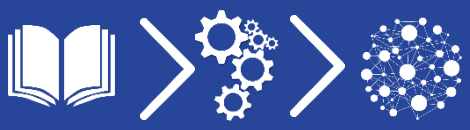
Application

Application pilote pour la recherche

Visa TM

Vers une infrastructure de services avancés de text mining

Vers une infrastructure de services avancés de text mining



2017 / 2019

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR,
DE LA RECHERCHE
ET DE L'INNOVATION

Application pilote pour la recherche

Exemple de l'écologie microbienne

I L'application démontre l'utilité de l'approche de composition de workflows pour un développement rapide et un résultat de qualité dans un domaine à fort enjeu scientifique et économique I

Description du Document

Application pilote pour la recherche

Lot	Application
Participants	MaIAGE (INRA)
Date de livraison	31/10/2019
Nature : Rapport	Version : 1.0

Contributeurs

	Nom	Organisation
Rédaction	Claire Nédellec Estelle Chaix Mouhamadou Ba Robert Bossy	MaIAGE (INRA)
Coordination	Claire Nédellec	MaIAGE (INRA)
Relecture	Louise Deléger Sandra Dérozier	MaIAGE (INRA)

SOMMAIRE

Avertissement

Ce document contient des descriptions des résultats du projet Visa TM. Certaines parties peuvent être soumises à des droits de propriété intellectuelle. Avant réutilisation du contenu, il est nécessaire de contacter le consortium pour approbation.

Acronymes et sigles

TDM	Text and Data Mining
OMTD	OpenMinTeD
IBF	Institut Français de Bioinformatique
CIRM	Centre de Ressources microbiennes
CNIEL	Centre national interprofessionnel de l'économie laitière
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH
STLO	Science et Technologie du Lait et de l'Œuf
SPO	Science pour l'Œnologie, Montpellier
GMPA	Génie et Microbiologie des Procédés Alimentaires
URTAL	Unité de Recherche Technologie et Analyses Laitières
SECALIM	SECurité des ALIments et Microbiologie
Micalis	MICrobiologie de l'ALImentation au service de la Santé
PESV	Plate-forme d'épidémiosurveillance en santé végétale
PMC	PubMed Central
NCBI	National Center for Biotechnology Information
GBIF	Global Biodiversity Information Facility
JGI	Joint Genome Institute
GOLD	Genomes Online Database
SRA	Sequence Read Archive
EFSA	European Food Safety Authority
ANSES	Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail German Federal Institute for Risk Assessment

Mirri	Microbial Resource Research Infrastructure
UMLS	Unified Medical Language System
QPS	Qualified presumption of safety
NER	Named Entity Recognition
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
URGI	Unité de Recherche Génomique Info
MGP	MetaGenoPolis
D2KAB	Data to Knowledge in Agronomy and Biodiversity

Résumé publiable

L'application démontre l'utilité de l'approche de composition de workflows de *text mining* et sa connexion à des applications métiers pour un développement rapide et un résultat de qualité dans un domaine à fort enjeu scientifique et économique, celui de la microbiologie. Le document analyse les points forts (qualité des productions) et limitations de l'approche (accès aux corpus). Il détaille également des perspectives de généralisation dans les domaines de l'agriculture et des sciences du vivant dans un contexte de science ouverte.

Introduction

Le volet application du projet Visa TM porte sur des réalisations concrètes qui illustrent la pertinence d'une infrastructure modulaire permettant le développement de la fouille de textes au profit de la communauté de recherche dans un cadre de science ouverte. Ce volet a pour but de démontrer la facilité de déploiement de nouveaux services de fouille de textes à base de composants de trois infrastructures (bibliothèques numériques, TDM et portails de ressources sémantiques) ainsi que la qualité et la pertinence des applications développées, dans deux domaines d'application scientifique.

Les applications pilotes sont porteuses d'enjeux scientifiques fondamentaux et appliqués, sociaux et économiques. Elles sont au nombre de deux : (1) la définition de corpus thématiques par une approche terminologique (2) la découverte de connaissance en écologie microbienne par extraction d'information présentée ici.

Ce document décrit la seconde application et son contexte scientifique. Dans cette application, les informations extraites automatiquement par l'application de fouille de textes complètent les informations expérimentales ou analytiques obtenues par d'autres approches. La disponibilité dans le domaine de la microbiologie de très nombreuses informations et données ouvertes, centralisées et représentées dans des standards interopérables facilite l'intégration des données textuelles et en augmente l'impact, justifiant le choix de ce domaine.

Nous avons défini le public visé et le scénario d'utilisation (section 2). La section Réalisation présente la démarche suivie, les corpus et ressources sémantiques exploitées et présente des exemples d'utilisation. La section Bilan met en évidence les principes généraux réapplicables à d'autres applications d'extraction d'information pour la recherche.

Analyse du besoin et spécification

1. Description du besoin

L'application *text mining pour la recherche* a pour but de démontrer par quels moyens les processus de fouille de textes peuvent être utiles à la recherche quand ils sont intégrés dans le processus de travail du chercheur. Le domaine de spécialité choisi est celui de la biodiversité microbienne. Le besoin traité ici est celui de la découverte et l'agrégation de connaissances décrites dans la littérature scientifique (documents scientifiques et champs de texte libre des bases de données), en particulier en microbiologie des aliments.

Plus précisément l'application facilite la découverte et l'agrégation des données relatives aux taxa microbiens dans les aliments et relatives à leurs propriétés, en particulier dans les aliments fermentés tels que les fromages et la viande.

Les microorganismes alimentaires participent à la construction des qualités sensorielles, nutritionnelles et sanitaires des aliments. Les informations visées par l'application doivent contribuer à enrichir la connaissance pour transformer, préserver et piloter des fonctionnalités avec une valeur ajoutée sociale et économique forte que ce soit pour des aliments d'importance culturelle ou économique (pain, vin, fromages, charcuterie,...). Il s'agit aussi de mieux contrôler les conditions d'altération des produits, de gérer les risques pour le consommateur et de limiter les pertes économiques.

Concrètement, par exemple :

- l'aide à l'identification de souches qui sont présentes dans les échantillons et formuler des hypothèses quant à leur provenance,
- l'aide à la sélection de souches pour de nouveaux produits alimentaires innovants (ex. fermentation, production de nutriments), ou de nouveaux moyens de conservation (biopréservation),
- l'aide à l'identification de souches ayant des propriétés positives (probiotiques) ou négatives sur la santé (pathogènes).

1.1 Utilisateurs

La communauté visée est celle des chercheurs en microbiologie et les organismes publics avec le soutien des infrastructures publiques de bioinformatique y compris les agrégateurs de données et les développeurs d'applications, et des entreprises agroalimentaires. Les utilisateurs finaux sont les scientifiques en microbiologie alimentaire.

Les scientifiques impliqués dans le cas d'utilisation sont demandeurs d'une application qui facilite la comparaison et l'exploitation unifiée des informations des micro-organismes et de leurs habitats et phénotypes à partir de diverses sources, littérature et bases de données, et grâce à une représentation formelle et unifiée.

.

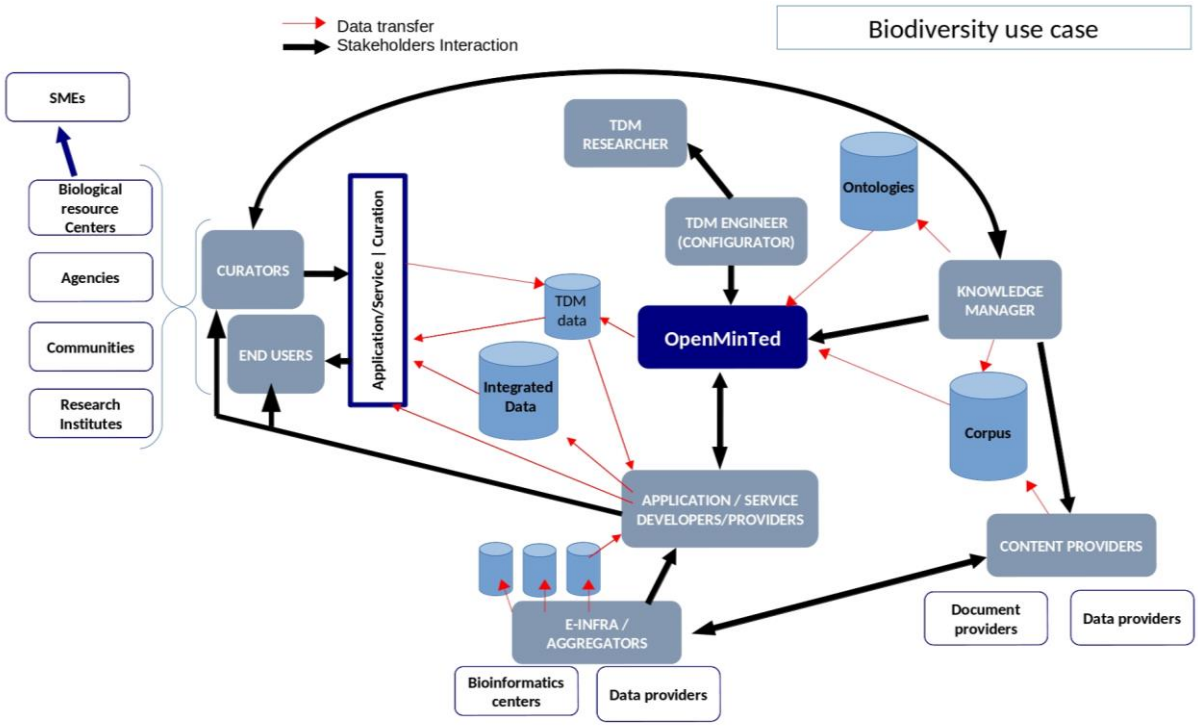


Figure 1. Carte des interactions entre les acteurs de l'application écologie microbienne.

L'analyse des besoins s'appuie sur plusieurs projets de recherche et industriels impliquant l'équipe Bibliome-MaIAGE de l'INRA, la plateforme bioinformatique Migale-INRA, de nombreux chercheurs microbiologistes de l'INRA, des représentants des Centre de Ressource Biologique (collections de souches microbiennes INRA CIRM et DSMZ) et de partenaires de l'INRA (CNIEL, filière de l'interprofession laitière) qui représentent différentes catégories d'acteurs (Figure 1).

Ces projets sont principalement,

- le projet H2020 OpenMinTeD [voir le document OpenMinTeD_D4.2: Community Requirements Analysis Report section Microbial Biodiversity],
- le projet CNIEL-INRA FoodMicroBiome Transfert[1]
- et les projets du Métaprogramme Méta-omiques et écosystèmes microbiens de l'INRA (GT Food) : OntoBiotope [Nédellec et al., 2017], Florilege [Falentin et al., 2017] et EnovFood [Falentin et al., 2019].

La typologie des profils concernés se compose de :

- chercheurs et ingénieurs en fouille de textes
- scientifiques microbiologistes
- plateformes bioinformatiques
- agrégateurs de contenus

[1] http://maiage.jouy.inra.fr/?q=fr/node/848

Chercheurs et ingénieurs en fouille de textes

Ce sont les membres de l'équipe Bibliome-MaIAGE de l'INRA dans notre application. De manière générale, les chercheurs et ingénieurs en fouille de textes sont impliqués dans la valorisation et le transfert d'outils de fouille de textes innovants leur adaptation et dans la combinaison de ces outils pour la réalisation du *workflow* de fouille de textes. Ils contribuent à l'automatisation de la constitution de corpus. Ils réalisent les tests d'outils et du workflow, ils contribuent à l'intégration du workflow de fouille de textes dans l'application cliente. L'application visée nécessite des outils d'analyse sémantique très avancés pour identifier les entités des textes, les catégoriser et les relier pour produire une base de connaissance.

Scientifiques microbiologistes (*end-users* dans le schéma de la Figure 1)

Ces scientifiques ont pour objectif de mieux connaître et contrôler la présence et les interactions de microorganismes dans les aliments en faisant le lien entre génotype, phénotype et environnement. Ce sont des laboratoires de recherche fondamentale et appliquée, ici, les laboratoires STLO, SPO, GMAP, TAL, URTAL, LRF, CIRM BIA, SECALIM, Micalis des projets sus-mentionnés. Ils couvrent les domaines des écosystèmes fromagers, des écosystèmes viande et produits marins, écosystèmes en œnologie et boulangerie, des interactions matrices laitières.

Ce sont aussi des industriels et agences dans un objectif de sécurité sanitaire, des industriels dans un objectif d'innovation (biopréservation, probiotiques, etc.), ici représenté par le CNIEL (*Centre national interprofessionnel de l'économie laitière*).

Plateformes de bioinformatique (*Bioinformatics centers* dans le schéma), ici Migale-INRA.

Elles offrent un service générique et public pour les projets de bioinformatique, par exemple, de métagénomique, du dépôt des analyses des échantillons (séquences), jusqu'à l'analyse de leur écosystème. Elles créent et maintiennent des systèmes d'information ouverts en réseau. Ces bioinformaticiens sont des développeurs d'application et de services sur des infrastructures métiers. Ils sont informaticiens non spécialistes de la fouille de textes et ils développent des d'applications qui utilisent des fonctions de fouille de textes. L'application pilote contribue à la définition d'un cadre de travail qui leur permet facilement de trouver les fonctions de fouille de textes qui leurs sont utiles pour les réutiliser, les adapter et les combiner. Ils sont au cœur du déploiement et de l'exploitation de cette application.

Agrégateurs (*e-infra aggregators* dans le schéma) et *Content Providers*

Ils souhaitent valoriser des contenus en particulier *Open Access*, des publications (PMC, Istex, OpenAire ici), et des ressources sémantiques (AgroPortal ici). Les agrégateurs sont également consommateurs de services de fouille de textes pour la curation de l'information produites par les projets de bioinformatique (ex. Migale, GOLD-JGI) ou des centres de ressources biologiques (ex. DSMZ, CIRM BIA).

1.2 Service attendu par les microbiologistes

Les méthodes de fouille de textes réduisent le temps d'analyse et d'interprétation grâce à la synthèse qu'elles permettent de faire en amont sur l'ensemble des connaissances publiées sur les habitats et phénotypes microbiens.

Les informations pertinentes extraites des textes portent sur les relations entre les micro-organismes, leurs habitats et certaines propriétés des micro-organismes, les phénotypes. Une fois normalisées à l'aide de bases de connaissances formelles (ontologies, taxonomie), ces données sont intégrables avec celles provenant d'autres sources (par exemple, les données expérimentales, les données scientifiques), en vue d'une utilisation par les chercheurs dans ces domaines.

Le service attendu doit intégrer toutes les sources pertinentes (publications, base de données) sans restriction. L'interface d'accès à l'information doit permettre une interrogation intuitive et efficace et un export des résultats dans un format standard.

L'application Florilège développée dans le cadre du projet OpenMinTeD répond en grande partie à ces besoins. Elle est décrite dans les livrables publics[2] *OpenMinTeD_D4.2_Community Requirements Analysis Report Microbial Biodiversity*, *OpenMinTeD_D9.2 Community-Driven Applications Design Report* et *OpenMinTeD-D9.4 Application Software Release*. Nous en donnons ici un résumé et précisons les éléments complémentaires apportés par le projet Visa TM.

1.3 Spécification du corpus et données

Les entités et relations à extraire

L'application réalisée extrait les mentions textuelles de quatre types d'entités, les microorganismes, leurs habitats, leurs phénotypes et les lieux géographiques.

Microorganismes

Les entités microorganismes sont des empans contigus de texte qui contiennent un nom de taxon sans ambiguïté, à tous les niveaux taxonomiques, du phylum à la souche. Ils sont associés à la catégorie la plus spécifique et unique de la ressource taxonomique du NCBI. Si une souche ou un groupe de souches n'est pas référencé par le NCBI, il est attribué au taxid le plus proche dans la taxonomie. La figure 2 en donne un exemple.

Streptococcus salivarius is the principal commensal bacterium of the oral cavity in healthy humans.

Figure 2. Exemple d'entités Microorganisme et Habitat

Habitat

Les entités habitat sont des empans possiblement non contigus de texte qui contiennent une mention complète d'un habitat potentiel pour les microorganismes. Les entités d'habitat se voient attribuer un ou plusieurs concepts de la sous-partie habitat de l'ontologie OntoBiotope

[2] http://openminted.eu/deliverables/

de référence. Les concepts assignés sont aussi spécifiques que possible. OntoBiotope définit les habitats de micro-organismes les plus pertinents dans tous les domaines considérés par l'écologie microbienne (hôtes, environnement naturel, environnements anthropisés, nourriture, médical, etc.). La figure 2 en donne un exemple. Les entités d'habitat sont rarement des entités figées (noms propres), ce sont généralement des expressions adjectivales et nominales incluant des propriétés et des modificateurs. Il y a de rares cas d'habitats référencés par des adjectifs ou des verbes. Les empans de texte sont généralement contigus mais certains sont discontinus par exemple pour les conjonctions.

Phénotype

Les entités Phénotype sont des empans possiblement non contigus de texte qui contiennent une mention complète d'un phénotype de microorganisme. Les entités phénotype se voient attribuer un ou plusieurs concepts de la sous-partie phénotype de l'ontologie OntoBiotope de référence. Les concepts assignés sont aussi spécifiques que possible. OntoBiotope définit les phénotypes de micro-organismes de tous types (morphologique, source d'énergie, relation avec l'hôte, *etc*.). Les entités Phénotype sont rarement des entités référentielles, ce sont généralement des expressions nominales incluant des propriétés et des modificateurs. Les empans de texte sont généralement contigus mais certains sont discontinus par exemple pour les conjonctions. La figure 3 en donne un exemple.

[…] mesophilic heterofermentative Lactobacilli […]

Figure 3.Exemple d'entités Microorganisme et Phénotype

Géographique

Les entités géographiques sont des lieux géographiques et d'organisation désignés par des noms officiels. La figure 4 en donne un exemple.

[…] sheep and goats in Europe […]

Figure 4. Exemple d'entités Géographique

L'application réalisée extrait également des relations entre ces entités, la relation *Lives in* et la relation *Exhibits*.

Relation Lives_In

La relation Lives_in a deux arguments, le microorganisme et le lieu où il vit (soit un Habitat, soit une entité géographique). Si un habitat est une partie d'hôte et que l'hôte est mentionné dans le texte, les deux habitats participent à deux relations distinctes. La figure 5 en donne un exemple.

Relation Exhibits

La relation Exhibits relie des entités de type Microorganism aux entités Phenotype. La figure 5 en donne un exemple.

La figure 5 donne un exemple des entités, de leur normalisation et des deux types de relation.

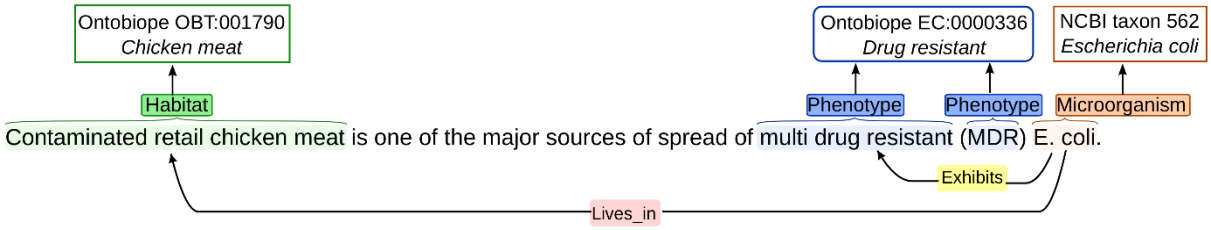


Figure 5. Exemple d'entités, de relations et de normalisation.

Le document Bacteria Biotope Annotation Guidelines[3] [Bossy et al., 2019] décrit très précisément les règles d'annotation manuelle et par là-même, d'annotation automatique des informations à extraire du texte.

