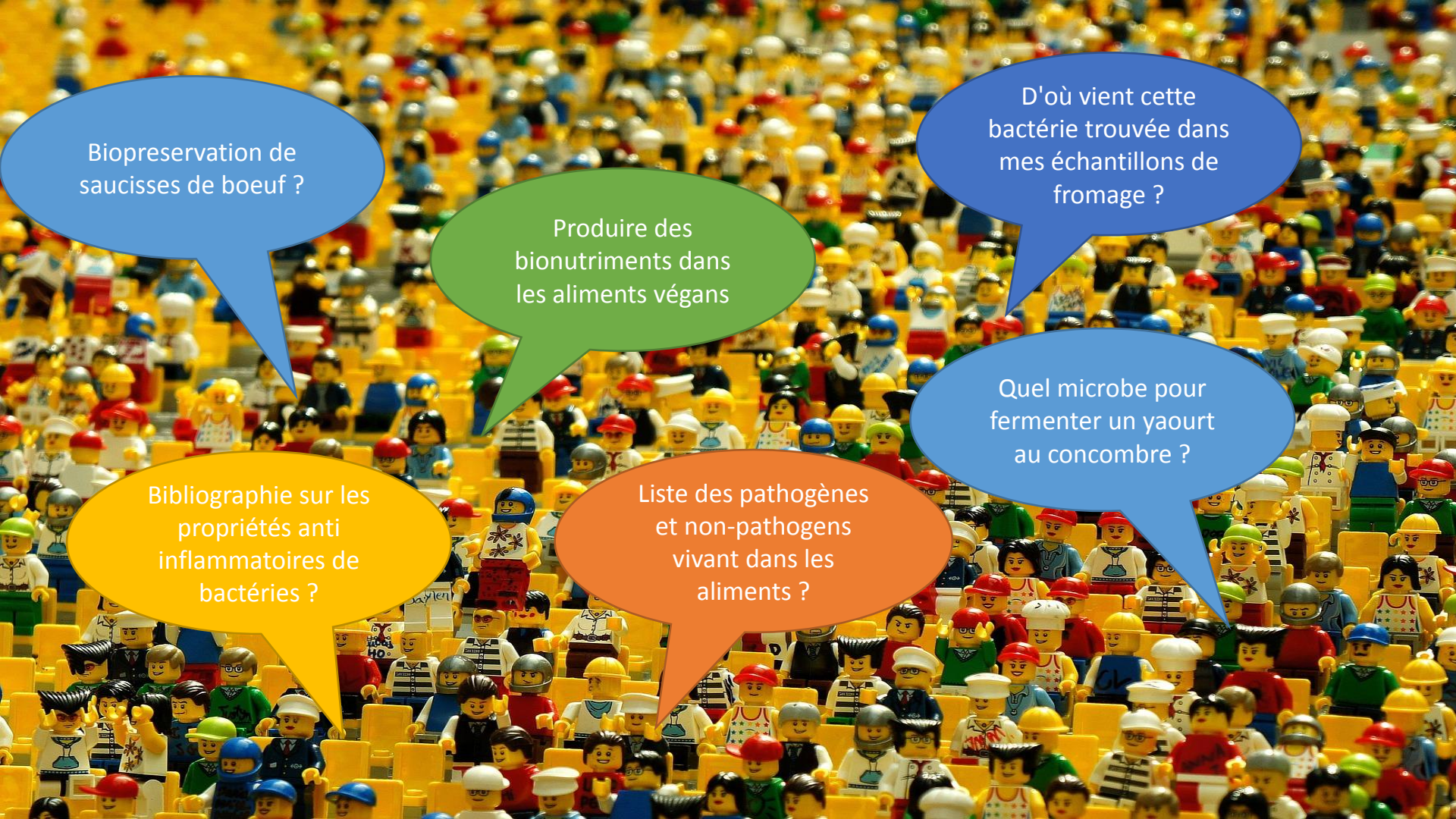


INRAE

- Exploiter les données textuelles de souches microbiennes pour de nouveaux produits alimentaires fermentés

Claire Nédellec (MaIAGE, dépt MathNum)



Biopreservation de saucisses de boeuf ?

Produire des bionutriments dans les aliments végans

D'où vient cette bactérie trouvée dans mes échantillons de fromage ?

Quel microbe pour fermenter un yaourt au concombre ?

Bibliographie sur les propriétés anti inflammatoires de bactéries ?

