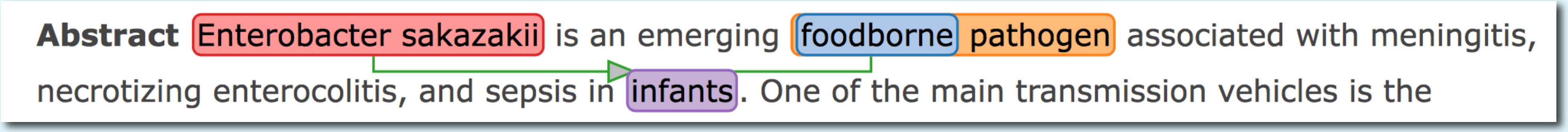
The TDM-based service developed by INRA

- automatically extracts information
- from massive amount of documents in all microbiology domains
- interoperable with experimental data using shared ontologies.

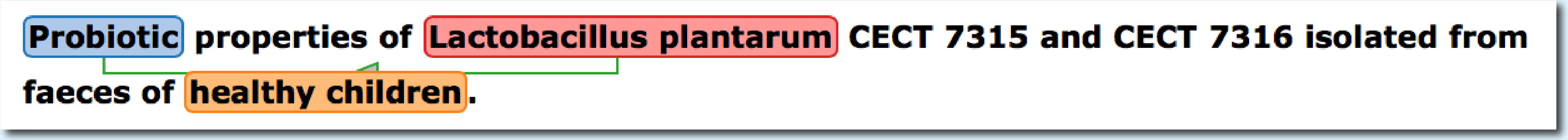
What information

- *who* is living: taxa, strain, species, families
- *where*: habitats of all kinds
- *how*: environment requirements, phenotype

Q?: bacteria livesin food and human



Abstract Enterobacter sakazakii is an emerging foodborne pathogen associated with meningitis, necrotizing enterocolitis, and sepsis in infants. One of the main transmission vehicles is the

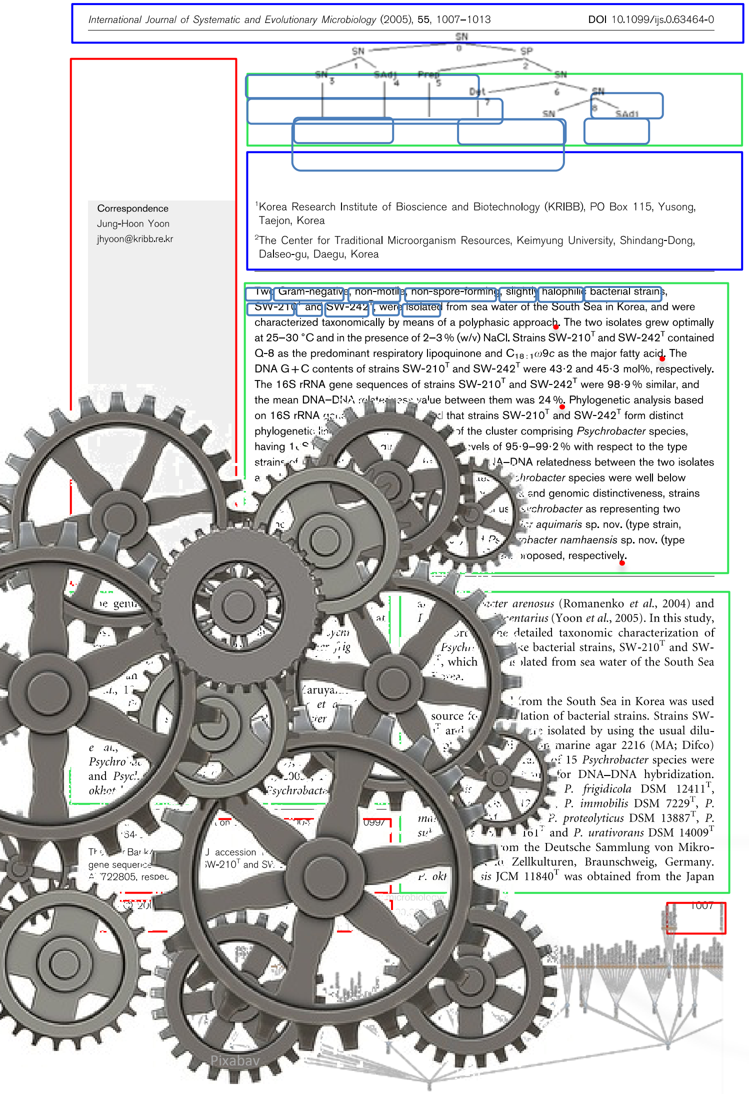


Probiotic properties of Lactobacillus plantarum CECT 7315 and CECT 7316 isolated from faeces of healthy children.