Les sources de données

Les données produites par l'application de fouille de textes sont intégrées dans l'application cliente Florilège. A terme, l'ambition de l'application Florilège est d'intégrer des informations de sources diverses. La figure 6 présente le schéma général de l'objectif de l'application. La figure 7 détaille les sources.

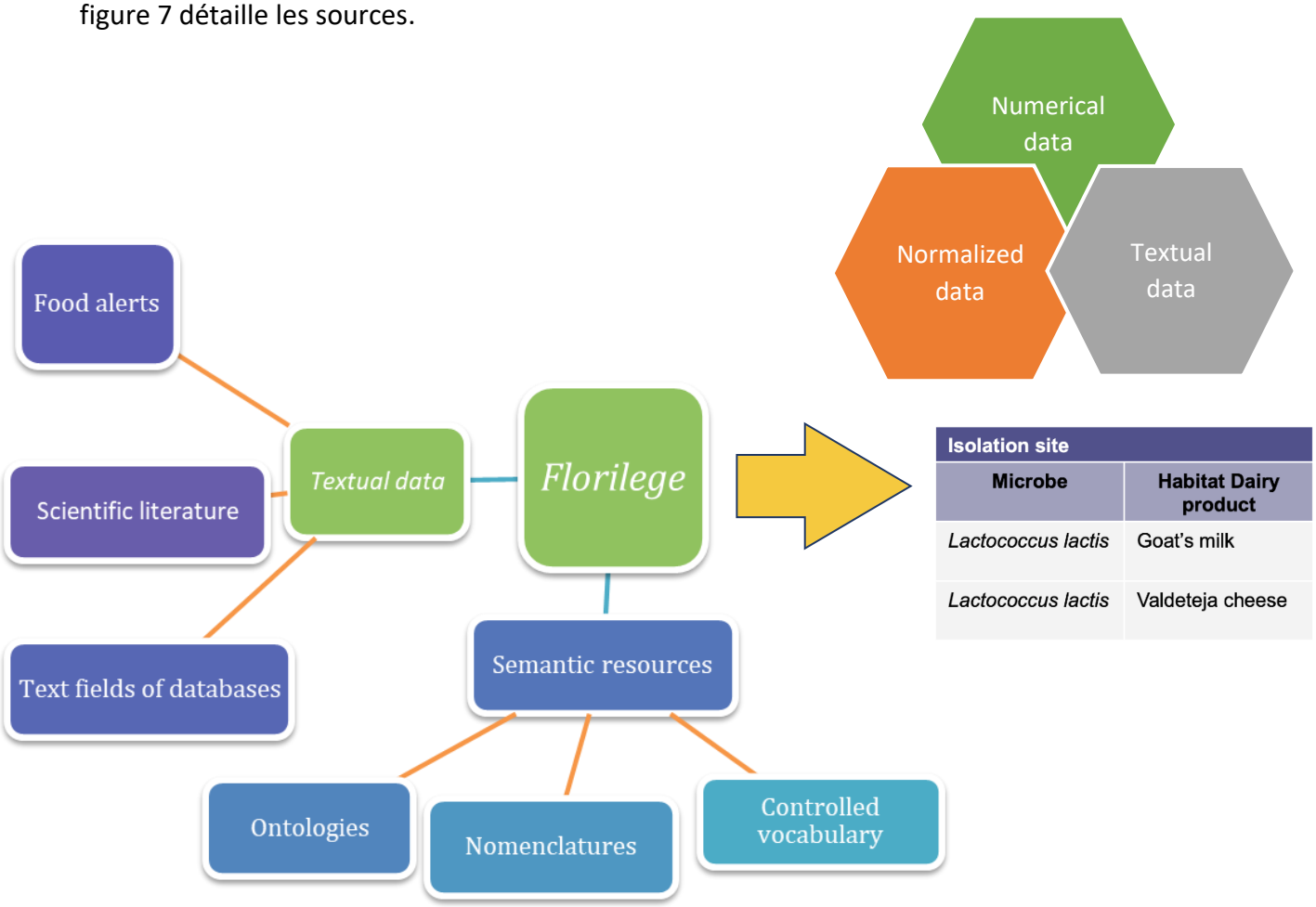


Figure 6. Schéma général de l'intégration de données.

[3] https://drive.google.com/file/d/1G0po_xlRjQCZ-qxuA_4PLdipXU6rtYTp/view

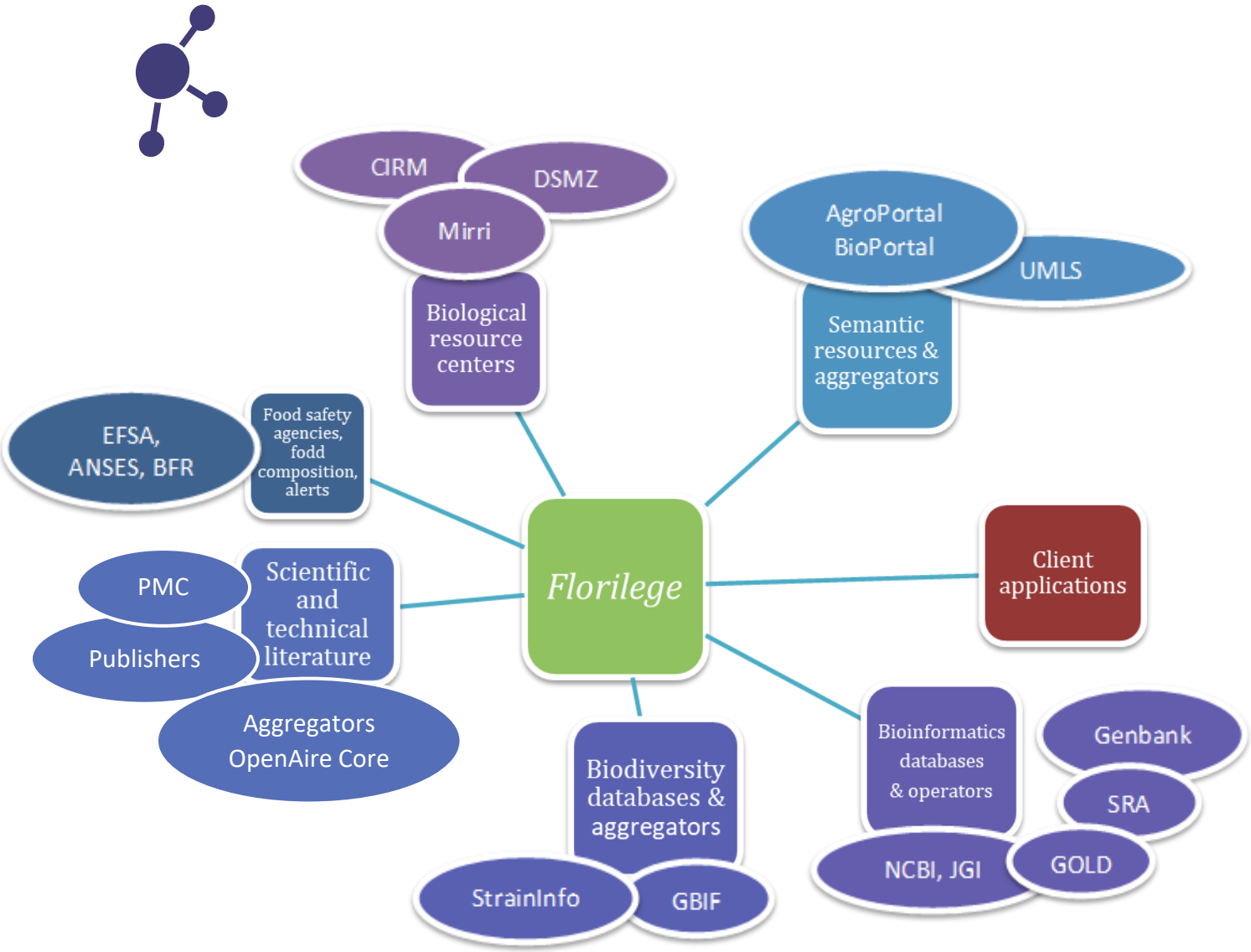


Figure 7. Schéma des sources de données pertinentes de la base Florilège.

Les sources de données sont de plusieurs natures.

Ressources sémantiques

Les ressources sémantiques de l'application sont utilisées pour guider l'extraction d'information et normaliser les entités, c'est-à-dire leur assigner une catégorie de référence. Ce sont :

> la taxinomie des espèces du NCBI[4] pour normaliser les taxa par leur ID. Les taxa des phylla de microorganismes sont listés dans la table 1.
> l'ontologie OntoBiotope[5] pour normaliser habitats et phénotypes [Nédellec et al, 2018].

[4] https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
[5] http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE

> Microbe	ID MeSH term	Food domain	ID domain
Alveolata	B01.043	Diet, Food and Nutrition	G07.203
Amoebozoa	B01.046	Food Analysis	E05.362
Nematoda	B01.050.500.500.294		J01.576.423.850.100
Choanoflagellata	B01.175	Food and Beverages	J02
Cryptophyta	B01.206	Food Industry	J01.576.423
Diplomonadida	B01.237	Food Microbiology	H01.158.273.540.274.332
Euglenozoa	B01.268		N06.850.601.500.249.300
Fungi	B01.300		N06.850.425.200
Haptophyta	B01.400		N06.850.460.400.300
Mesomycetozoea	B01.500	Food Packaging	J01.576.423.200.375
Oxymonadida	B01.625		J01.576.423.850.600
Parabasalidea	B01.630		J01.576.761.400
Glaucophyta	B01.650.232	Food Quality	J01.576.423.850.730
Chlorella	B01.650.940.150.469		N06.850.601
Prototheca	B01.650.940.150.634		
Volvocida	B01.650.940.150.925		
Volvox	B01.650.940.150.950		
Desmidiales	B01.650.940.800.150.200		
Retortamonadidae	B01.675		
Rhizaria	B01.680		
Stramenopiles	B01.750		

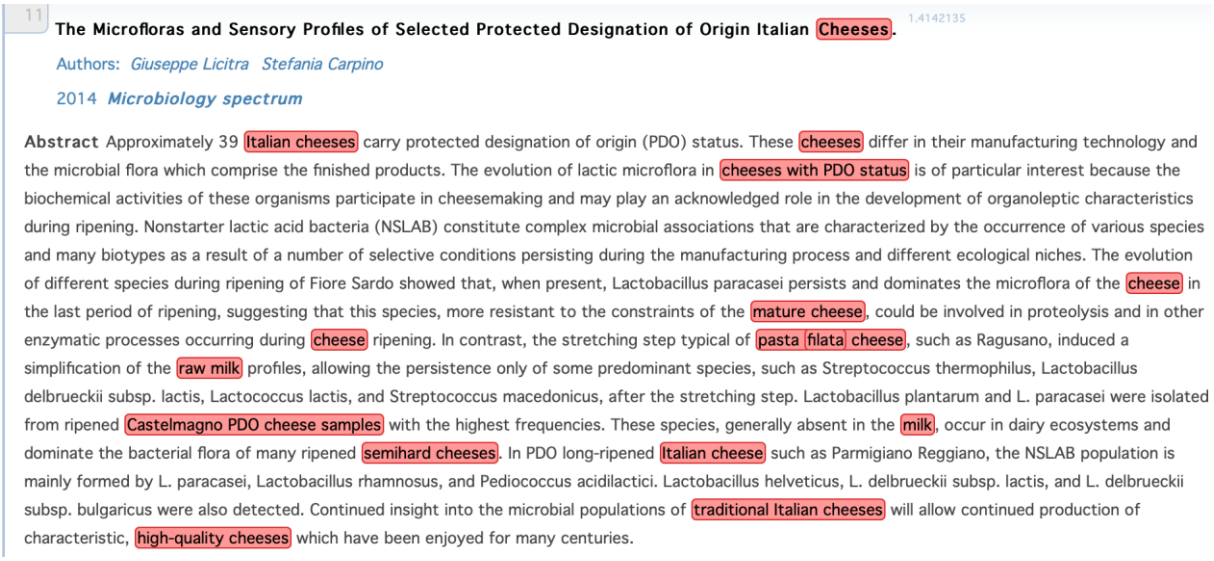
Crenarchaeota	B02.075
Euryarchaeota	B02.200
Korarchaeota	B02.500
Nanoarchaeota	B02.600
Bacteria	B03
Viruses	B04

Tableau 1. Critère de sélection des entrées PubMed

Documents

Le corpus de documents scientifiques : dans la version courante il s'agit de **références PubMed** sélectionnées par les taxa microbiens sous forme de mots-clefs MeSH dans l'interface de requête PubMed, à laquelle sont ajoutées les restrictions sur les articles de journaux en langue anglaise : "AND (lang:eng) AND (type:D016428*) ". Le détail est décrit dans [Chaix et al., 2018].

Le choix de PubMed comme source documentaire est justifié par sa très large couverture thématique en microbiologie, comparée aux autres agrégateurs comme WoS (Web of Science). Les microbiologistes souhaitent également accéder aux informations des articles complets que l'application de fouille de textes peut analyser. Les grandes difficultés rencontrées d'ordre technique et légales pour accéder à et analyser l'ensemble des documents nécessaires limitent la source de l'application de septembre 2019 aux références PubMed. La figure 8 donne un exemple de telle référence.



11 1.4142135

The Microfloras and Sensory Profiles of Selected Protected Designation of Origin Italian Cheeses.

Authors: *Giuseppe Licitra Stefania Carpino*

2014 ***Microbiology spectrum***

Abstract Approximately 39 Italian cheeses carry protected designation of origin (PDO) status. These cheeses differ in their manufacturing technology and the microbial flora which comprise the finished products. The evolution of lactic microflora in cheeses with PDO status is of particular interest because the biochemical activities of these organisms participate in cheesemaking and may play an acknowledged role in the development of organoleptic characteristics during ripening. Nonstarter lactic acid bacteria (NSLAB) constitute complex microbial associations that are characterized by the occurrence of various species and many biotypes as a result of a number of selective conditions persisting during the manufacturing process and different ecological niches. The evolution of different species during ripening of Fiore Sardo showed that, when present, Lactobacillus paracasei persists and dominates the microflora of the cheese in the last period of ripening, suggesting that this species, more resistant to the constraints of the mature cheese, could be involved in proteolysis and in other enzymatic processes occurring during cheese ripening. In contrast, the stretching step typical of pasta filata cheese, such as Ragusano, induced a simplification of the raw milk profiles, allowing the persistence only of some predominant species, such as Streptococcus thermophilus, Lactobacillus delbrueckii subsp. lactis, Lactococcus lactis, and Streptococcus macedonicus, after the stretching step. Lactobacillus plantarum and L. paracasei were isolated from ripened Castelmagno PDO cheese samples with the highest frequencies. These species, generally absent in the milk, occur in dairy ecosystems and dominate the bacterial flora of many ripened semihard cheeses. In PDO long-ripened Italian cheese such as Parmigiano Reggiano, the NSLAB population is mainly formed by L. paracasei, Lactobacillus rhamnosus, and Pediococcus acidilactici. Lactobacillus helveticus, L. delbrueckii subsp. lactis, and L. delbrueckii subsp. bulgaricus were also detected. Continued insight into the microbial populations of traditional Italian cheeses will allow continued production of characteristic, high-quality cheeses which have been enjoyed for many centuries.

Figure 8. Entrée de la base PubMed. Les habitats sont surlignés.

De nombreuses **bases de données spécialisées** contiennent des champs textuels décrivant des habitats et phénotypes microbiens. Dans l'application de septembre 2019, trois sources principales ont été intégrées. Le document *OpenMinTeD-D9.4 Application Software Release Final* détaille les sources et licences.

> le champ commentaire de **GenBank** FEATURES Location/Qualifiers isolation_source des RNA 16S de plus de 800 paires de base. La taille du RNA 16S est un élément de fiabilité de l'identification du taxon. La figure 9 donne un exemple des "isolation sources" pour *Halomonas sp*. Elle illustre la diversité des descriptions des milieux. Les étiquettes sont des exemples des catégories associées automatiquement par l'application de fouille de textes.

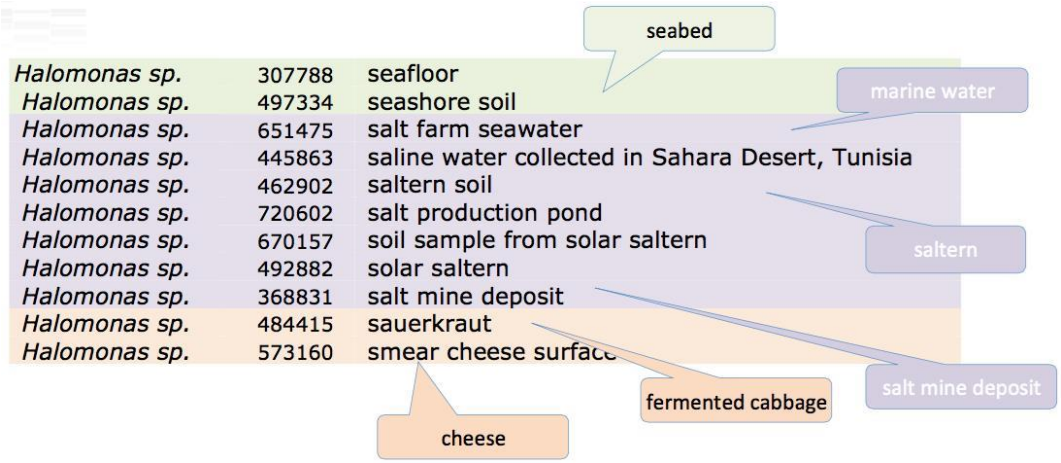


Figure 9. Entrée de la base de données GenBank

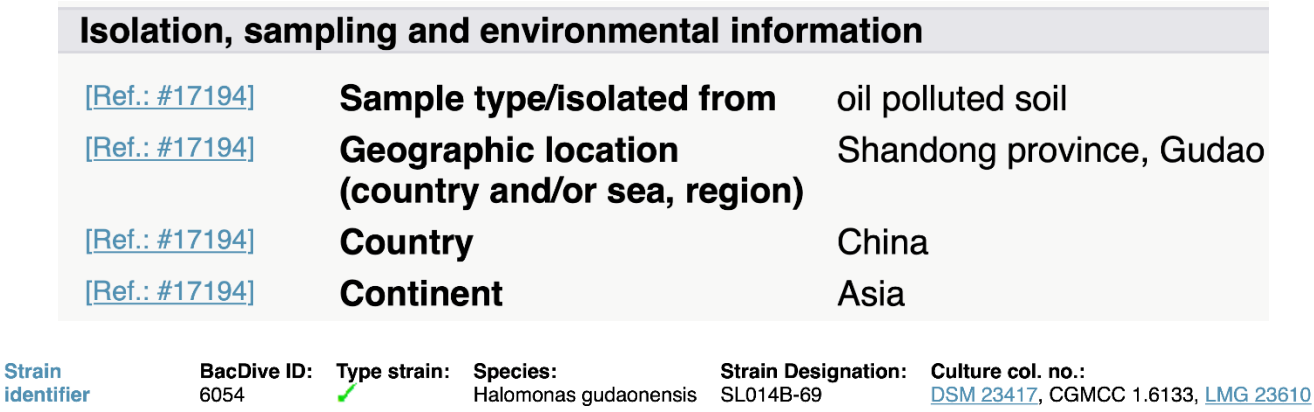
> Le champ Source des **CIRM**, Centre International de Ressources Microbiennes créé en 2004 par l'INRA autour de ses collections de micro-organismes (bactéries, levures, champignons filamenteux). Le CIRM conserve plus de 15 000 souches de bactéries associées aux plantes, bactéries pathogènes, bactéries d'intérêt alimentaire, de levures et de champignons filamenteux. Dans la version de septembre 2019, Florilège exploite 2385 entrées. La table 2 représente un extrait du CIRM levures.