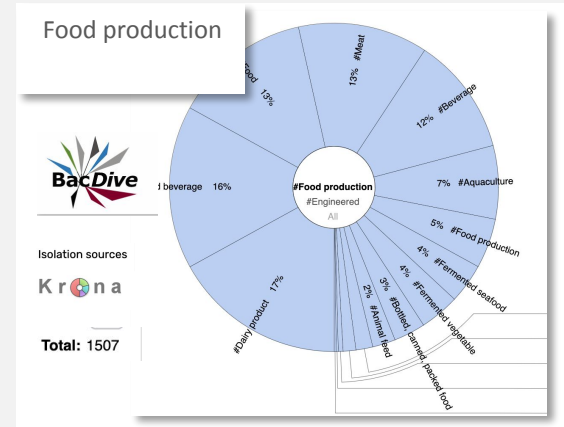
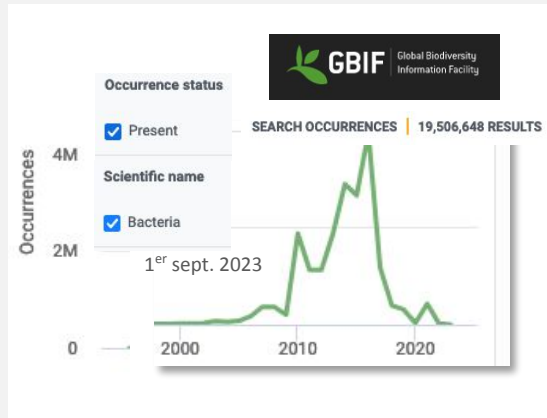
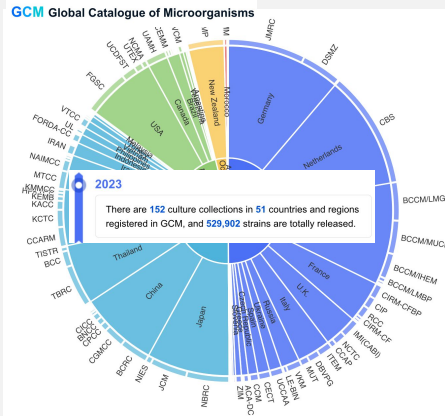
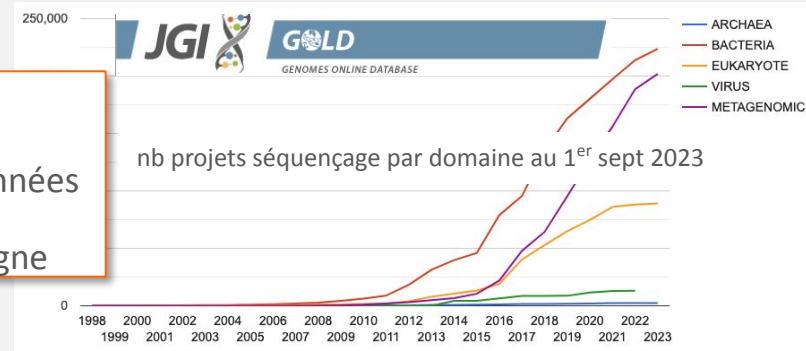
Liste des pathogènes et non-pathogens vivant dans les aliments ?

● Microorganismes, aliments et littérature

Des descriptions textuelles des écosystèmes, habitats, propriétés dans des millions de documents et bases de données