http://bibliome.jouy.inra.fr/demo/food/alvisir/webapi/search

A high-precision machinery

- Artificial intelligence
- Natural Language Processing
- Machine Learning



International Journal of Systematic and Evolutionary Microbiology (2005), **55**, 1007–1013 DOI 10.1099/ijs.0.63464-0

Correspondence
Jung-Hoon Yoon
jhyoon@kribb.re.kr

[1]Korea Research Institute of Bioscience and Biotechnology (KRIBB), PO Box 115, Yusong, Taejon, Korea

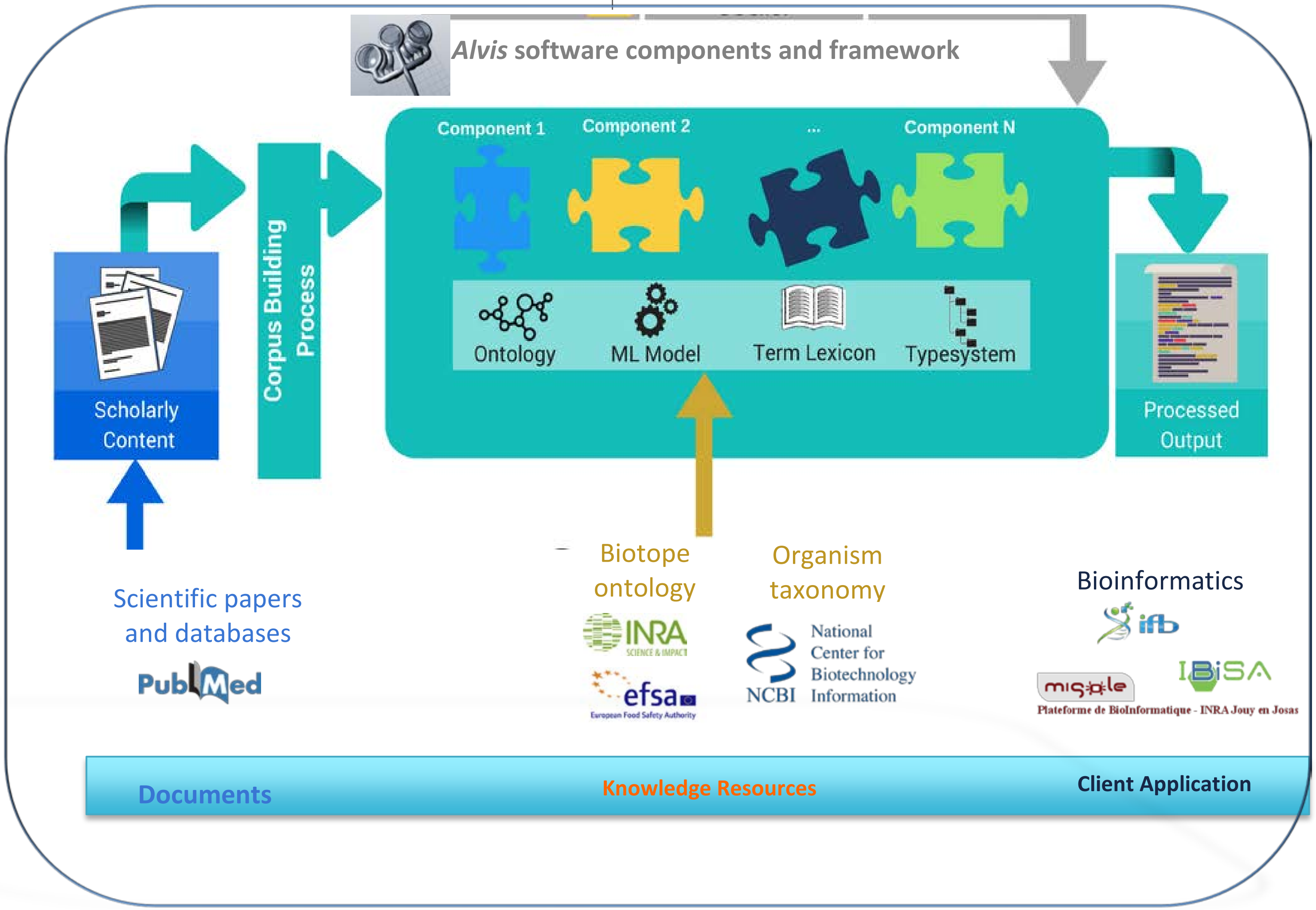
[2]The Center for Traditional Microorganism Resources, Keimyung University, Shindang-Dong, Dalseo-gu, Daegu, Korea