Espèce	Source	Environnement
Saccharomycopsis capsularis	pasture soil	Environment
Saccharomycopsis fibuligera	chalky bread	Food
Saccharomycopsis crataegensis	fallen hips of hawthorne (Crataegus sp.)	Plant
Saccharomycopsis crataegensis	fallen hips of hawthorne (Crataegus sp.)	Plant
Saccharomycopsis vini	grape must	Oenology
Saccharomycopsis vini	grape must	Oenology

Saccharomycopsis malanga	ragi-tapi, fermented-food starter	Food
Saccharomycopsis capsularis	pollen on bee (Xylocopa caffra)	Plant
Saccharomycopsis vini	grape must	Oenology
Saccharomycopsis fibuligera	chalky bread	Food

Tableau 2. Exemple d'entrées de la base de données CIRM Levure

> Le champ Sample type/isolated de **BacDive**, the *Bacterial Diversity Metadatabase* fournit des informations sur la biodiversité des bactéries et archae. Il dépend de DSMZ de l'Institut Leibniz DSMZ - la collection nationale allemande de Microorganismes et de culture de Cellules (*Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH*) (en abrégé DSM et DSMZ) fondé en 1969. La version de septembre 2019 contient 19 913 entrées. La Figure 10 en donne un exemple.



Isolation, sampling and environmental information

[Ref.: #17194]	**Sample type/isolated from**	oil polluted soil
[Ref.: #17194]	**Geographic location (country and/or sea, region)**	Shandong province, Gudao
[Ref.: #17194]	**Country**	China
[Ref.: #17194]	**Continent**	Asia

Strain identifier	BacDive ID:	Type strain:	Species:	Strain Designation:	Culture col. no.:
	6054	✓	Halomonas gudaonensis	SL014B-69	DSM 23417, CGMCC 1.6133, LMG 23610

Figure 10. Entrée de la base de données BacDive de DSMZ.

Le tableau 3 synthétise la description des différents corpus de documents. La dernière ligne représente les corpus annotés manuellement et utilisés pour l'entraînement et l'évaluation de l'extraction automatique d'information.

Nom	Spécification du corpus	Utilisation	Licence
Microbial Biodiversity corpus	Entrées Pubmed sur les microbes [Chaix et al. 2018]	Extraction de l'information pertinente.	licence PMC
GenBank corpus	texte libre du champ *Isolation* field de GenBank, séquences RNA16S (taille > 800bp)	Extraction de l'information pertinente.	licence Publique NCBI
CIRM corpus	Entrées libre du champ *Source*	Extraction de l'information pertinente.	CC-BY 3.0
BacDive corpus		Extraction de l'information pertinente.	licence DSMZ
Bacteria Biotope corpus	BioNLP-ST'11[6] [Bossy et al., 2012], BioNLP-ST'13[7] [Bossy et al, 2015], BioNLP-ST'16[8] [Deléger et al., 2016], BioNLP-OST'19 [Bossy et al, 2019][9]	Évaluation des résultats automatiques	license CC-BY-SA v3.0. (INRA)

Tableau 3. Corpus documentaire de l'application Florilège.

[6] http://2011.bionlp-st.org/home/bacteria-biotopes
[7] http://2013.bionlp-st.org/tasks/bacteria-biotopes
[8] http://2016.bionlp-st.org/tasks/bb2
[9] https://sites.google.com/view/bb-2019

Architecture

L'application est composée de plusieurs éléments comme le montre la Figure 11. Le workflow de fouille de textes est déployé dans OpenMinTeD pour extraire les informations à partir de textes. Les applications clientes AlvisIR et Florilège déployées sur la plateforme IFB INRA Migale permettent aux utilisateurs d'accéder facilement aux informations produites par le workflow de fouille de textes.

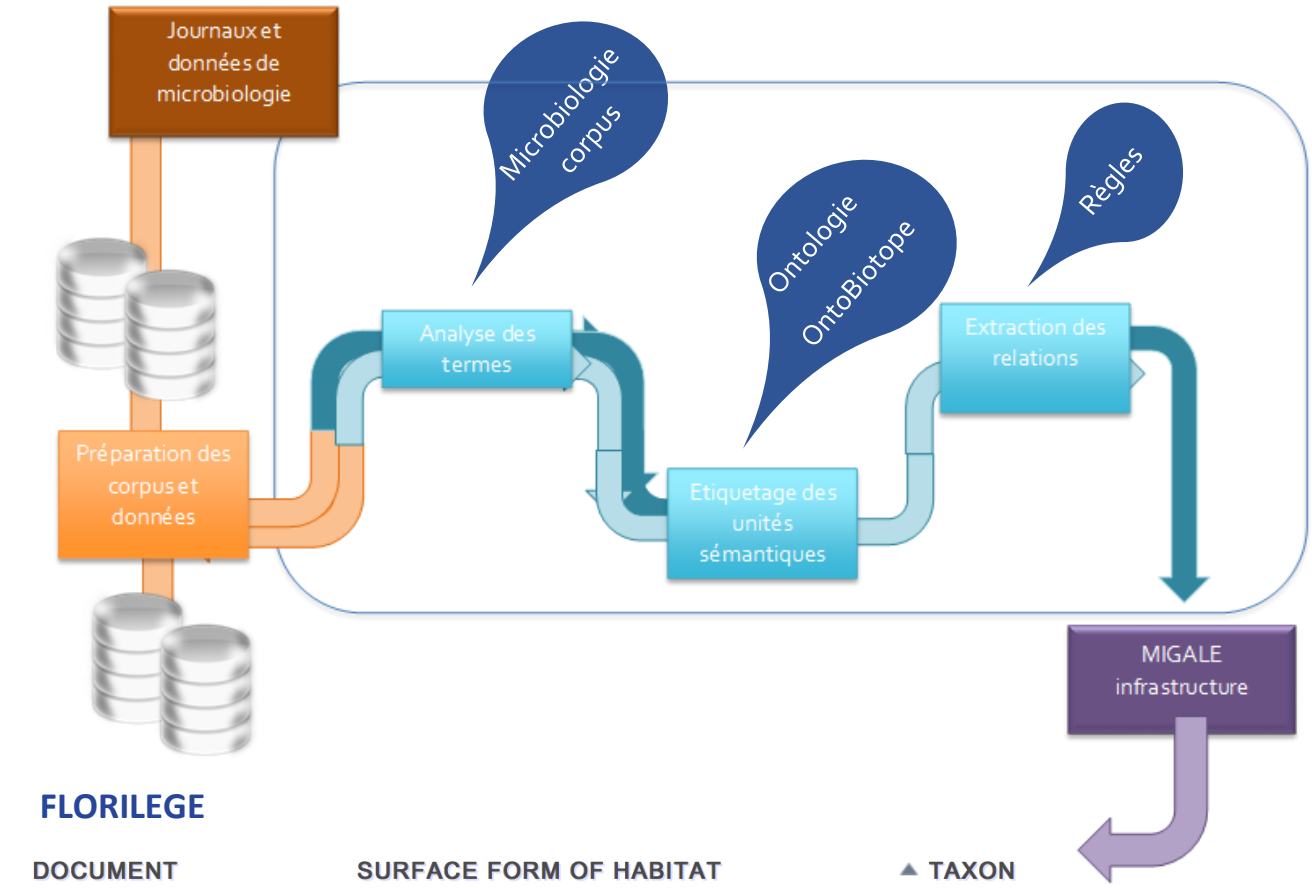
Le workflow de fouille de textes réalise plusieurs étapes d'extraction d'information:

- Il récolte et convertit préalablement en texte, les articles scientifiques et champs de texte libre des bases de données. La sélection des documents pertinents est réalisée à l'aide de critères superficiels (listes de mots clés) décrits ci-dessus.
- Par un processus de traitement automatique de la langue il prétraite les documents par tokénisation, segmentation en phrases, lemmatisation, étiquetage morphosyntaxique et éventuellement, analyse syntaxique.
- Les outils de reconnaissance et de catégorisation des entités du workflow permettent de détecter les entités pertinentes (les taxa, habitats, phénotypes et lieux géographiques) dans les documents et de les mettre en correspondance avec des catégories, des concepts des ressources sémantiques formelles (taxonomie et ontologie). Le workflow utilise une combinaison de méthodes en TAL, adaptées au domaine grâce aux ressources sémantiques et par apprentissage automatique. Les méthodes sont entraînées sur un corpus de référence annoté manuellement par des experts en microbiologie alimentaire. Dans une étape suivante, les relations entre les entités identifiées sont extraites grâce à des méthodes combinant TAL et apprentissage automatique.

Les mécanismes et interfaces permettant de gérer le cycle de vie du workflow (créer, modifier, exécuter, etc…) sont offerts par la plateforme OpenMinTeD. Elle contient notamment une bibliothèque de modules de traitement et autres ressources impliquées dans la fouille de textes. Elle est également dotée d'interfaces adaptées pour assurer l'accès et la gestion de l'ensemble des ressources par les utilisateurs.

Les résultats du workflow de fouille de textes sont exploités par les applications clientes. Au-delà des informations extraites, le workflow de traitement partage deux types d'informations avec ces applications clientes : les références de la base de connaissances (taxons, habitat et phénotypes) et les métadonnées pour l'affichage d'informations textuelles (références bibliographiques). Elles proviennent de ressources externes et nécessitent la synchronisation des mises à jour entre le workflow de fouille de textes et les applications clientes. Les références de la base de connaissances servent à indexer et formuler des requêtes. Les métadonnées bibliographiques servent à l'affichage du texte primaire.

Ces deux applications clientes Florilege et AlvisIR prennent en charge deux types d'interactions différents correspondants à deux usages différents. Le premier sous forme de base de données et de listes exportables, le second sous forme de moteur de recherche bibliographique avec des interfaces de recherche et de visualisation avancées.



FLORILEGE

DOCUMENT	SURFACE FORM OF HABITAT	▲ TAXON
PMID: 3972448	dairy products	Clostridium beijerinckii
PMID: 11265191, 10419207, 10528719	bakery, high moisture bakery products, high moisture-high pH bakery products	Clostridium botulinum

AlvisIR

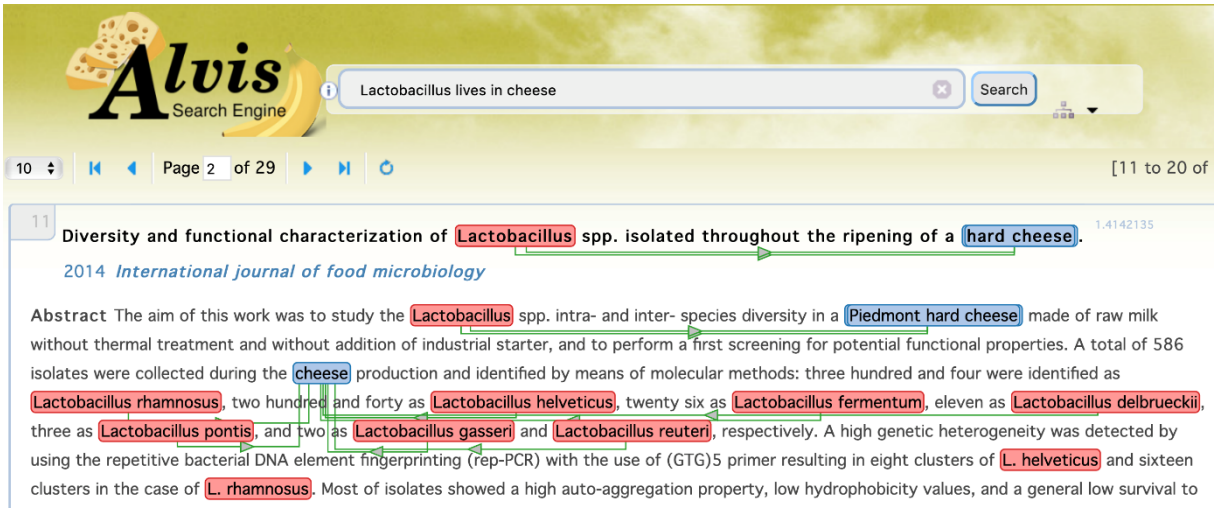


Figure 11. Schéma général des étapes d'extraction d'information

Réalisation

3.1 Corpus et données

Les données traitées en septembre 2019 sont décrites dans le tableau 3.

Source	Nombre de textes
Références PubMed	2 333 943
GenBank	65 536
BacDive	24 150
CIRM	2383

Tableau 4. Nombre de documents par source.

	Taxinomie du NCBI	OntoBiotope	Lexique d'expressions latine
Version	Non Versionnée	La version courante publique est The current public version is BioNLP-OST'19	V0.5
Lien vers la ressource	ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz	http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE	ressource incluse dans AlvisNLP
Lien vers la documentation	N/A	http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE	N/A
Utilisation	Reconnaissance et normalisation des taxa	Reconnaissance et normalisation des habitats et phénotypes	prétraitement linguistique (segmentation en mot, étiquetage morpho syntaxique
Format des données	Format propriétaire	obo	Texte
Taille	36Mb compressé, ~1.4M taxa, 2M noms	3,601 classes	37KB; 1889 entrées
Licence	UMLS Metathesaurus License: http://uts.nlm.nih.gov/licence.html	CC-BY-SA license v3.0	Apache License v2

Tableau 5. Licences et utilisation des ressources sémantiques.

Le tableau 4 décrit les ressources sémantiques utilisées pour extraire et catégoriser les entités des textes, puis pour indexer et requêter les données de la base de connaissance Florilège.

Le tableau 5 donne le nombre d'entités et de relations différentes extraites automatiquement par l'application de fouille de textes et indexées dans l'application cliente Florilège.

Source	Total Taxa	Taxa Uniques	Total Habitats	Habitats Uniques	Total Phénotype	Phénotypes Uniques	Relation Exhibits	Relation Lives in
Références PubMed	8 581 967	6 3367	26 804 948	2 592	3 372 543	256	48 458	588 752
GenBank	2 214	167	1 853	119	N/A	N/A	N/A	64 584
BacDive	144 008	50 054	135 060	911	N/A	N/A	N/A	19 913
CIRM	61 400	12 349	28 117	990	N/A	N/A	N/A	620

Tableau 6. Données de la base Florilège

La base Florilège indique également le Statut QPS pour 13468 taxons pour les relations taxon - habitat et 780 taxons pour les relations taxon - phénotype. Le statut QPS est une donnée externe. Le statut QPS ou Présomption d'innocuité reconnue est accordé par l'EFSA à un microorganisme qui répond aux critères suivants :

- son identité taxonomique doit être clairement définie
- le corpus de connaissances disponibles doit être suffisant pour pouvoir établir sa sécurité
- l'absence de propriétés pathogènes doit être établie et justifiée
- son utilisation prévue doit être clairement décrite
- Il est particulièrement utile pour la conception de produits alimentaires contenant des microorganismes.

3.2 Traitements

L'application pilote en microbiologie est composée d'outils génériques, réutilisés dans de nombreuses autres applications dans le traitement de documents de langue anglaise. L'adaptation au domaine est réalisée au moyen de l'exploitation des ressources sémantiques externes et d'une ressource interne : le lexique de l'outil ToMap composé des têtes non sémantiques dans le domaine (ex. *Sample* dans *intestinal sample*) [Golik et al., 2011].

Les composants interopérables sont connectés sous forme d'un workflow AlvisNLP, encapsulés dans un container Docker et déployés sur la plateforme OpenMinteD selon la spécification de la documentation d'OpenMinteD[10]. Le composant Docker est entièrement fonctionnel en tant qu'image docker autonome et répond aux exigences de la plateforme. Il a

[10] https://guidelines.openminted.eu

été enregistré dans le registre OMTD avec les métadonnées pertinentes, y compris les conditions de licence pour l'utilisation du composant. Il a été mis à jour afin de traiter le format XMI et peut prendre en entrée les documents XMI résultant de la conversion des corpus PDF créés avec le corpus builder d'OMTD au format XMI. Les ressources nécessaires ont été mises à disposition dans l'infrastructure OMTD ou encapsulées dans l'image du Docker. Une application a été créée à partir du composant Docker sur la plate-forme OMTD et l'exécution a été testée avec succès.

3.3 Outils et Composants

Sauf indication contraire, tous les composants de l'application de fouille de textes proviennent de la bibliothèque de composants d'AlvisNLP. Ils partagent tous les mêmes caractéristiques opérationnelles :

- Version : 0.5.
- Licence : Licence Apache v2.
- Format d'entrée et de sortie : représentation interne d'AlvisNLP, sauf indication contraire. Les composants disponibles sur OpenMinTeD acceptent et produisent du XMI format d'échange d'OMTD en utilisant le système de type proxy AlvisNLP.
- Mode de déploiement : Module AlvisNLP.
- Exigences relatives à l'environnement système : AlvisNLP[11]
- Exigences et limites de performance : RAM pour la taille du corpus.

L'INRA ajoute continuellement des composants AlvisNLP ou des assemblages de composants AlvisNLP sur OpenMinTeD. La liste des composants actuellement disponibles est tenue à jour[12].

La figure 12 présente l'architecture générale, décomposée dans les figures suivantes.

[11] https://bibliome.github.io/alvisnlp/

[12] https://github.com/openminted/alvis-docker/tree/master/openminted-components

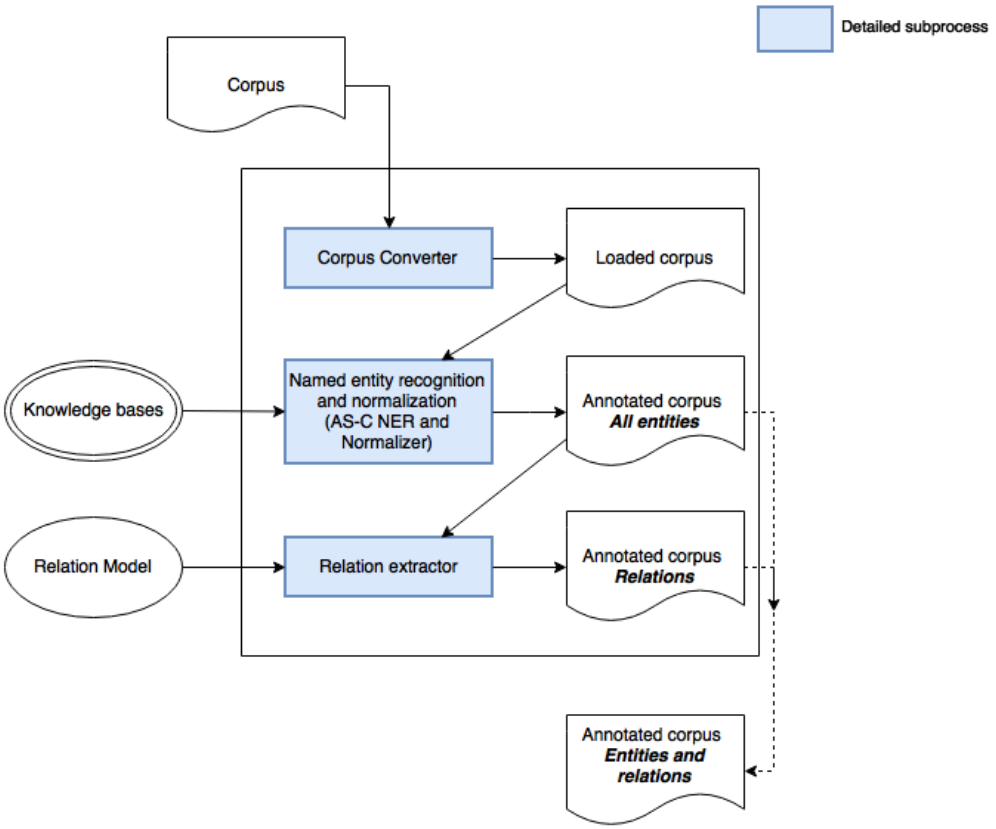


Figure 12. Architecture générale de l'application fouille de textes

La figure 13 détaille l'architecture de la reconnaissance des entités et de leur normalisation.