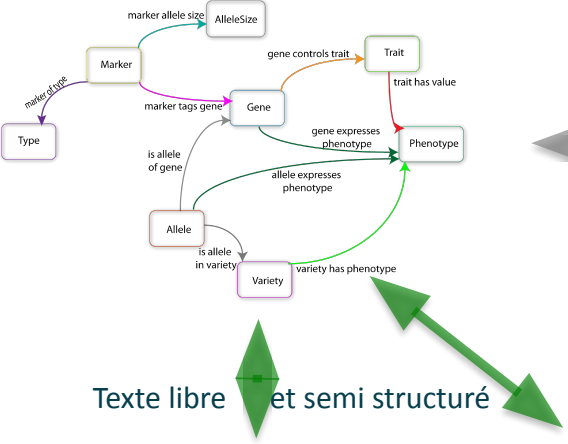


Une croissance exponentielle des données et des sources en ligne

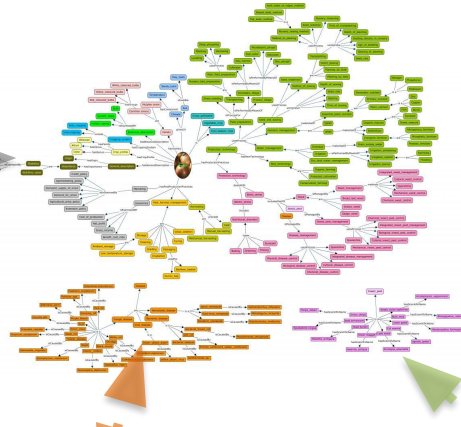


Intégrer les informations

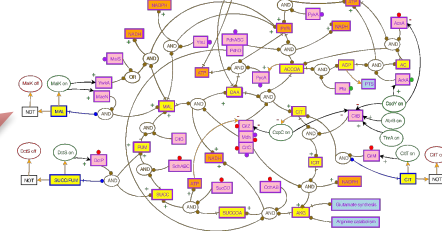
Modèle de connaissances du texte



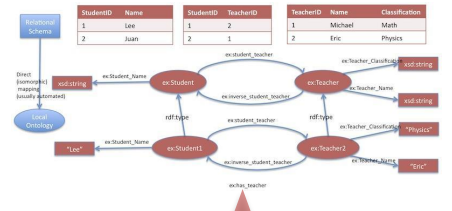
Une représentation formelle (ontologie) pour la gestion et le raisonnement



Modèle dynamique



Modèle de données



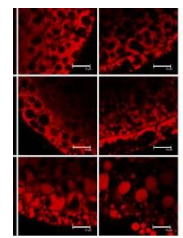
Texte libre et semi structuré

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a method for producing a starter for saki, comprising the steps of: (a) providing a yeast strain; (b) culturing the yeast strain in a medium containing a substrate and a nutrient source; (c) harvesting the yeast cells; and (d) drying the yeast cells to form a powder.

- Sample type/isolated from irradiated ground pork and beef
- Sample type/isolated from stool of breast-fed infant
- Sample type/isolated from "Moto" starter of saki
- Sample type/isolated from Sake starter/Moto
- Sample type/isolated from Moto (starter of saki)