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 °C and in the presence of 2–3 % (w/v) NaCl. Strains SW-210^T and SW-242^T contained Q-8 as the predominant respiratory lipoquinone and $C_{18:1}\omega9c$ as the major fatty acid. The DNA G+C contents of strains SW-210^T and SW-242^T were 43·2 and 45·3 mol%, respectively. The 16S rRNA gene sequences of strains SW-210^T and SW-242^T were 98·9 % similar, and the mean DNA–DNA relatedness value between them was 24 %. Phylogenetic analysis based on 16S rRNA [illegible] that strains SW-210^T and SW-242^T form distinct phylogenetic [illegible] of the cluster comprising *Psychrobacter* species, having 16S [illegible] levels of 95·9–99·2 % with respect to the type strains of [illegible] DNA–DNA relatedness between the two isolates a[illegible] *Psychrobacter* species were well below [illegible] and genomic distinctiveness, strains [illegible] *Psychrobacter* as representing two [illegible] *aquimaris* sp. nov. (type strain, [illegible] *Psychrobacter namhaensis* sp. nov. (type [illegible] proposed, respectively.

[illegible] *arenosus* (Romanenko *et al.*, 2004) and [illegible] *marincentarius* (Yoon *et al.*, 2005). In this study, [illegible] the detailed taxonomic characterization of [illegible] *Psychrobacter*-like bacterial strains, SW-210^T and SW-[illegible], which [illegible] isolated from sea water of the South Sea [illegible] Korea.

[illegible] from the South Sea in Korea was used [illegible] source for [illegible] isolation of bacterial strains. Strains SW-[illegible] and [illegible] were isolated by using the usual dilu-[illegible] on marine agar 2216 (MA; Difco) [illegible] of 15 *Psychrobacter* species were [illegible] for DNA–DNA hybridization. [illegible] *P. frigidicola* DSM 12411^T, [illegible] *P. immobilis* DSM 7229^T, *P.* [illegible] *P. proteolyticus* DSM 13887^T, *P.* [illegible] 161^T and *P. urativorans* DSM 14009^T [illegible] from the Deutsche Sammlung von Mikro-organismen und Zellkulturen, Braunschweig, Germany. *P. okh*[illegible]*sis* JCM 11840^T was obtained from the Japan

1007



OntoBiotope in an European Open Science perspective

OntoBiotope service becomes an application of OpenMinTeD text-mining infrastructure
Benefits from

- **Full-text paper collection aggregation, standardisation**
    - *OpenAire*, *CORE*
- Guaranteed **computational resources** in a secure environment, virtual machines and monitoring capabilities
    - *Okeanos service of GRNet in EGI federated cloud*.
- **Semantic resources** aggregation, uniform access, standard representation, update
    - *AgroPortal* (integration *in progress,* Visa TM project)

OntoBiotope in an European Open Science perspective

Deployment of OntoBiotope TDM on OpenMinTed infrastructure offers to the scientific communities