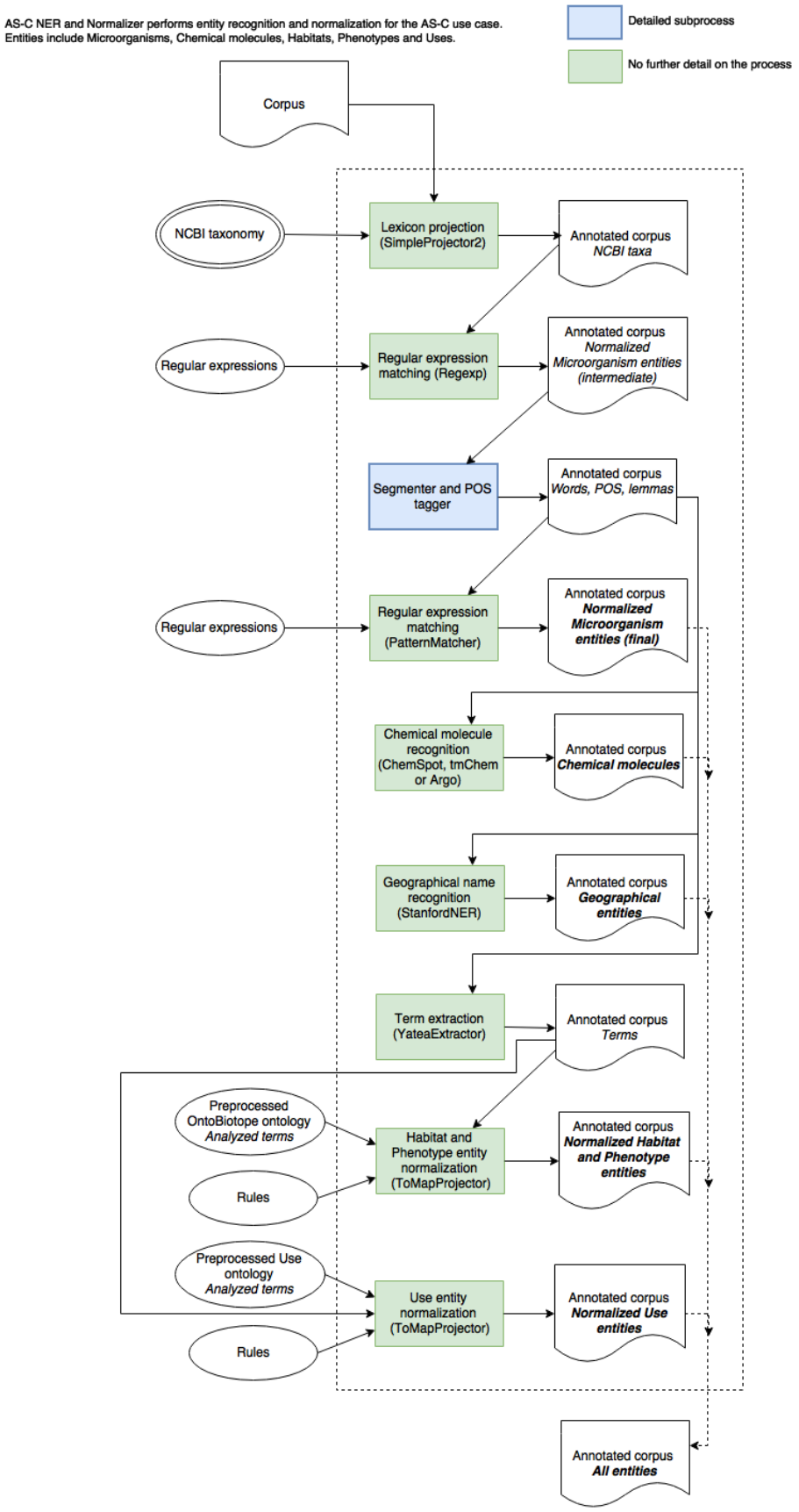


Figure 13. Architecture du traitement de reconnaissance et de normalisation des entités

Outils de construction de corpus

Ce sont les outils de la tâche Corpus Converter de la figure 13 ci-dessus.

XML Reader

- Tâche : Convertit les documents de PubMed au format XML (Corpus Converter).
- Lien vers la documentation : https://bibliome.github.io/alvisnlp/reference/module/XMLReader
- Format d'entrée : XML, les schémas XML spécifiques sont supportés par la configuration
- Ressources et modèles statiques utilisés : la structure du XML d'entrée est gérée par une feuille de style XSLT fournie par l'utilisateur. Quelques XSLTs sont disponibles avec la distribution AlvisNLP pour les schémas les plus utilisés (PubMed, PMC, ScienceDirect, HTML, Springer).

Outils de prétraitement de corpus

Ce sont les outils de la tâche Segmenter POS Tagger de la figure 14 ci-dessous.

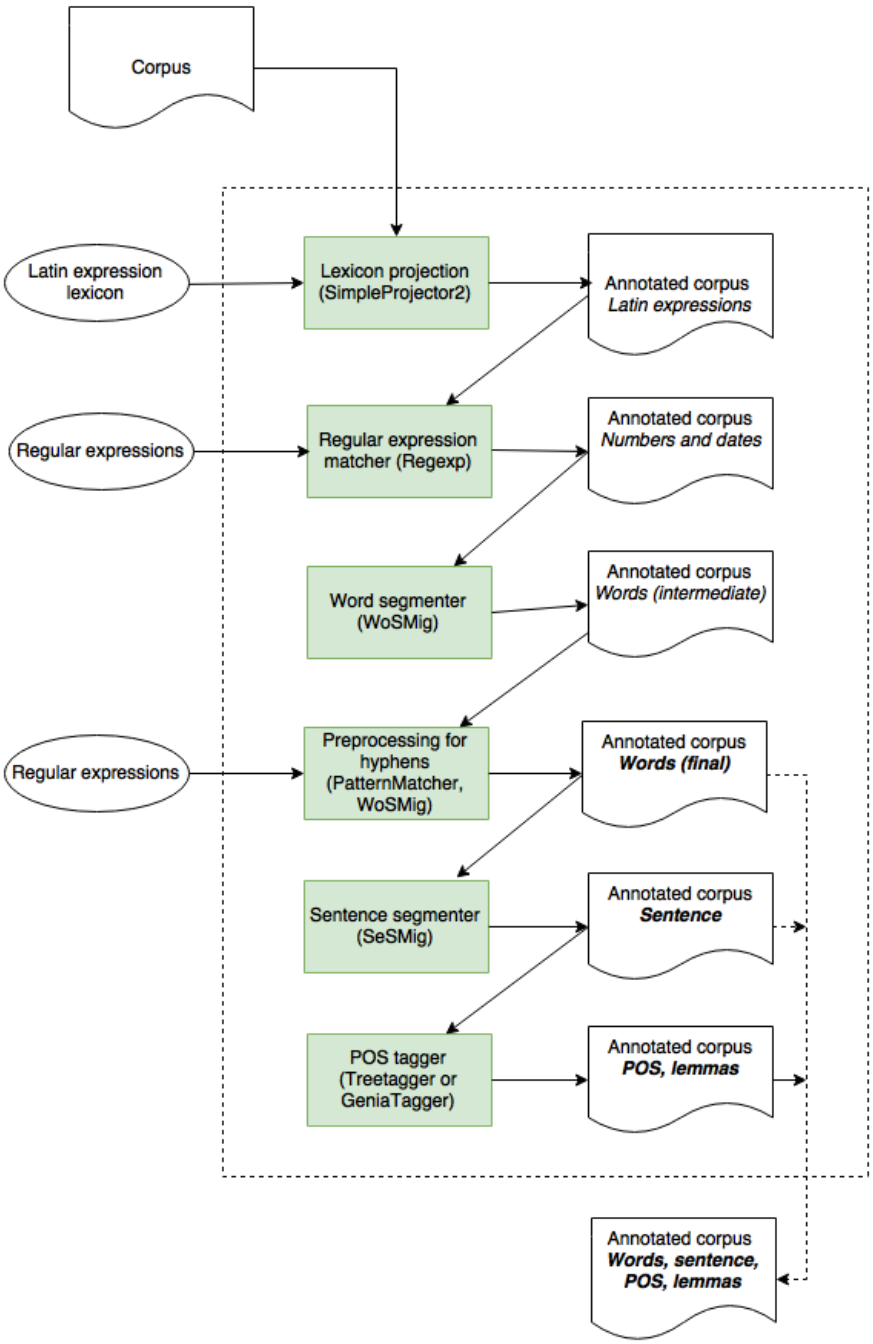


Figure 14. Architecture de la segmentation en mots et en phrases puis étiquetage morpho-syntaxique

WoSMig

- Tâche: Segmentation de mots (Tokenisation) dans Segmenter-POS-Tagger
- Lien vers la documentation : https://bibliome.github.io/alvisnlp/reference/module/WoSMig

SeSMig

- Tâche: Segmentation de phrases dans Segmenter-POS-Tagger
- Lien vers la documentation : https://bibliome.github.io/alvisnlp/reference/module/SeSMig

Genia Tagger

- Tâche : Étiquetage morphosyntaxique et lemmatisation dans Segmenter-POS-Tagger
- Lien vers la documentation : http://www.nactem.ac.uk/GENIA/tagger/
- https://bibliome.github.io/alvisnlp/reference/module/GeniaTagger
- Configuration requise pour l'environnement système (système d'exploitation, librairies installées) : AlvisNLP, Genia Tagger v3.0.2 (http://www.nactem.ac.uk/GENIA/tagger/)
- Utilisation de ressources et de modèles statiques : Modèles POS tag/chunk/EN

Outils de reconnaissance d'entités et de normalisation

Ces outils sont ceux décrits dans la figure 15.

StanfordNER

- Tâche : reconnaissance d'entités nommées (entités géographiques)
- Lien vers la documentation : http://nlp.stanford.edu/software/CRF-NER.shtml https://bibliome.github.io/alvisnlp/reference/module/StanfordNERhttp://nlp.stanford.edu/software/CRF-NER.shtml
- Contraintes relatives à l'environnement système : AlvisNLP et Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/)
- Utilisation de ressources et de modèles statiques : NER CRF model

TabularProjector

- Tâche: Projection de lexique. (REN et normalisation)
- Lien vers la documentation : https://bibliome.github.io/alvisnlp/reference/module/TabularProjector
- Format du vocabulaire : fichier texte tabulaire (une entrée par ligne)
- Ressources et modèles statiques utilisés : peuvent être utilisés avec n'importe quel lexique (dans le format requis)

RegExp

- Tâche : applique une expression régulière sur le contenu des sections et crée une annotation pour chaque correspondance. (REN et normalisation)
- Lien vers la documentation : https://bibliome.github.io/alvisnlp/reference/module/RegExp
- Utilisation de ressources et de modèles statiques : Expressions régulières

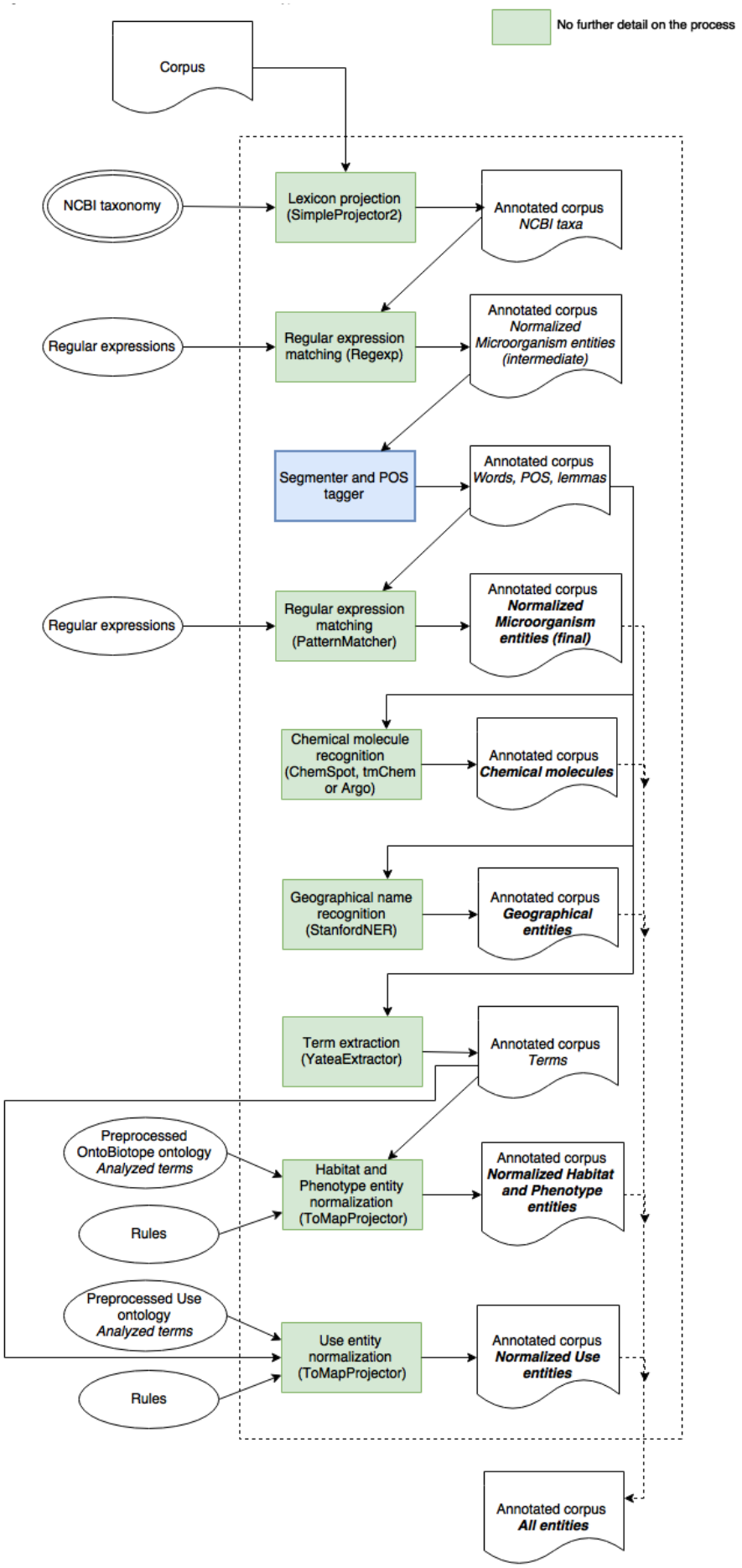


Figure 15. Architecture de la reconnaissance et de la normalisation d'entités

PatternMatcher

- Tâche: applique un motif spécifique sur le contenu des sections et crée une annotation pour chaque correspondance. (REN et normalisation)
- Lien vers la documentation :
- https://bibliome.github.io/alvisnlp/reference/module/PatternMatcher
- Utilisation de ressources et de modèles statiques : Motifs d'expression régulière

YaTea Extracteur de termes

- Tâche: Extraction de termes pour la REN
- Lien vers la documentation :
- https://bibliome.github.io/alvisnlp/reference/module/YateaExtractor
- Formats d'entrée/sortie : Représentation interne d'Alvis, termes extraits en XML propriétaire
- Configuration requise pour l'environnement système (système d'exploitation, librairies installées, interprètes de langue) : AlvisNLP/ML, Lingua : Yatea (https://metacpan.org/pod/distribution/Lingua-YaTeA/lib/Lingua/YaTeA.pm)
- Ressources statiques et modèles utilisés : modèles d'extraction de termes

ToMapProjector

- Tâche: Normalisation d'entités avec des concepts d'ontologie ou de terminologie (REN et normalisation)
- Lien vers la documentation :
- https://bibliome.github.io/alvisnlp/reference/module/TomapProjector Formats d'entrée/sortie : Représentation interne AlvisNLP, analyse du lexique sérialisé XML, extraction des termes extraits par YaTeA en XML
- Mode de déploiement : Module AlvisNLP
- Exigences pour l'environnement système (OS, et limites (mémoire utilisée, vitesse de traitement, multi-threading) : RAM pour la taille des corpus et lexiques
- Ressources statiques et modèles utilisés : extraction de termes, lexique pré-analysé/ontologie

ToMapTrain

- Tâche : Analyser les termes d'une ontologie/un lexique afin de les utiliser avec le module de normalisation (ToMapProjector).
- Lien vers la documentation :
- https://bibliome.github.io/alvisnlp/reference/module/TomapTrain
- Formats d'entrée/sortie : Représentation interne AlvisNLP, lexique au format OBO, analyse lexicale sérialisée XML
- Exigences pour l'environnement système (OS, et limites (mémoire utilisée, vitesse de traitement, multi-threading) : RAM pour la taille des corpus et lexiques
- Ressources statiques et modèles utilisés : ontologie/lexique à analyser

3.4 Services

Téléchargement de la taxonomie NCBI

> Adresse : ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz
> Tâches : récupérer la dernière version de la taxonomie NCBI
> Formats : propriétaire, tar.gz
> Conditions d'utilisation (en particulier limitations de performance) : https://uts.nlm.nih.gov/license.html libre d'utilisation, pas OA/OSS
> Institution d'accueil : NCBI
> Type d'interface : Téléchargement FTP
> Commentaire : non inclus dans le workflow OMTD

3.5 Autres

PubMed

> Adresse : https://www.ncbi.nlm.nih.gov/pubmed
> Tâches : identifier les listes de publications scientifiques pertinentes (Corpus Builder) à partir des requêtes des utilisateurs.

3.6 Interfaces de données

L'application prend en entrée trois types de bases de connaissances encapsulées dans l'image Docker : Taxonomie NCBI, l'ontologie OntoBiotope et le lexique d'expressions latines.
La taxonomie du NCBI et l'ontologie Ontobiotope sont exportées par l'application de fouille de textes vers l'application cliente.
L'application utilise les corpus en entrée :

> Les corpus (PubMed, GenBank, BacDive, CIRM) sont gérés par le composant Corpus Builder
> Les corpus utilisés pour l'évaluation sont ceux de Bacteria Biotope

L'application exporte des annotations textuelles vers les applications clientes, de courts extraits (snippets) et des informations bibliographiques par l'outil Exportation d'annotations.

3.7 Interfaces utilisateurs

OpenMinTeD

L'interface utilisateur primaire est celle de la plateforme OpenMinTeD[13] qui permet de réutiliser l'application sur différents corpus et d'en exploiter les résultats au format XMI. La figure 16 montre l'écran d'accueil.