Image



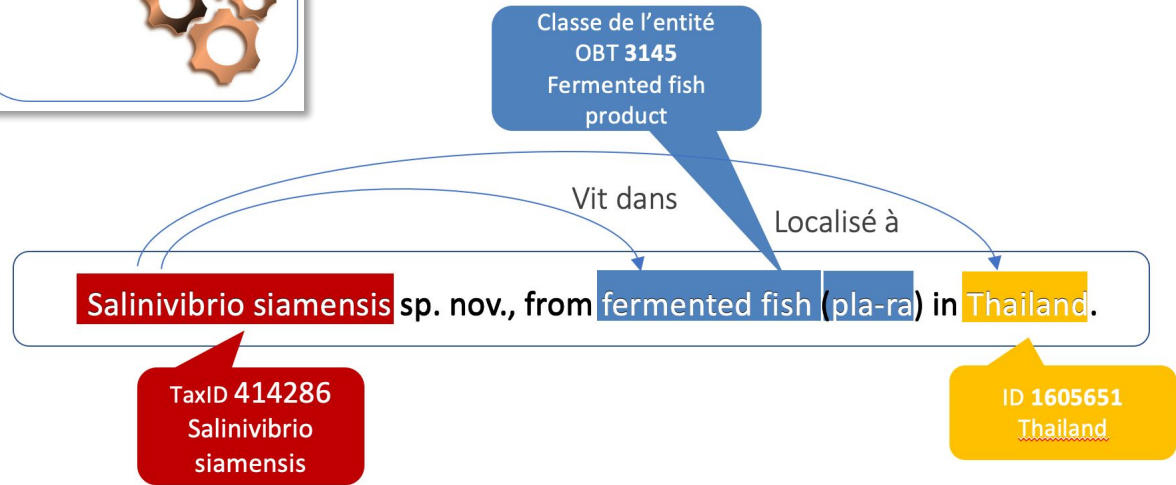
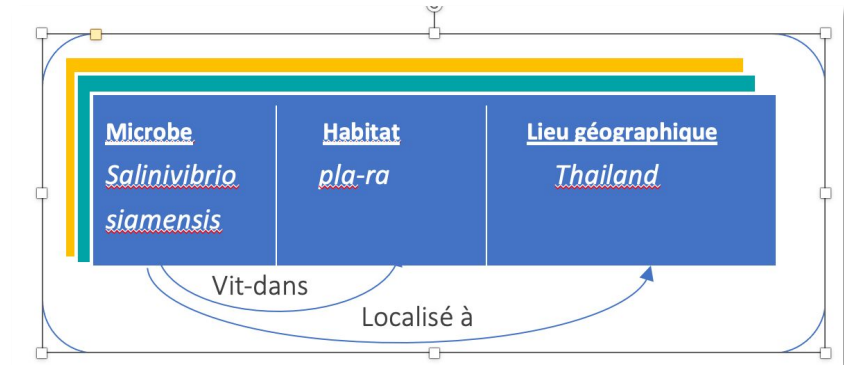
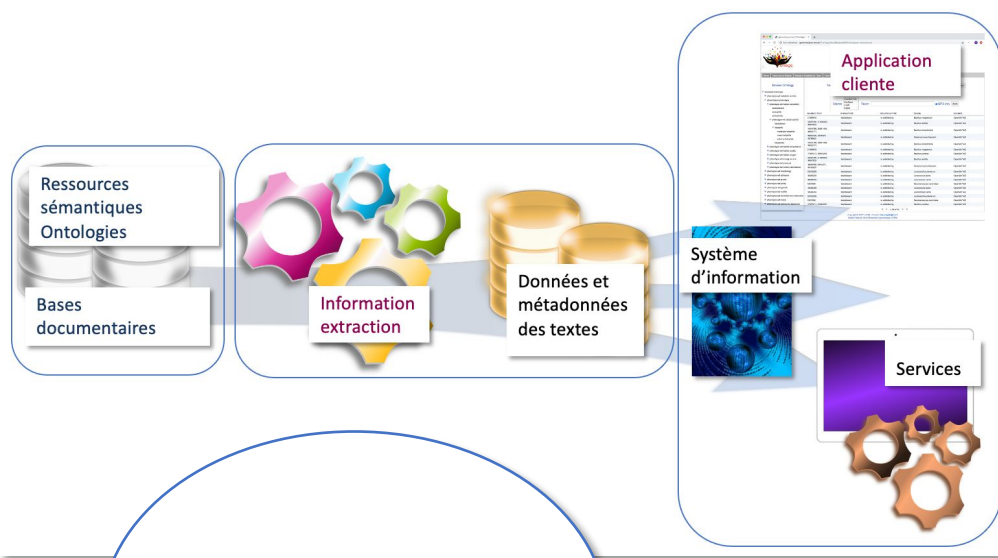
Extraction d'information textuelle
13 sept 2023 / Ferment'IA / C. Nédellec

Données

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

● Extraction d'information



INRAE

Extraction d'information textuelle
13 sept 2023 / Ferment'IA / C. Nédellec

Des données et des référentiels divers et dispersés



● Méthodes et verrous



Traitement automatique de la langue

- la révolution des réseaux de neurones profonds (*deep learning*)
- dans les domaines de spécialités, peu ou pas exemples d'apprentissage annotés manuellement
- stratégies de transfert et d'exploitation de connaissances externes

Focus sur deux méthodes

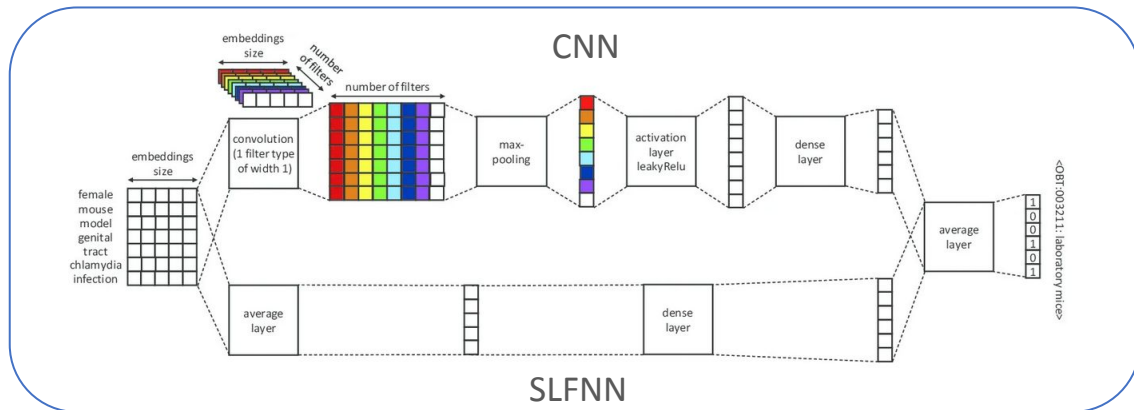
C-Norm : normalisation d'entités avec peu d'exemples