Innovative research

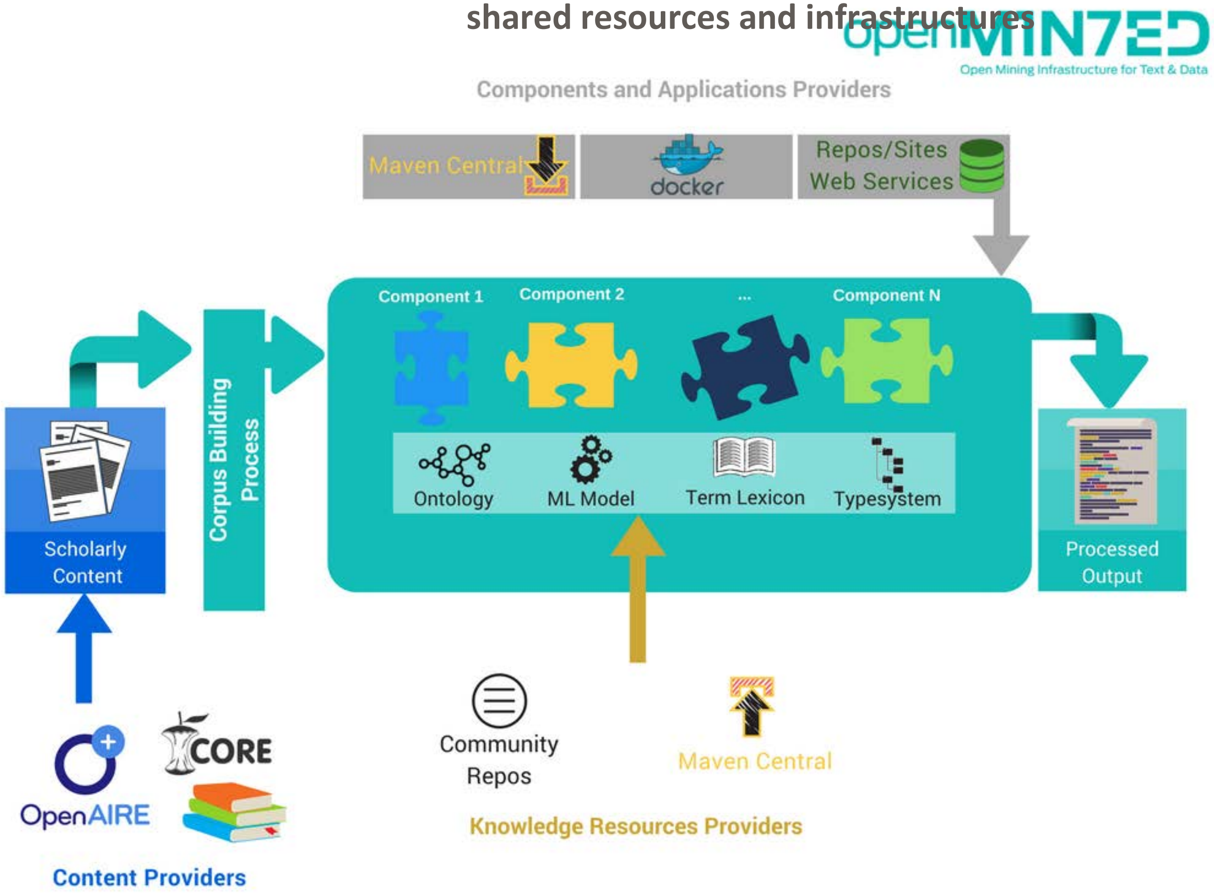
national, European, international

shared resources and infrastructures

A fully open access in a unified framework to the service, the processing workflows, the input data, the TDM final and intermediate results