[13] https://services.openminted.eu/home

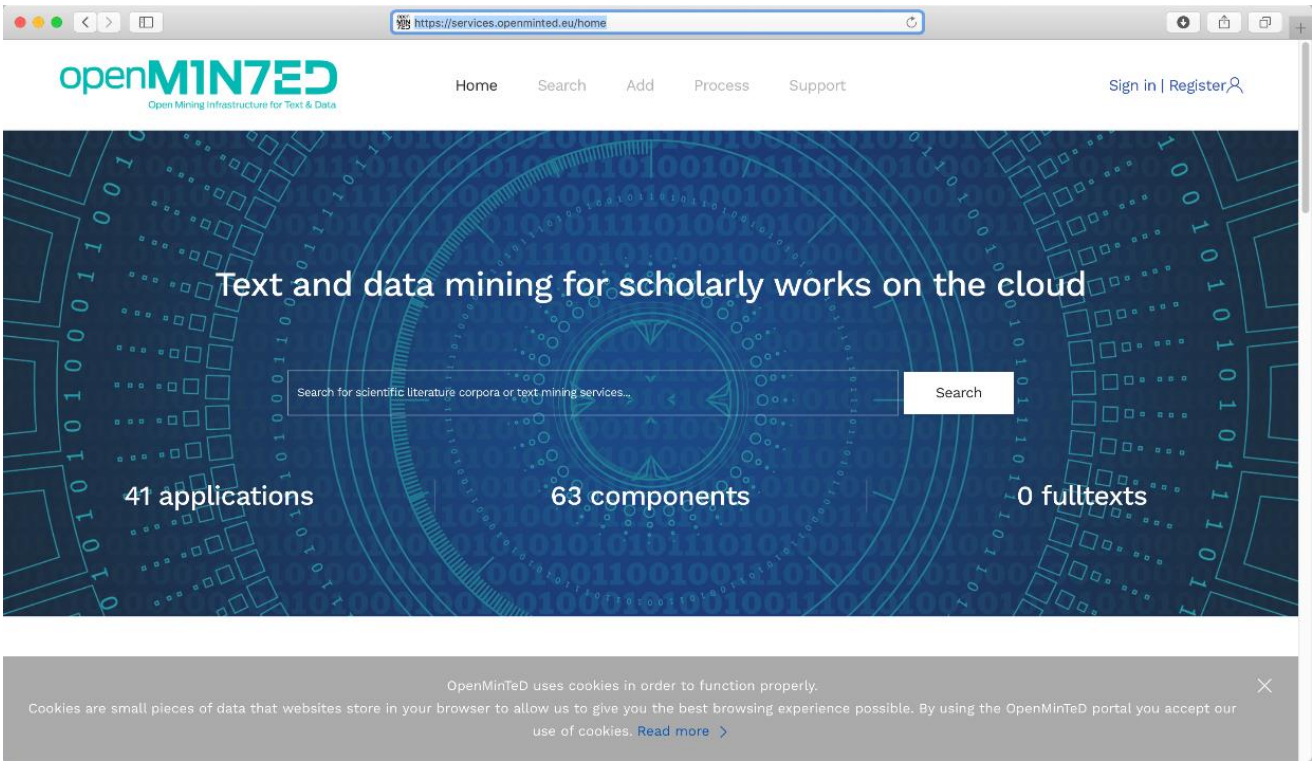


Figure 16. Écran d'accueil de la plateforme de service OpenMinTeD.

La figure 17 représente le résultat de la recherche avec le mot clef Microbe.

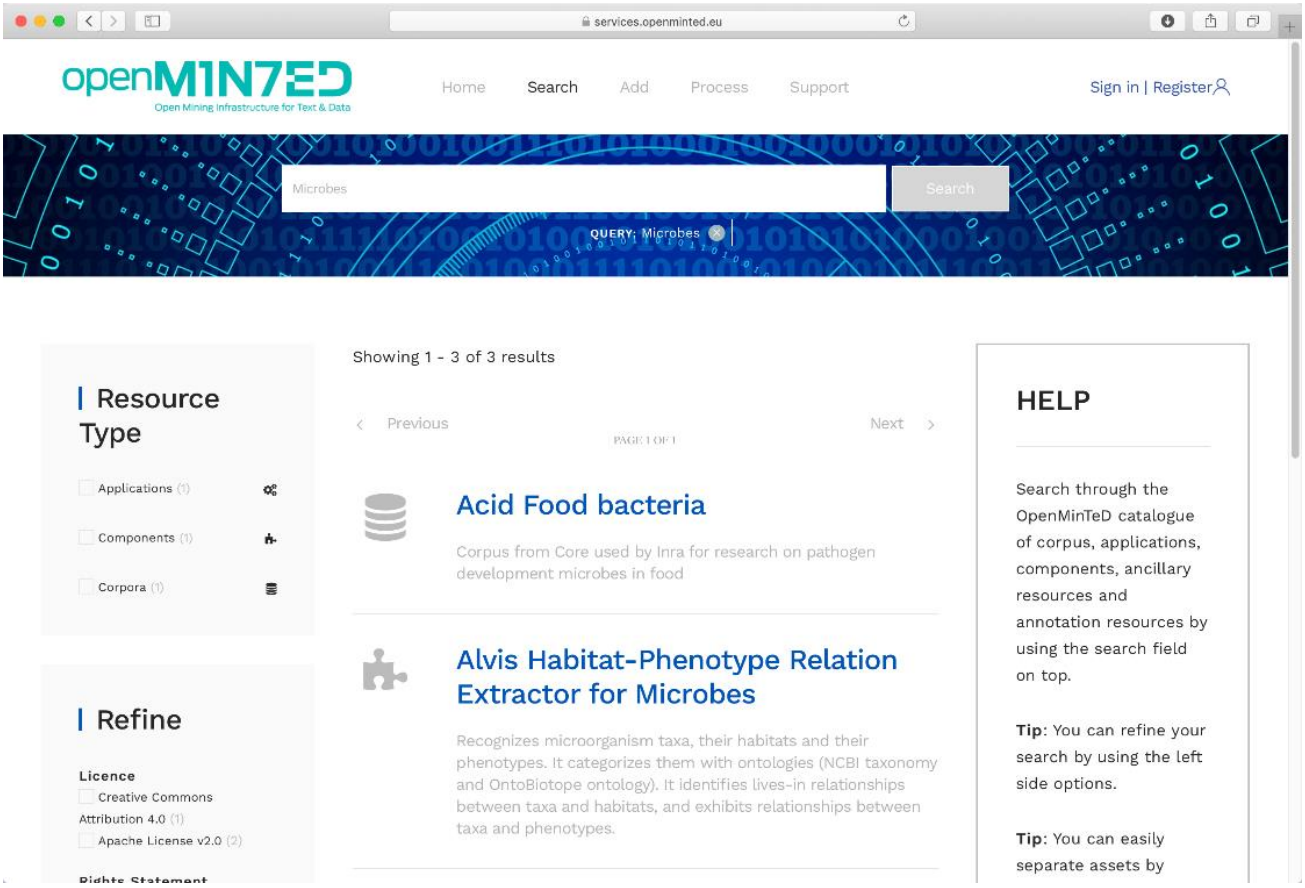


Figure 17. Résultat de la recherche d'application ou de corpus avec le mot clef Microbe sur la plateforme de service OpenMinTeD.

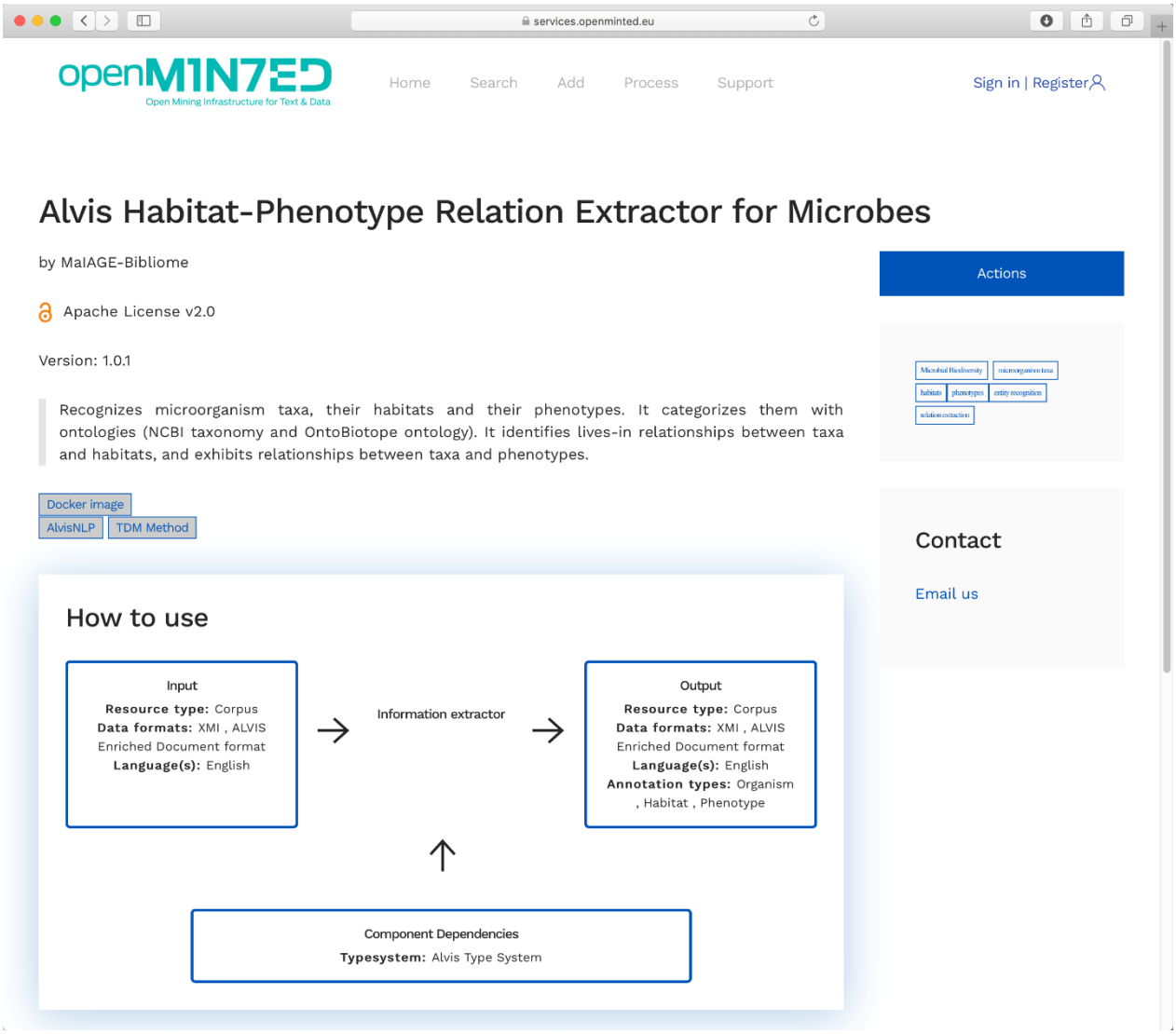


Figure 18. Description de l'application Microbiologie sur la plateforme de service OpenMinTeD

Pour exécuter l'application, il faut être connecté via la fédération d'identité EduGain ou par Facebook puis cliquer sur "Process".

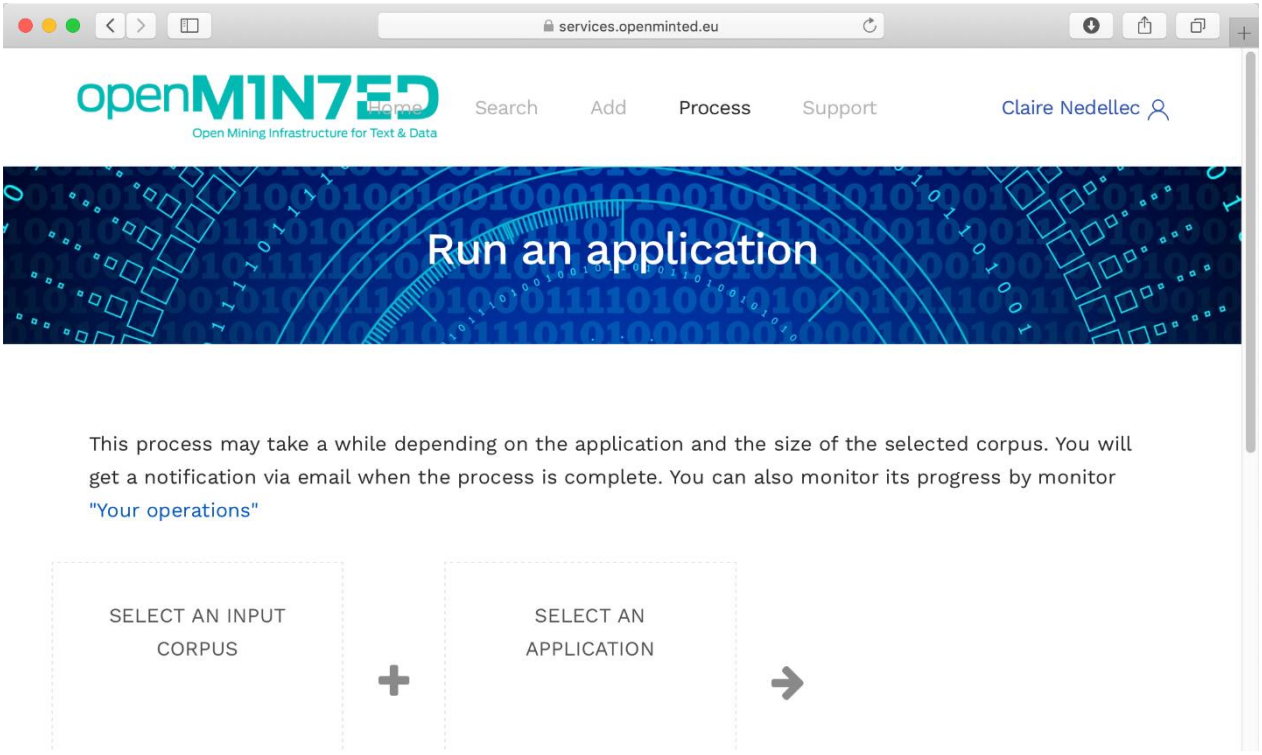


Figure 19. Accès à la fonction d'exécution Process sur la plateforme de service OpenMinTeD

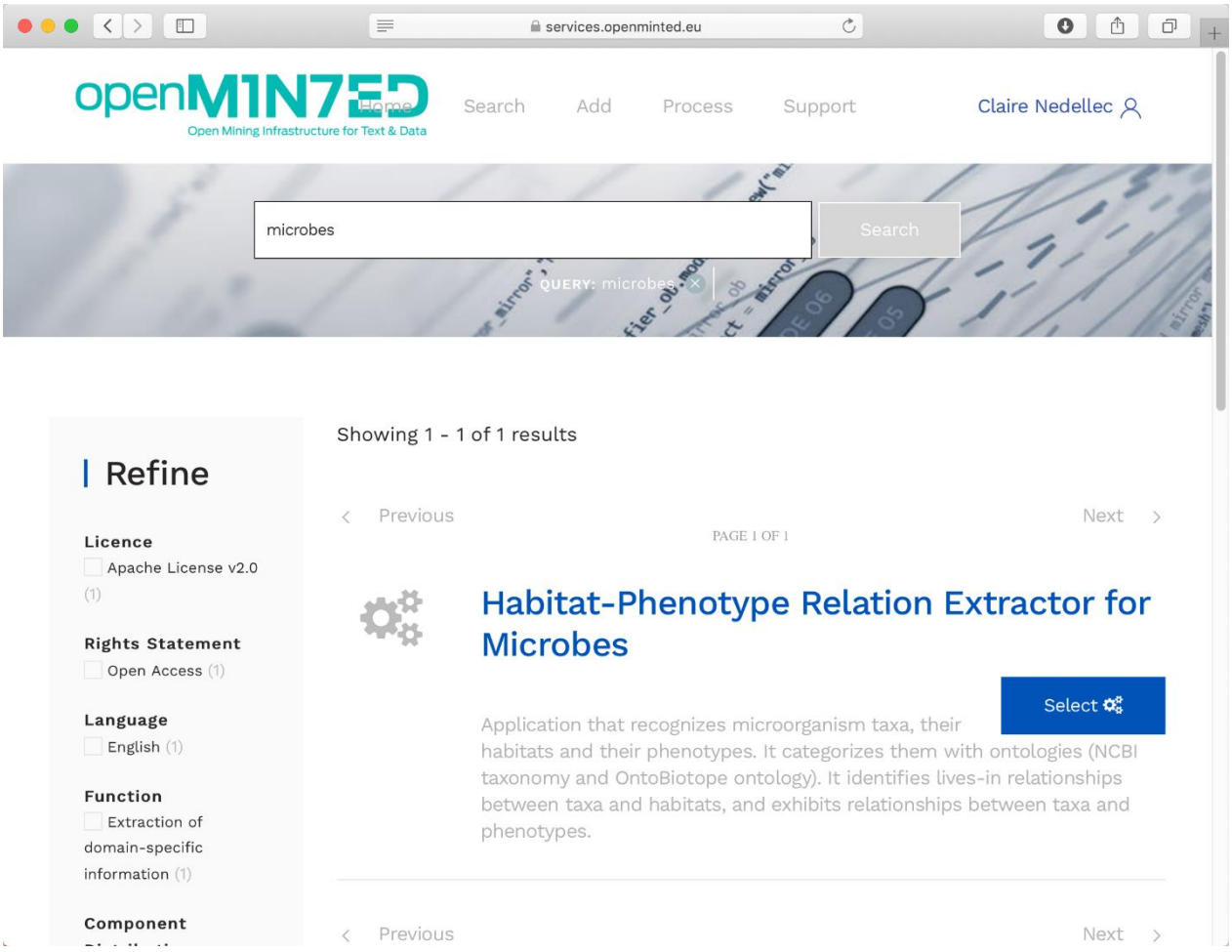


Figure 20. Choix de l'application Microbiologie

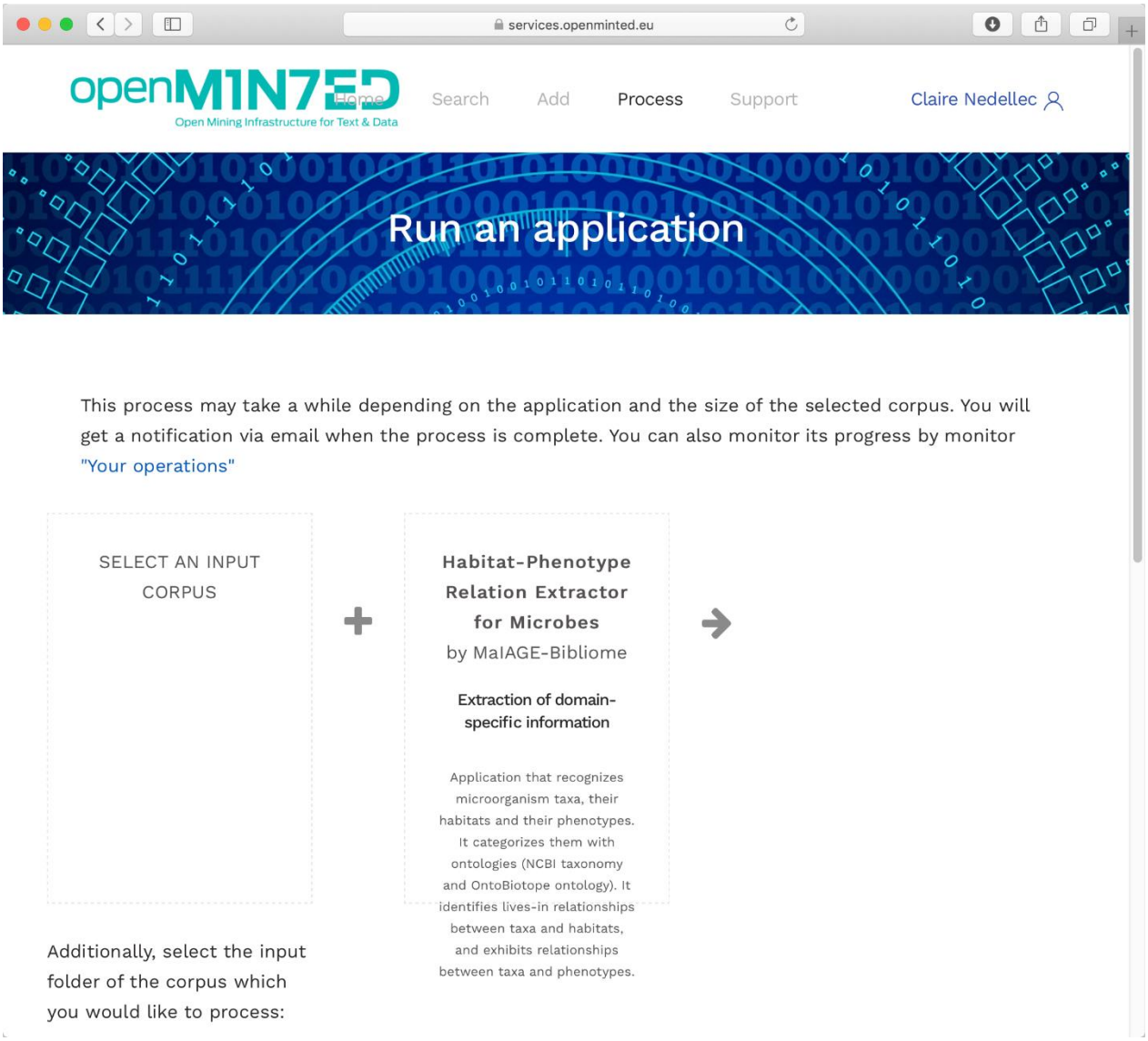


Figure 21. Exécution de l'application Microbiologie sur la plateforme de service OpenMinTeD