KB-PubMedBERT : extraction de relations à l'aide d'une base de connaissance externe

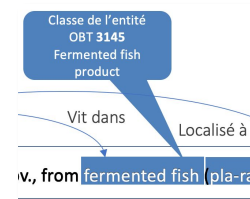


● Normalisation des entités avec C-Norm

Ferré, A., et al. C-Norm: a neural approach to few-shot entity normalization. *BMC Bioinformatics* (2020).
<https://doi.org/10.1186/s12859-020-03886-8>



Problème d'apprentissage fortement multi-classes et à faible nombre d'exemples (*few-shot*)



C-Norm

- Supervision faible
- Intégration des connaissances ontologiques
- Sémantique distributionnelle (Word2Vec)

Les résultats expérimentaux [Ferré et al., 2020] surpassent ceux des autres méthodes sur la tâche *BB-norm task* de Bacteria Biotope'19, [Bossy et al., BioNLP-OST 2019] avec ~3000 classes d'habitats et ~ 400 classe de phénotypes

	Distance de Wang	F1 strict
<u>Habitats</u>	C-Norm 77,7	60,4
[Deng et al., 2019]	PADIA 68,4	48,8
<u>Phénotypes</u>	C-Norm 88,1	70,0
	PADIA 75,8	61,8



INRAE

Extraction d'information textuelle
 13 sept 2023 / Ferment'IA / C. Nédellec

● Omnicrobe application

Dérozier S, et al., Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. *PLoS ONE* 2023.

https://omnicrobe.migale.inrae.fr

Omnicrobe Search Web services About

Search relations by habitat TSV Download Filter Selection

139 relations for the habitat "cheese"

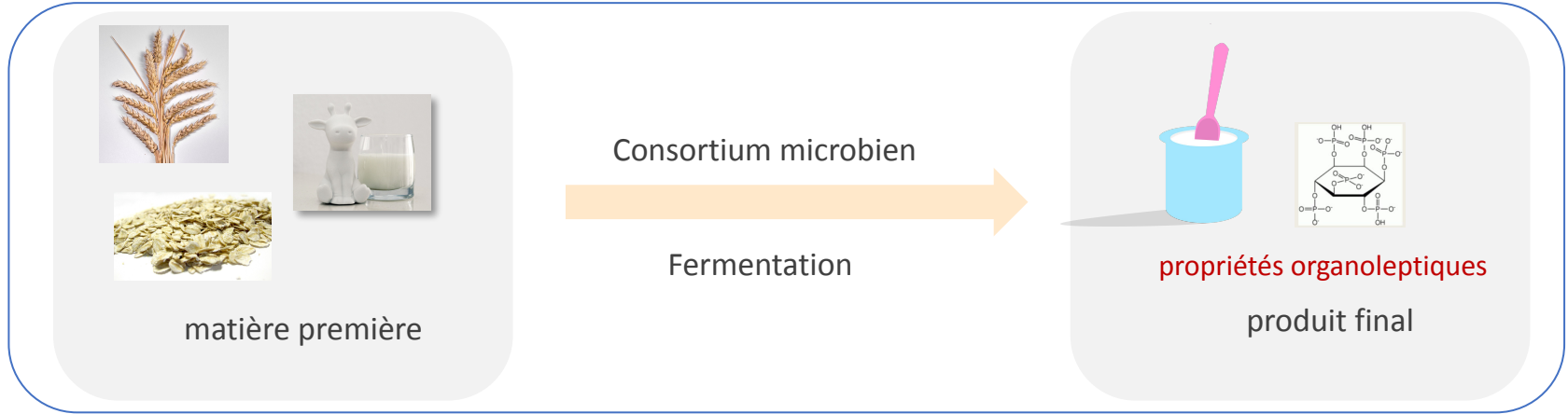
Source PubMed GenBank CIRM-BIA CIRM-Levures Taxon QPS only Apply