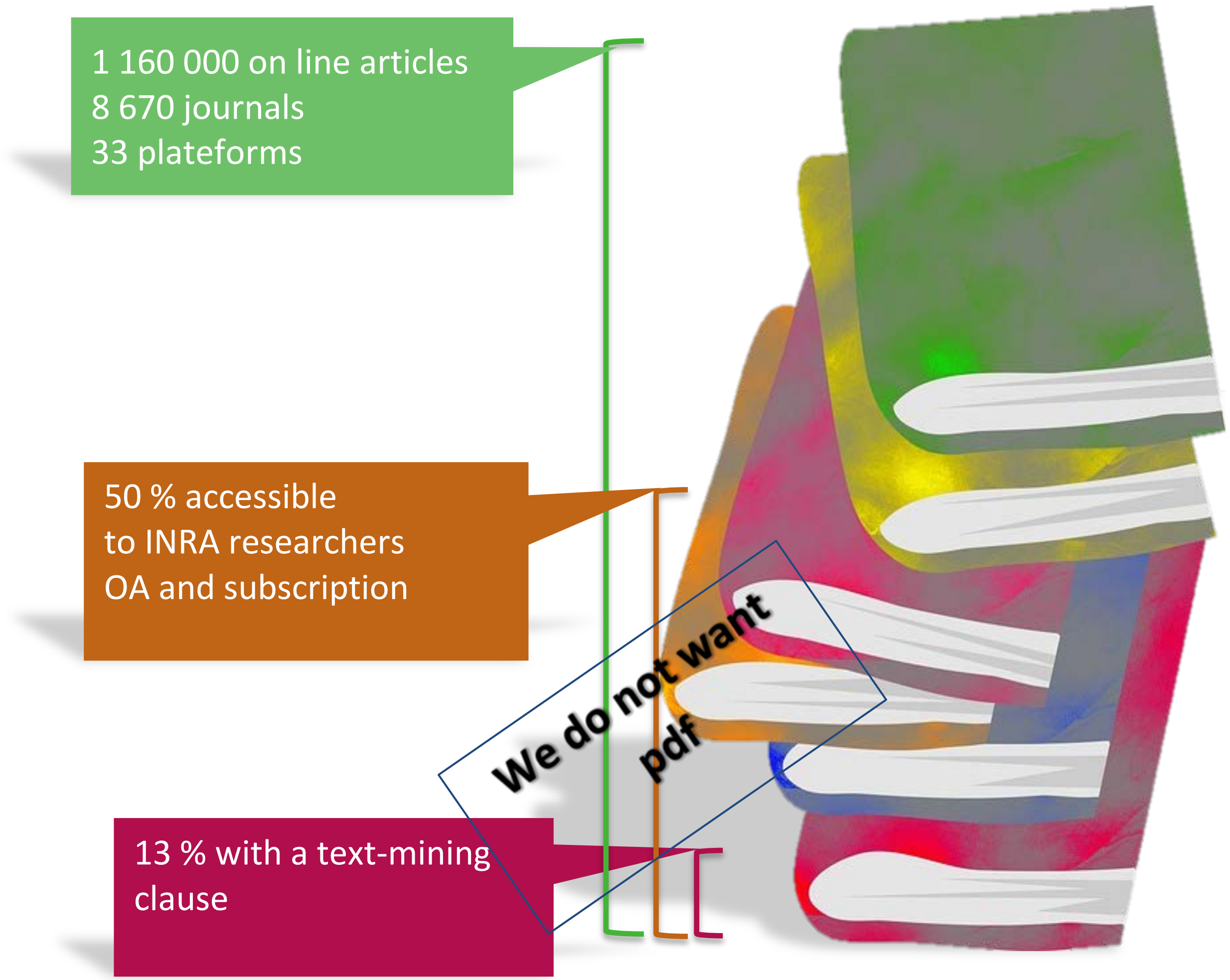
Reproducibility and adaptability.
Non TDM specialists (bioinformaticians) can reuse the workflows and replace subparts from component catalogue

The technological choices (Galaxy framework, Docker and Maven) make OMTD evolutive and interoperable.



Barriers and opportunities in the European Science Cloud

Extend document sources, remove legal and technical hindrances



What food microbiology researchers need is *all* the scientific public information

- Not only the one that is Open Access
- Not only the one that is automatically findable
- Not only the one that is in a standard parsable format

Barriers and opportunities in the European Science Cloud

Facilitating access and use of high-level services in AgroFood.

In Agriculture and Life Sciences, text-mining is not the end of the story

TDM services and results will be combined and integrated into wider data analysis applications

Inteleave text-mining and experimental data analysis

Experts may intervene in the analysis.

Improve OMTD virtual research environment to support collaboration and sharing data, knowledge models and workflows

FAIR data

Federating more data and knowledge from all fileds of food microbiology : genetics, biological resources, health, nutritional information, industrial process, distribution, retail, cooking