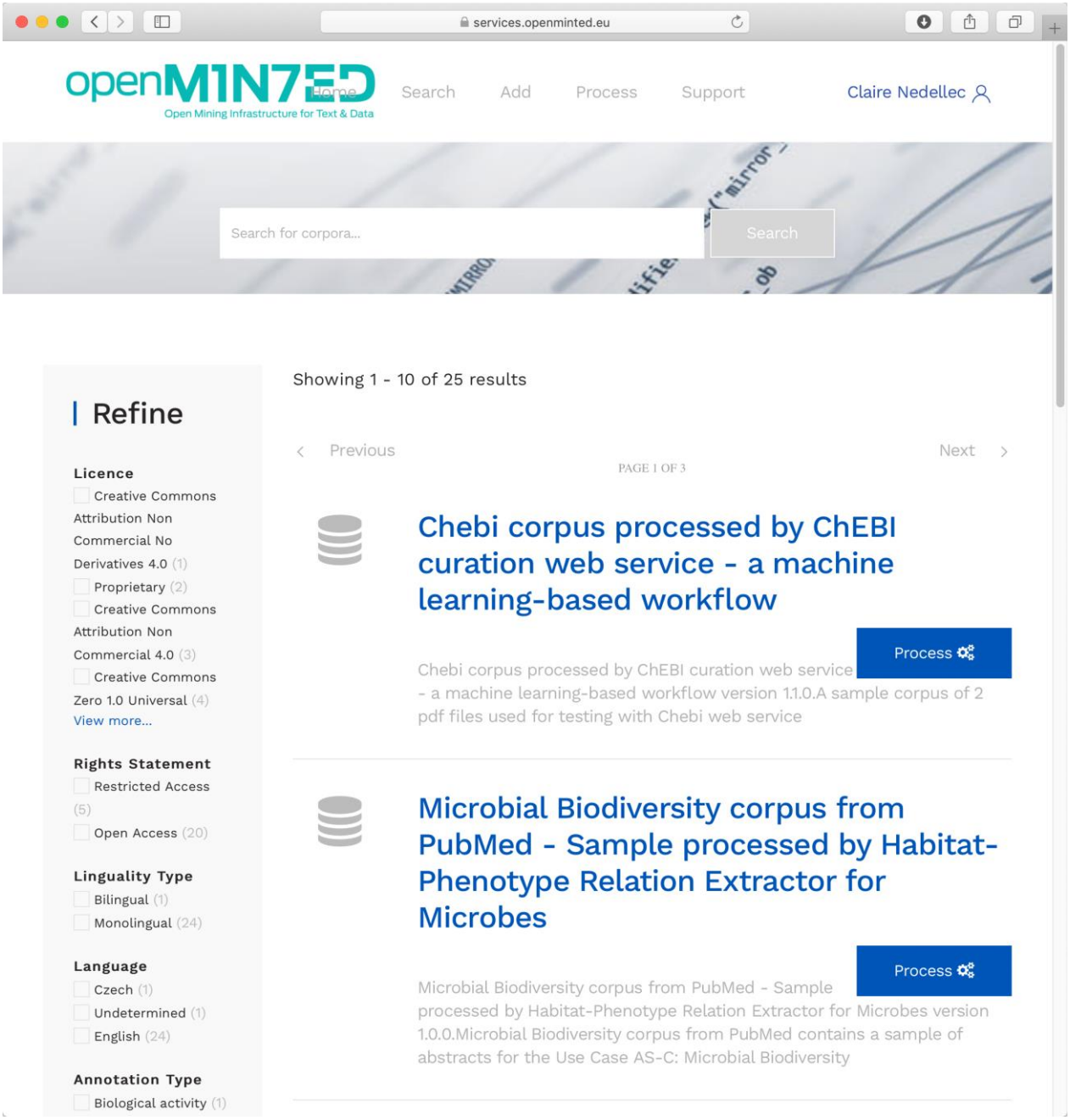


Figure 22. Choix d'un corpus pour l'application Microbiologie sur la plateforme de service OpenMinTeD

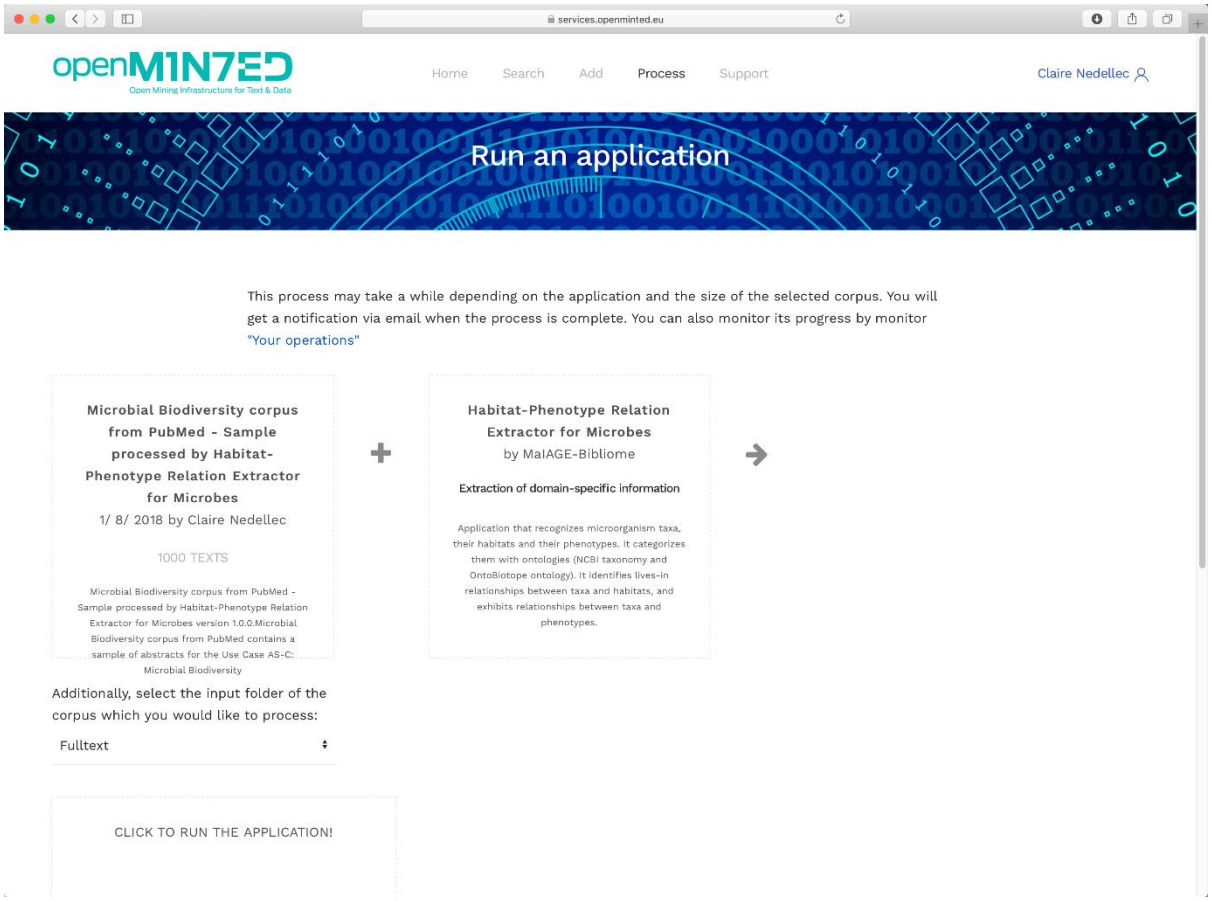


Figure 23. Exécution de l'application Microbiologie avec le corpus choisi sur la plateforme de service OpenMinTeD

Interfaces des applications clientes

Pour une recherche plus précise dans les données, nous avons développé deux applications clientes pour l'application de fouille de textes qui présentent des interfaces utilisateur plus riches.

- Interaction de l'utilisateur final avec l'application client Florilège : interrogation et navigation et exportation des résultats de l'interrogation (Figure 24).
- Interaction de l'utilisateur final avec l'application du moteur de recherche sémantique AlvisIR : interrogation, navigation ontologique et sélection des facettes. La figure 25 donne un exemple d'une requête de relation signifiant "quels Lactobacillus vivent dans le fromage".

La section suivante décrit les usages et leur impact.

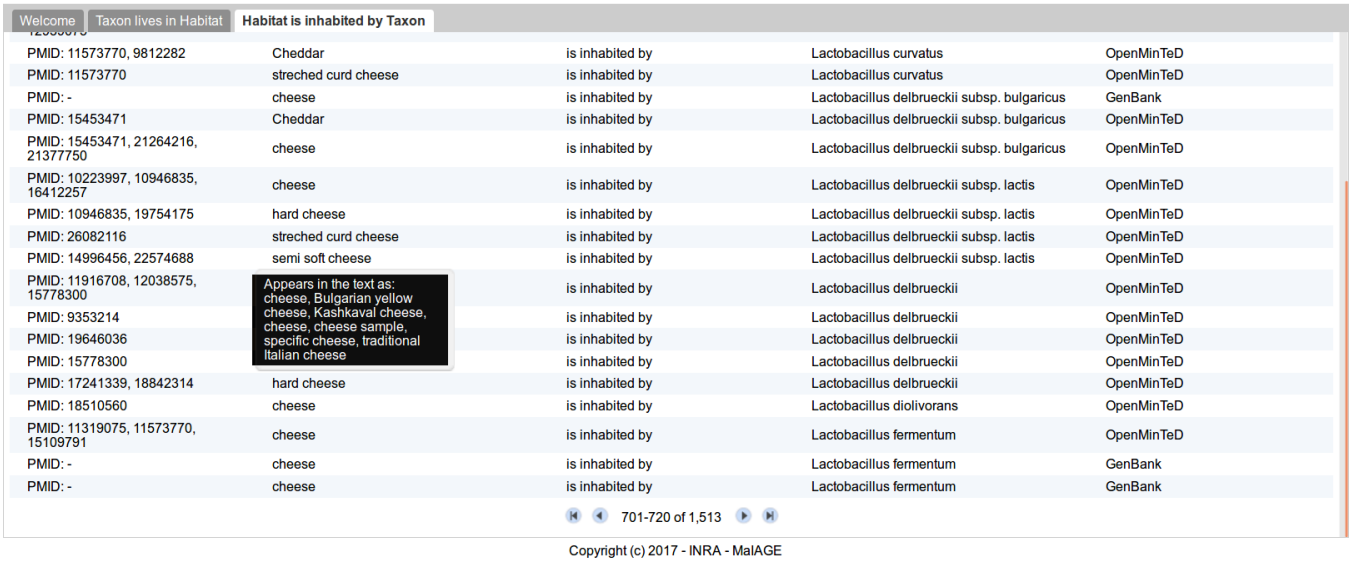


Figure 24. Copie d'écran de l'application cliente Florilège.

Figure 25. Copie d'écran de l'application cliente Alvis Food Semantic Search Engine

3.8 Applications Clientes

Les deux applications clientes de l'application fouille de textes sont génériques. Le moteur de recherche sémantique AlvisIR développé et maintenu par l'unité MaIAGE est réutilisé dans de très nombreuses applications. Son déploiement est très rapide (1/2 journée). L'application

cliente Florilège est nouvelle mais basée sur des technologies non spécifiques de bases de données relationnelles de manière à favoriser la généralisation de l'approche.

Application cliente Florilège

L'application Cliente Florilège est maintenue par la plateforme Migale. Son interface utilisateur est une application web accessible par un navigateur[14].

Technologies

Les technologies utilisées ici sont traditionnelles : les données sont gérées par le SGBD PostgreSQL. Les données sont reçues de l'application fouille de textes sous forme de fichier structuré. L'interface est développée en GWT. L'indexation hiérarchique est réalisée en indexant les données par le chemin jusqu'à la racine, dans la hiérarchie.

Interface

L'écran d'accueil (Figure 26) donne accès à quatre onglets qui permettent d'interroger les données des deux relations Taxon lives in Habitat et Taxon exhibits Phenotype, en ciblant l'un ou l'autre des arguments.

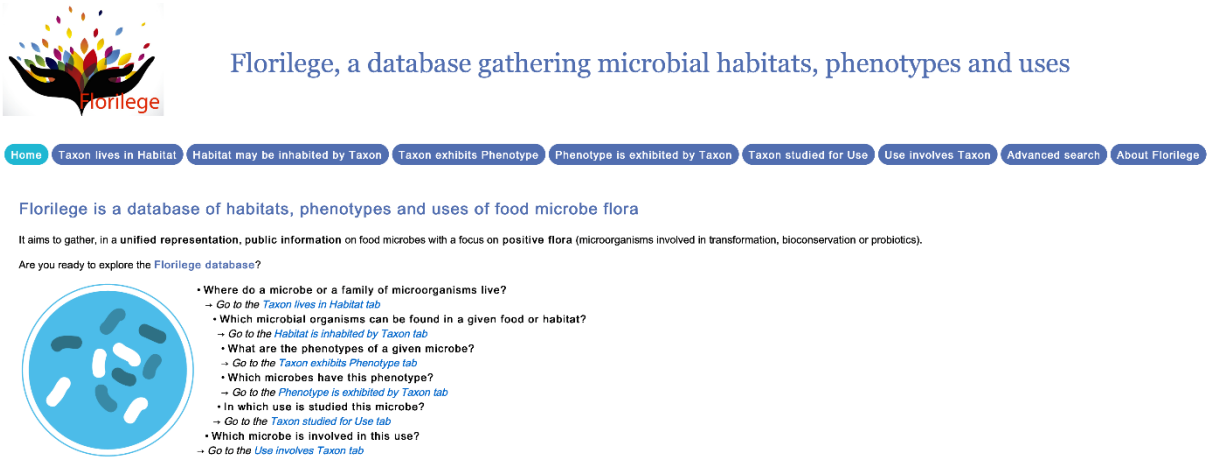


Figure 26. Écran d'accueil de l'application cliente Florilege

Le premier onglet, Taxon lives in Habitat permet d'interroger la base par une requête sur un taxon, famille, espèce, souche. La Figure 27 donne un exemple d'interrogation avec *Lactococcus plantarum*.

La colonne Source Text indique les références des documents. Elles sont cliquables et ouvrent une nouvelle fenêtre affichant le texte source. La colonne Taxon donne le nom normalisé. Le passage de la souris ouvre une fenêtre avec les formes extraites des documents source. La colonne Habitat donne les noms normalisés, Le passage de la souris ouvre une fenêtre avec les formes extraites des documents source. La colonne Source indique le nom de la base d'où proviennent les documents de la ligne.

[14] http://migale.jouy.inra.fr/Florilege/

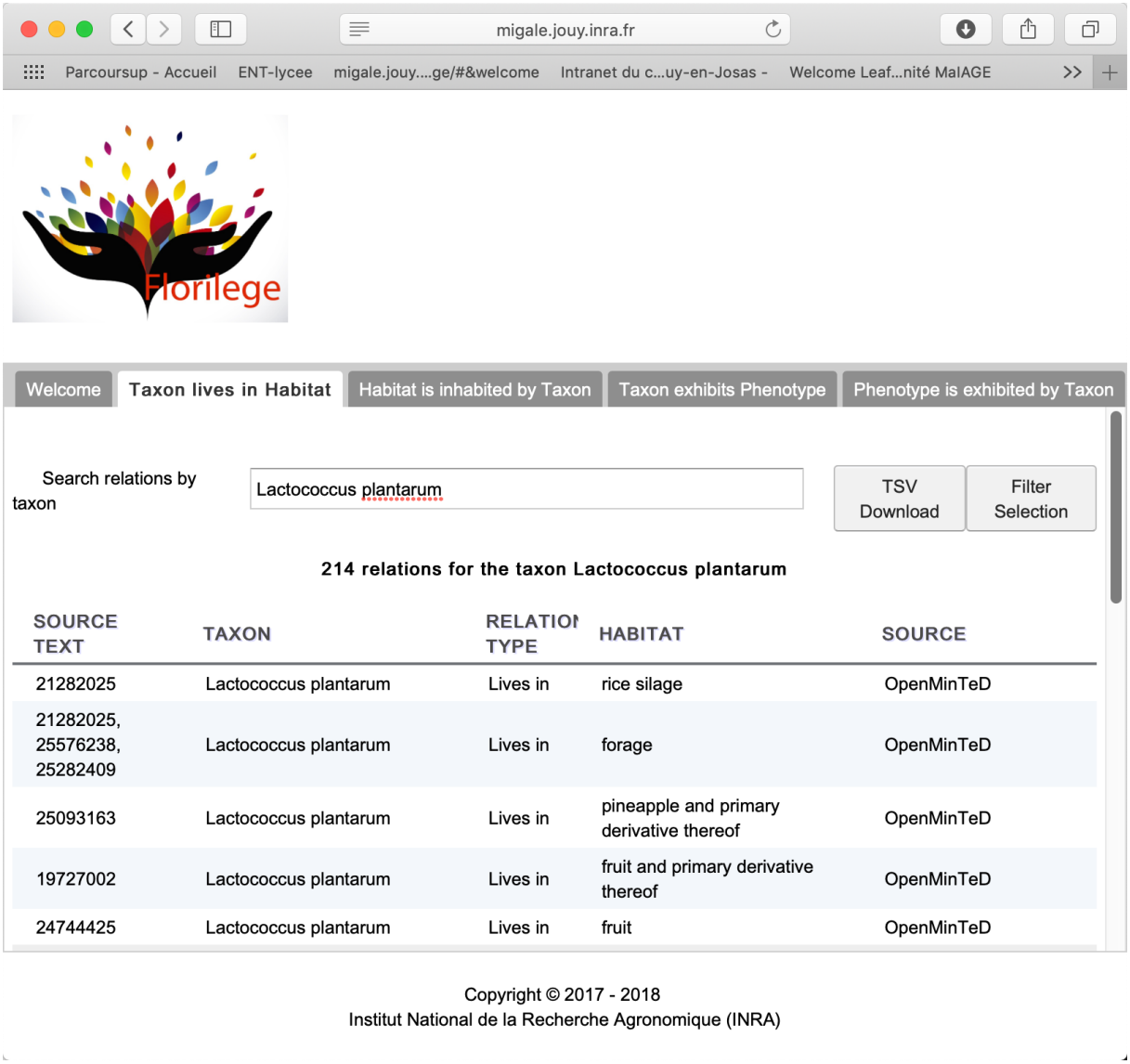


Figure 27. Écran de l'onglet Taxon lives in Habitat

Le bouton Filter Selection permet de sélectionner un habitat pour le microorganisme, ou une source de données, ou les microorganismes QPS only, c'est-à-dire inoffensifs. La figure 28 donne un exemple avec l'habitat *fermented food*. Les résultats filtrés sont maintenant *fermented milk*, *vinegar*, etc.