SOURCE TEXT	HABITAT	RELATION TYPE	TAXON	QPS	SOURCE
23035691	ripened cheese	may be inhabited by	Lactobacillus casei	✓	PubMed
27375244, 18538879	cheese	may be inhabited by	Lactobacillus casei group	✓	PubMed
7424219, 15975679, 30171516	cheese	may be inhabited by	Lactobacillus coryniformis	✓	PubMed
30625157	cheese	may be inhabited by	Lactobacillus crispatus	✓	PubMed
11573770	stretched curd cheese	may be inhabited by	Lactobacillus curvatus	✓	PubMed
11573770, 9812282	Cheddar	may be inhabited by	Lactobacillus curvatus	✓	PubMed
27423415	Habitat: Maroilles Appears in the text as: Maroille	may be inhabited by	Taxon: Lactobacillus curvatus Appears in the text as: L._curvatus	✓	PubMed
9812282, 29550118, 22916884		may be inhabited by		✓	PubMed
9353214	ripened cheese	may be inhabited by	Lactobacillus delbrueckii	✓	PubMed
21377750, 29319527, 16091941	cheese	may be inhabited by	Lactobacillus delbrueckii	✓	PubMed
19646036	feta	may be inhabited by	Lactobacillus delbrueckii	✓	PubMed
15778300	cottage cheese	may be inhabited by	Lactobacillus delbrueckii	✓	PubMed

13 sept 2023 / Ferment'IA / C. Nédellec

● Conception d'un *yaourt végétal*

Application d'Omnicrobe à l'innovation alimentaire



Expérimentation

1. Sélection de jus végétaux (avoine, riz, soja, ..)
2. **Sélection de bactéries disponible dans la collection CIRM BIA, capable de les fermenter**
3. Implémentation des souches sélectionnées dans une fermentation « réelle »
4. Mesure du niveau d'acidité et de la variation des composants organoleptiques

Etape 2 : Omnicrobe pour la sélection des bactéries

- **habitat:** vegetable juice (-> soy, rice, oat beverage)
- **usage:** acidification activity
- **phénotype :** mesophilic or thermophilic
- **Qualified Presumption of Safety :** yes
- **source:** collection INRAE CIRM BIA

INRAE

Extraction d'information textuelle
13 sept 2023 / Ferment'IA / C. Nédellec

● Sélection de souches microbiennes

[Harle et al. *Diversity of the metabolic profiles of a broad range of lactic acid bacteria in soy juice fermentation*. Food Microbiol. 2020.]



- concentration rapide sur un sous-ensemble pertinent d'espèces parmi les candidats potentiels.
- un gain dans la recherche bibliographique et l'expérimentation



7 espèces à tester pour la production de yaourt

201 souches à tester

Test d'acidification : fermentation du jus de soja dans un medium expérimental

6 / 8 *Lactobacillus acidophilus*
3 / 6 *Lactobacillus casei*
1 / 20 *Lactobacillus delbrueckii*
0 / 16 *Lactobacillus helveticus*
7 / 9 *Lactobacillus paracasei*
8 / 9 *Lactobacillus paraplantarum*
40 / 44 *Lactobacillus plantarum*
13 / 46 *Lactococcus lactis*
56 / 57 *Streptococcus thermophilus*

Acidifying strains / # strains

● De nombreuses questions scientifiques

Traitement automatique de la langue (TAL)

Relations longue distance

Désambiguïsation des entités

Normalisation par de grandes ontologies (ex. taxonomie)

De nombreuses informations à extraire

Métabolisme, molécules, leur production et dégradation

Consortia microbiens

Interface TAL/Représentation des connaissances

Entity-linking

Qualité des données

Inférence sur propriétés (ex. phénotype, biotope) pour la découverte de connaissance





Migale
Valentin Loux
Mouhamadou Ba



Bibliome
Louise Deléger
Claire Nédellec
Robert Bossy
Arnaud Ferré
Anfu Tang
Estelle Chaix
Philippe Bessières



INSTITUT **DATAIA**
Science des données, Intelligence & Société

Labex **DigiCosme**
université PARIS-SACLAY

université
PARIS-SACLAY



LISN
Pierre Zweigenbaum



StatInfoMics
Sandra Dérozier

MATHÉMATIQUES ET INFORMATIQUE APPLIQUÉES
MaiAGE
DU GÉNOME À L'ENVIRONNEMENT



SPO
Delphine Sicard



STLO et CIRM-BIA
Hélène Falentin
Florence Valence

INRAE



Secalim
Monique Zagorec



Micalis
Pierre Renault
Bedis Dridi
Maarten van de Guchte



INRAE

Extraction d'information textuelle
13 sept 2023 / Ferment'IA / C. Nédellec