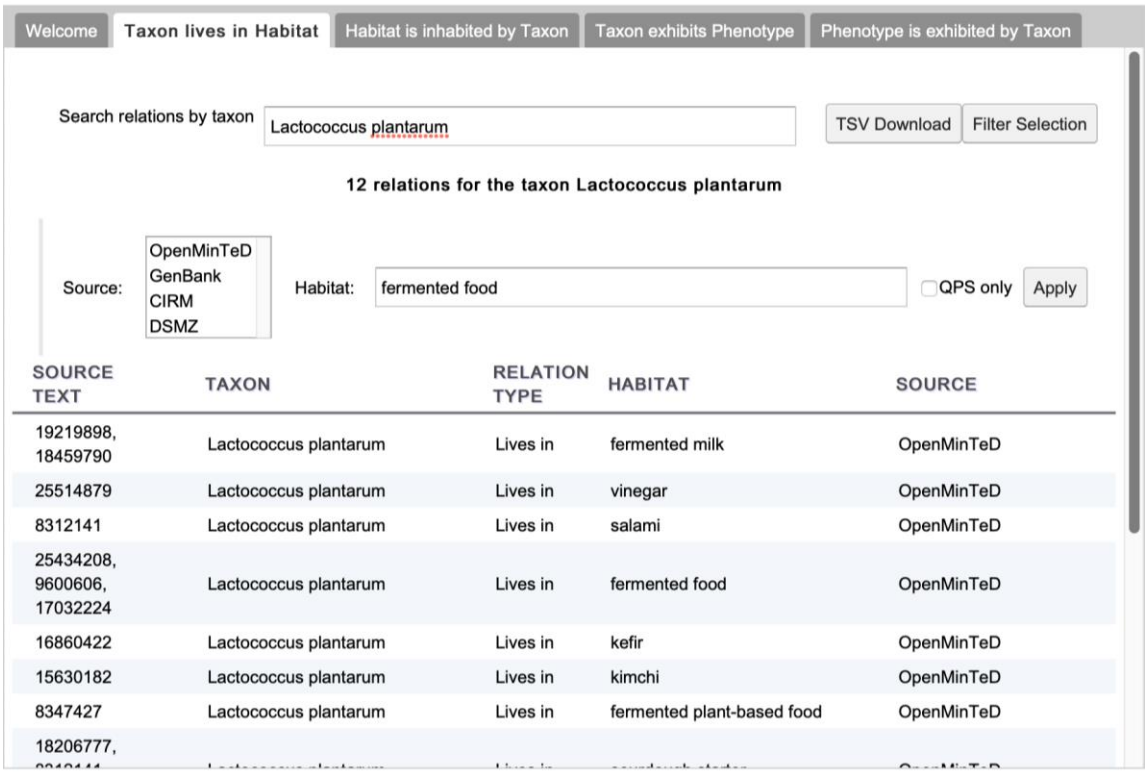


Figure 28. Écran de l'onglet Taxon lives in Habitat après sélection d'un habitat.

Le bouton TSV download permet de télécharger dans un format texte tabulé l'ensemble des réponses de la requête courante.

Le deuxième onglet, "Habitat is inhabited by Taxon" permet d'interroger la base par une requête sur un habitat à n'importe quel niveau de généralité. La Figure 29 donne un exemple d'interrogation avec *fermented food*.

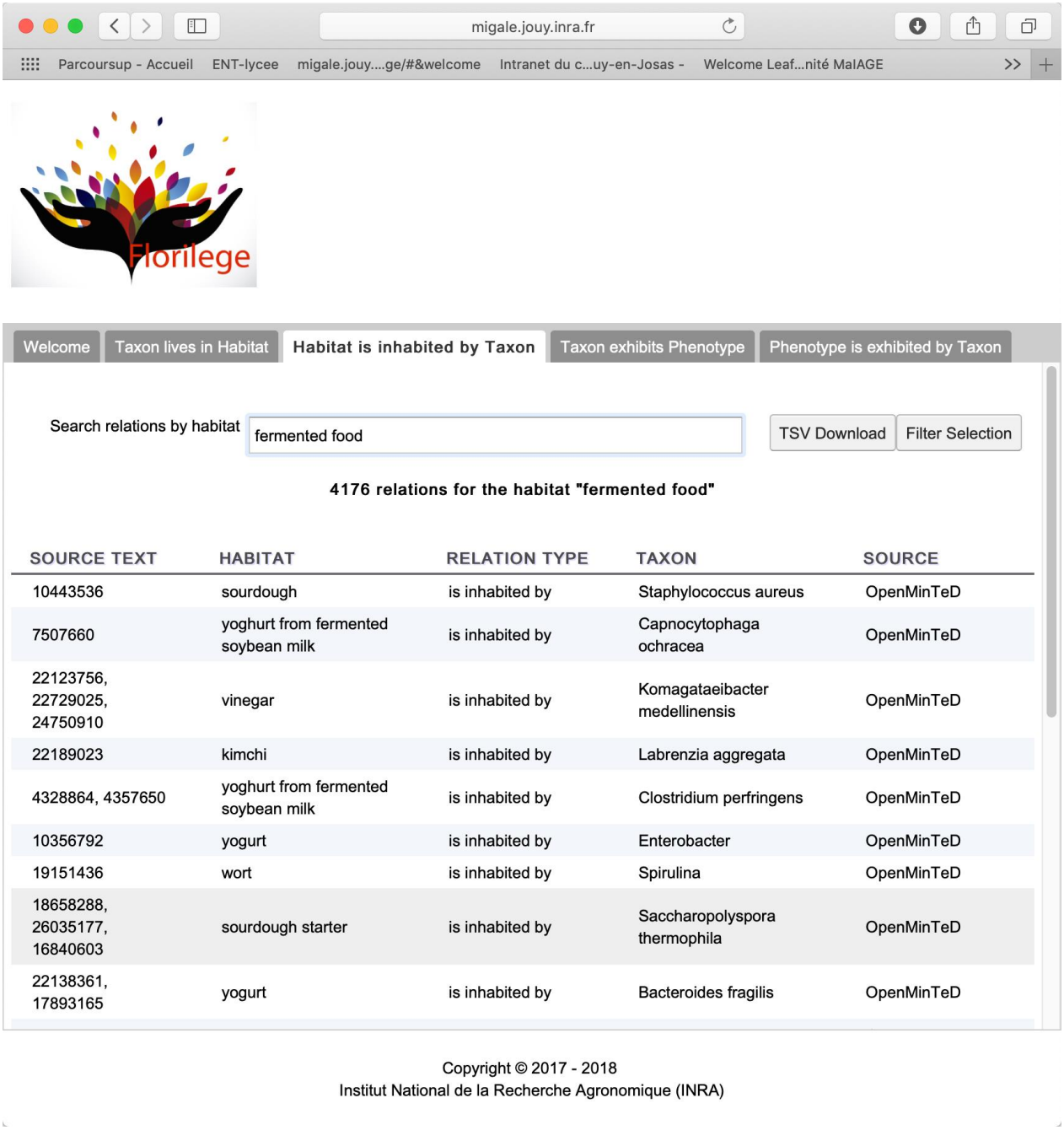


Figure 29. Écran de l'onglet Habitat is inhabited by Taxon.

Le troisième onglet, "Taxon exhibits Phenotype" permet d'interroger la base par une requête sur un phénotype de niveau de généralité quelconque. La Figure 30 donne un exemple d'interrogation avec *Mycobacterium tuberculosis*.

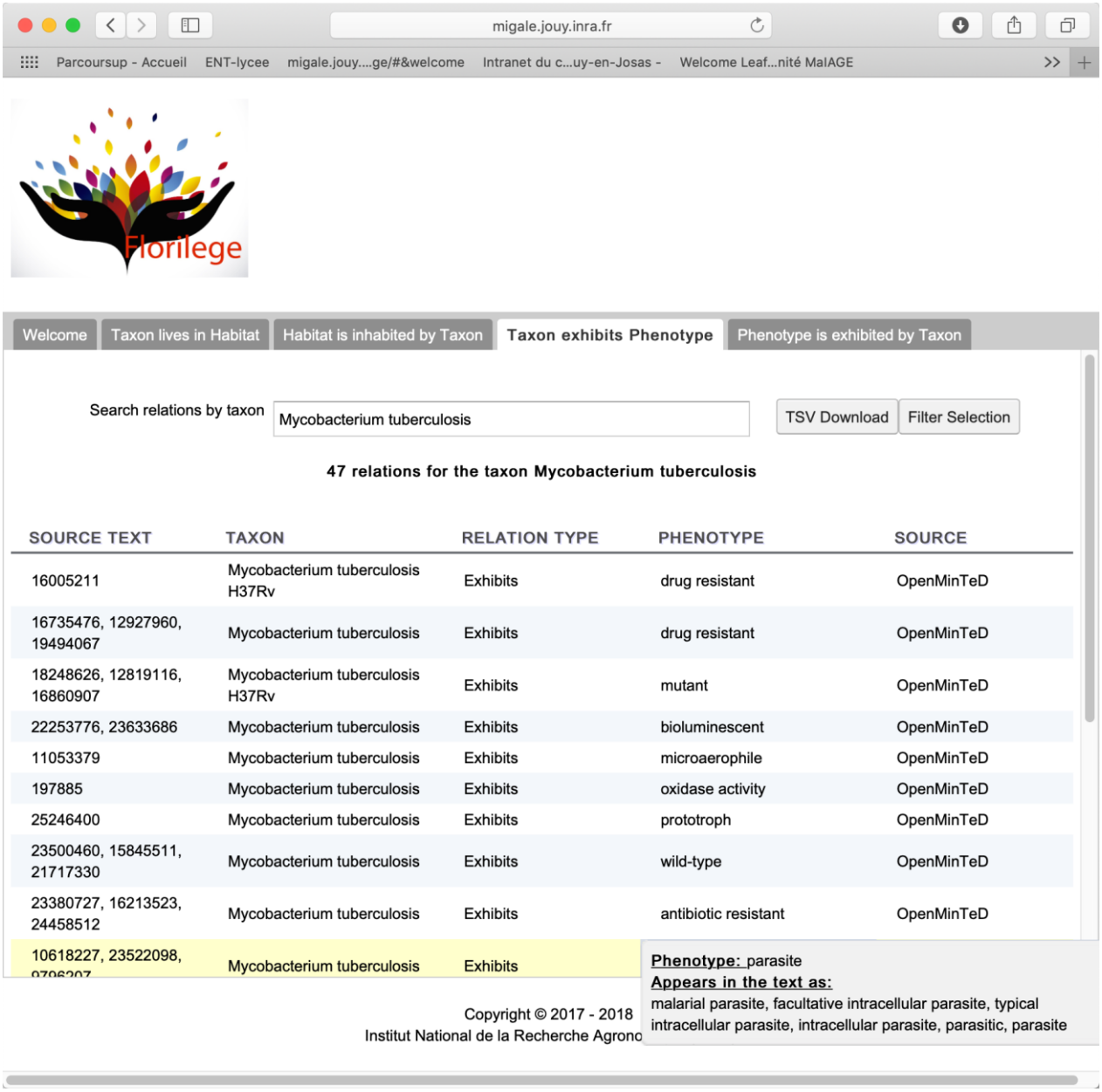


Figure 30. Écran de l'onglet Taxon exhibits Phenotype.

Le passage de la souris ouvre une fenêtre avec les formes extraites des documents source (*parasite* dans l'exemple de la figure 30).

Le quatrième onglet, "Phenotype is exhibited by Taxon" permet d'interroger la base par une requête sur un phénotype. La Figure 31 donne un exemple d'interrogation avec *mesophile* et une sélection du taxon *Lactococcus*.

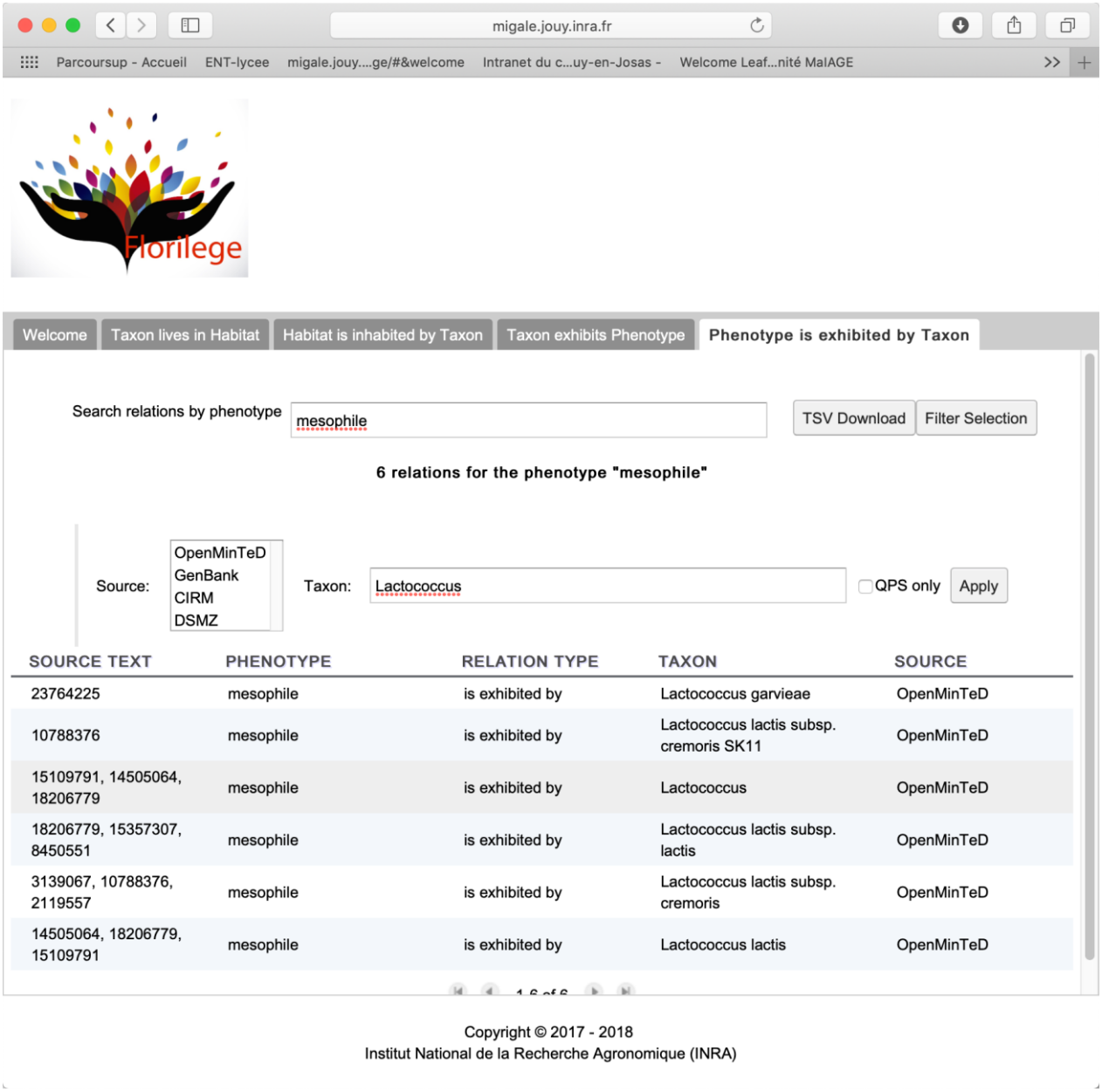


Figure 31. Ecran de l'onglet Phenotype is exhibited by Taxon.

Application cliente AlvisIR

L'application AlvisIR reçoit les données de l'application fouille de textes sous forme de fichier structuré. AlvisIR repose sur le moteur Lucene étendu pour traiter les hiérarchies et les relations.

L'écran principal du moteur de recherche AlvisIR[15] (Figure 32) permet de saisir une requête portant sur les différentes entités et leurs relations.

[15] http://bibliome.jouy.inra.fr/demo/ontobiotope/alvisir2/webapi/search

Figure 32. Écran d'accueil du moteur de recherche AlvisIR pour la microbiologie.

La figure 33 en donne un exemple avec la requête sur le taxon Xylella fastidiosa, pathogène connu de plantes, la relation lives in et l'habitat général "microbial habitat".

Figure 33. Exemple d'interrogation de relation avec le moteur de recherche AlvisiR.

La requête peut être formulée à partir de la navigation dans l'ontologie (Figure 34). Elle s'ouvre sur un clic sur l'arbre à droite du bouton "search". Plus d'informations sur l'interface utilisateur peuvent être trouvées dans le didacticiel sur la plate-forme FOSTER[16],[17].

[16] https://www.fosteropenscience.eu/node/2292
[17] https://www.fosteropenscience.eu/node/1858

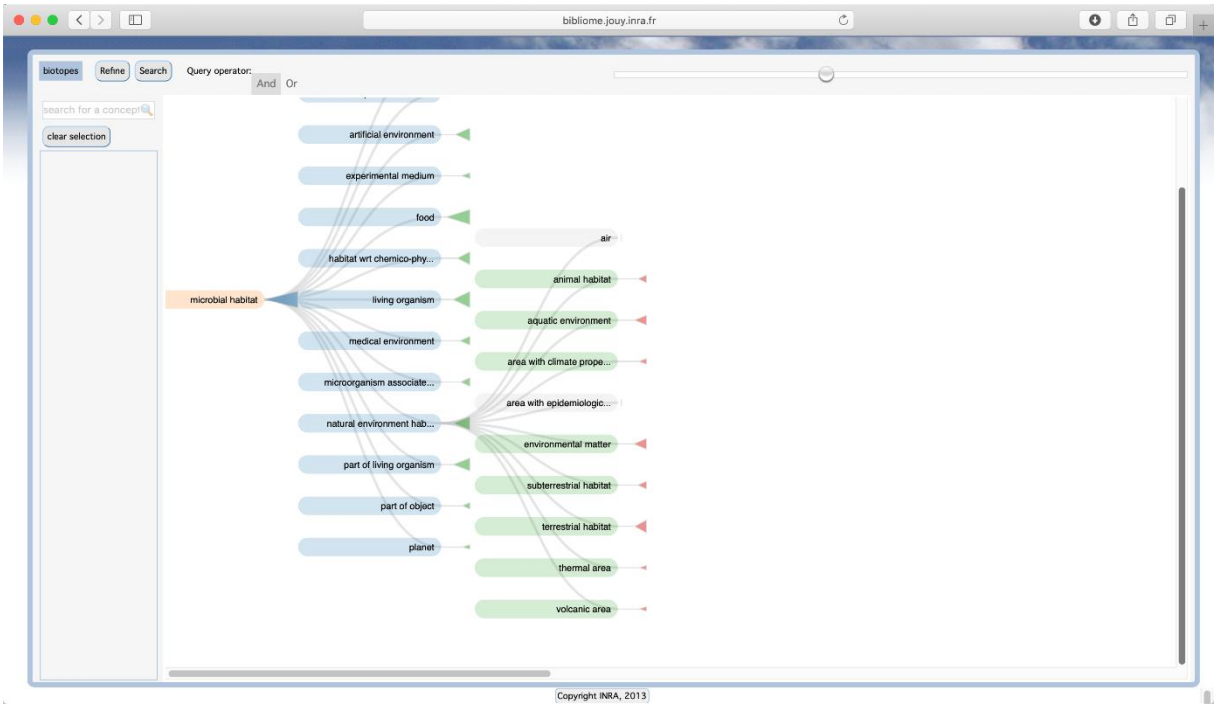
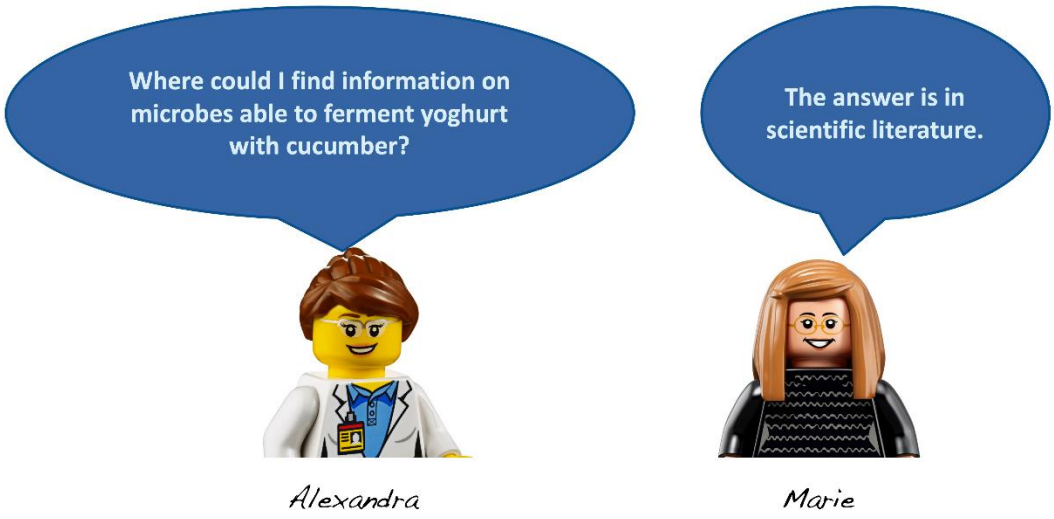


Figure 34. Exemple d'affichage de l'ontologie OntoBiotope avec le moteur de recherche AlvisiR.

3.9 Scénario d'utilisation

La vidéo en ligne[18] donne un exemple complet de l'utilisation de l'application fouille de textes sur la plateforme OpenMinTeD et avec l'application cliente Florilège.

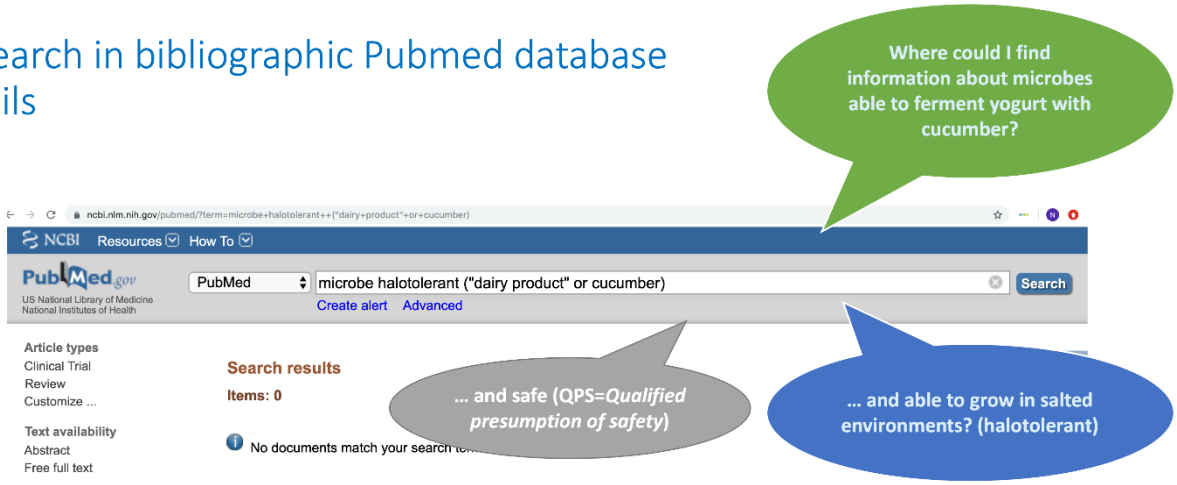
Elle met en scène Alexandra, biologiste et Marie, bioinformaticienne. Alexandra cherche des informations sur des microbes capable de fermenter un mélange de yaourt et de concombre pour un nouveau produit alimentaire. Marie sait qu'il n'y a pas d'ensemble de données directement disponible pour répondre à de tels besoins.



[18] https://visatm.inist.fr/wp-content/uploads/WP9-AgriBiodiv-Use-Case-demo-V4.mp4

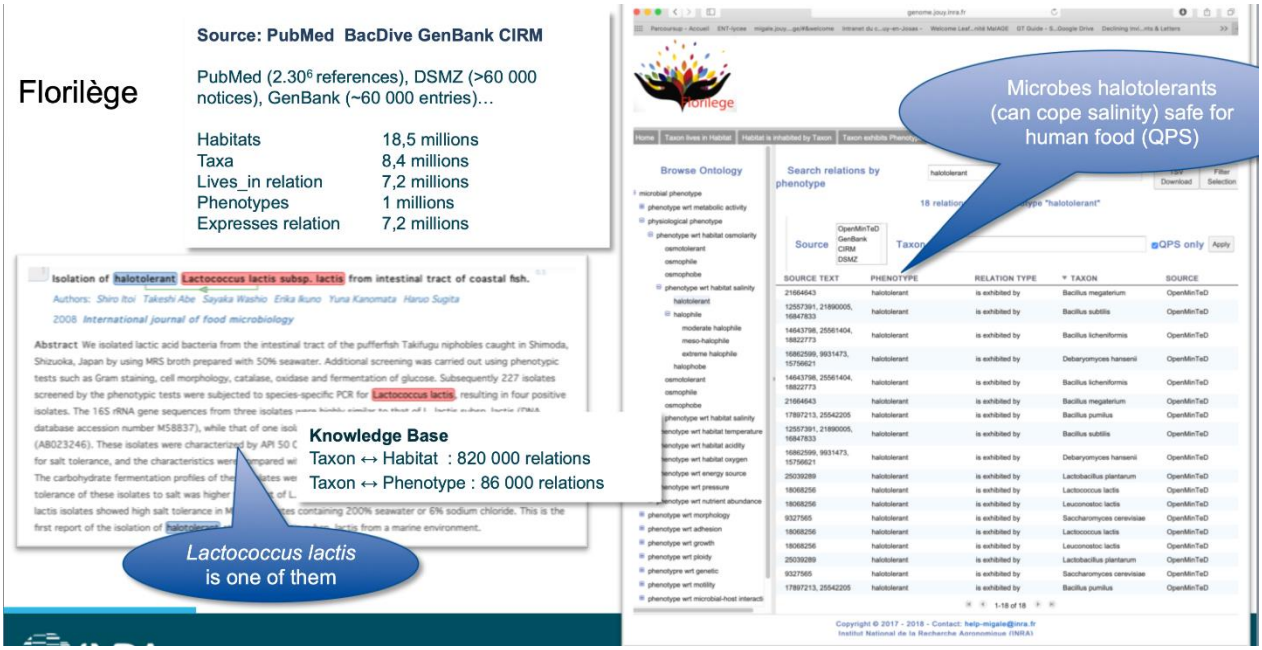
La recherche simple dans les moteurs de recherche classiques échoue. Répondre à la question de Marie nécessite l'analyse sémantique des entités, des catégories et des relations des articles scientifiques en microbiologie. C'est ce que fait l'extraction de l'information.

Search in bibliographic Pubmed database fails



La vidéo[19] illustre comment Marie explore la plateforme OpenMinTeD pour trouver une application pertinente puis l'exécute pour examiner si les résultats sont pertinents pour son besoin.

Marie montre ensuite à Alexandra comment interroger les applications clientes externes d'OpenMinTeD Florilege et AlvisIR, pour rechercher les taxa qui sont halotolérants et QPS (utilisable pour l'alimentation), par exemple *Lactococcus lactis subsp. lactis* puis en cherchant si les microbes proposés peuvent vivre dans du yaourt et du concombre, comme figuré dans les copies d'écran ci-dessous. *Lactococcus lactis subsp. lactis*. et bien d'autres répondent à ces critères.



[19] https://visatm.inist.fr/wp-content/uploads/WP9-AgriBiodiv-Use-Case-demo-V4.mp4

Figure 35. Exemple d'affichage pour le scénario.

Bilan

4.1 Évaluation des résultats de prédiction de l'application

Les résultats de prédiction des entités par l'application de fouille de textes sont mesurés grâce aux données de la compétition BioNLP-ST Bacteria Biotope 2016 [Deléger et al., 2016]. L'utilisation combinée de ToMap et de Contes (HONOR) permet d'obtenir les meilleurs résultats sur la tâche de normalisation, très supérieurs à l'état de l'art avec un score de 76% [REF PacLing].

Méthode	Score
Baseline	0.54
ToMap	0.66
BOUN	0.62
Turku	0.63
BOUNEL	0.66
CONTES (training)	0.61
CONTES (labels)	0.62
CONTES (training and labels)	0.70
HONOR (training)	0.72
HONOR (labels)	0.72
HONOR (training and labels)	0.76

Tableau 7. Résultats de la tâche de normalisation BioNLP-ST 2016 Bacteria Biotope.

4.2 Impact

En microbiologie

L'application de fouille de textes est exploitée à travers ses interfaces clientes Florilège et AlvisIR. Les projets d'utilisation directe portent sur des aspects très divers de recherche et d'innovation en microbiologie :

- l'identification de nouveaux microbes candidats pour concevoir des jus fermentés végétaux à base d'avoine et de lupin (unité STLO-INRA), projet en cours ENovFood
- la détection automatique de mentions de bioagresseurs, leur milieu et le lieu d'observation, à partir de documents d'alerte pour la plateforme de surveillance du végétal PESV.

- la conception d'un catalogue de microorganismes probiotiques et leur effets anti-inflammatoires (MGP, appel Qualiment).
- la conception d'un catalogue exhaustif des microbes pathogènes pour l'homme et présents dans l'intestin (société MaatPharma).

Technologie

La combinaison d'outils génériques de détection et de normalisation d'entités avec peu ou pas d'exemples comme montré ici ouvre la voie à de nombreuses applications d'extraction d'information dans les domaines scientifiques et techniques.

4.3 Bilan sur les données ouvertes et la réutilisation

Logiciels

Les logiciels et applications décrits ici sont tous distribués sous licence libre sur GitHub. L'interopérabilité assurée par le format Alvis et l'encapsulation dans un container *docker* sur OpenMinTeD facilite leur réutilisation. Ces logiciels sont aisément trouvables grâce à l'ontologie OMTD-Share qui indexe les composants.

Problèmes de corpus.

Le verrou principal de l'application est l'accès aux textes complets souhaité par les utilisateurs.

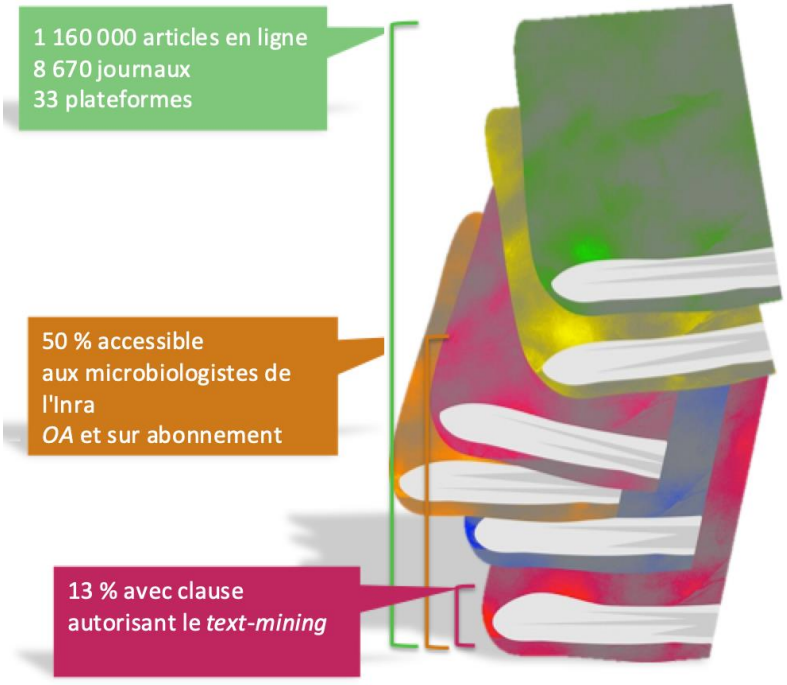


Figure 36. Composition du corpus et accès en 2016.

Seule la moitié des articles sont accessibles (libres ou sur abonnement INRA). Deux tiers environ des articles accessibles sont accessibles par des API sur des plateformes : PMC, ISTEX, Springer, Elsevier, Wiley. L'accès par l'API ne permet pas de télécharger l'ensemble des articles recherchés, ceci pour des raisons encore inexpliquées. Les autres articles ne sont accessibles

que par l'URL de la page de lecture dans des formats extrêmement variés, html, pdf, non documentés ou difficile à traiter pour l'extraction d'information (ex. pdf). Voir par exemple [Nédellec, 2016][20] sur la situation en 2016 pour plus de détails (Figure 37).

Flexibilité d'utilisation des ressources sémantiques

La version courante de la plateforme OpenMinTeD et sa connexion avec AgroPortal permet automatiquement de télécharger et de déclarer une nouvelle ressource sémantique (ontologie), mais pas de l'exploiter comme donnée externe d'une application. Elle doit être incluse dans le container Docker. Cette contrainte limite l'exploitation de notre application générique avec d'autres ressources pour des utilisateurs non spécialistes.

[20] https://docplayer.fr/64547867-Le-tdm-par-l-exemple-des-microbes-dans-mon-fromage.html

Généralisation du pilote dans un cadre de science ouverte

L'application pilote en microbiologie présente des propriétés générales qui permettent la réutilisation de ses technologies pour d'autres applications d'extraction d'information d'entités et de relation, et de normalisation. Le même workflow a par exemple été utilisé avec succès pour développer une application dans le domaine de la sélection du blé (*OpenMinTeD-D9.4 Application Software Release)*. La capacité de l'application à structurer les données pour les rattacher automatiquement à une ontologie permet l'intégration avec des données autres que textuelles dans un cadre de science ouverte, en particulier dans le domaine des sciences de la vie et de l'agriculture, domaines pilotes.

Les nouvelles techniques d'acquisition, d'échange et d'analyse à haut débit permettent aux chercheurs d'aborder des problématiques transversales et intégratives par la réutilisation de données produites à d'autres fins. L'approche modélisatrice apporte l'abstraction nécessaire à la réutilisation et à l'intégration de la connaissance et de données hétérogènes multi-échelles ainsi que des méthodes d'analyse dans des workflows flexibles. C'est cet objectif que permet d'atteindre la généralisation de l'application pilote pour le développement de services pour les microbes, plantes et animaux, adaptés à l'activité du chercheur, qui intègrent (1) données non structurées, expérimentales et analytiques et (2) services/traitement dirigés par les modèles de connaissances et ontologies suivant les principes FAIR.

Enjeux

L'intégration de données et de connaissance est au cœur de la bioinformatique pour la collecte, la curation, l'analyse et l'extraction de l'information. Le rôle des ontologies est essentiel comme référentiels pour désigner, représenter et lier les données et les processus.

Les connaissances disponibles sont extrêmement nombreuses (des millions dans les publications et les textes libres des bases de données) dans des domaines clefs des sciences de la vie (milieux, phénotypes des organismes, molécules produites, fonctions des gènes, réseaux de régulation). L'exploitation de cette connaissance pour interpréter, expliquer et comparer les résultats expérimentaux et analytiques requiert des méthodes de fouille de textes qui ont atteint aujourd'hui la maturité nécessaire à leur intégration pour des traitements à grande échelle comme illustré ici.

La diversité des questions d'analyse biologiques associées aux questions génotypiques, phénotypiques et fonctionnelles est telle qu'il est nécessaire de construire un cadre méthodologique et logiciel partagé. Le partage passe par la composition rapide de workflows d'analyse dédiés, à partir de composants logiciels facilement reconfigurables, et la réutilisation de ces workflows par les utilisateurs novices.

Forces et opportunités

Dans ce paragraphe, nous examinons les forces et opportunités de l'application dans un cadre de text mining et de Science Ouverte.

L'approche suivie par les équipes Migale, Bibliome et DIST de l'INRA peut servir d'exemple méthodologique et de fil conducteur pour la généralisation de l'approche en s'appuyant sur des preuves de concept faites dans le cadre d'applications en microbiologie alimentaire et en biologie des plantes.

Les besoins et bénéficiaires en Science de la vie sont bien identifiés (voir livrable D4.2 OpenMinTeD).

Les plateformes bioinformatiques, à l'interface de la recherche en biologie et de la recherche en bioinformatique et mathématique appliquée sont un lieu privilégié d'anticipation et de généralisation de solutions techniques avancées. La coordination grandissantes des structures de service existantes est une force, par exemple, celles des plateformes bioinformatiques INRA Migale, URGI et GenoToul dans le cadre du chantier de coordination et labellisation des infrastructures Inra, dont bioinformatique en lien avec le nouveau projet de l'IFB. La maturité technologique et organisationnelle sur l'interopérabilité, la conception de workflows et le partage de données locales et internationales est un atout pour généraliser la démarche du développement de l'application et la rationaliser.

Dans le domaine des sciences du vivant, la coordination Inra, nationale (IFB) et internationale (Elixir) sont des réseaux sur lesquelles s'appuyer. Les nombreux liens avec des infrastructures externes en fouille de textes (CLARIN), fourniture de contenu (OpenAire, etc.), calcul, sont à consolider et à étendre en bénéficiant de l'environnement EOSC.

La forte dynamique et structuration des communautés internationales en (1) text and data mining (TDM) appliqué au vivant (BioNLP) (2) en sémantique, modélisation et standardisation pour l'agriculture et l'alimentation, dont la plateforme H2020 OpenMinTeD sont des éléments porteurs.

Les problématiques "données"

Les sources primaires identifiées ici sont les données expérimentales et analytiques, les données textuelles (bases de données, littérature) et les ressources sémantiques (lexiques, ontologies). Les problématiques sont :

Seront à prendre en compte, la faible interopérabilité des données due au manque d'existence et d'adoption des standards, à une réutilisation limitée des ressources sémantiques et des modèles. Il sera nécessaire d'améliorer l'accessibilité et la réutilisabilité de ces ressources au sein de nos infrastructures.

Le raisonnement sur des données multi-échelles (de la molécule à l'environnement) et à des niveaux d'abstraction variés (de l'observation au trait phénotypique) nécessite la résolution de leur interopérabilité sémantique (formaliser leurs correspondances). Il faudra, à partir de cas concrets, aborder les questions de recherche sur les correspondances entre ontologies. C'est l'objectif du projet ANR D2KAB[21]

Des environnements homme-machine adaptées au travail et au domaine du chercheur devront être développés pour potentialiser les traitements sémantiques de manière

[21] http://www.d2kab.org

interactive (user-in-the-loop) et dynamique. L'association Intelligence Artificielle - bioinformatique est source d'innovation et l'EOSC apporte des moyens de renforcer les compétences et les technologies, par l'approche VRE notamment (ex AgInfra22).

Valeur ajoutée

Dans notre modèle d'organisation, la charge de la coordination et la rationalisation des traitements, dont la fouille de textes, et l'intégration des résultats sont transférées du chercheur à la plateforme de service, ce qui apporte gain de temps, concentration sur la question de recherche, meilleurs moyens de calcul, meilleure capacité à anticiper et à se mobiliser sur les fronts de science, pérennisation et mutualisation des workflows et résultats, reproductibilité, accompagnement technique et méthodologique…

Le chercheur accède et exploite plus efficacement les résultats de recherches complémentaires et comparables aux siens qui lui sont utiles, pour des approches systémiques et transversales.

Par la modélisation et l'adoption des principes FAIR implémentés ici, l'Inra augmente sa capacité à remobiliser et à exploiter la masse d'information non structurée, textuelles, sur des objets d'intérêt : génétique, phénotype, environnement, écosystèmes, conditions expérimentales, et leurs impacts sur la santé.

[22] https://aginfra.d4science.org/explore

Chaîne d'impact

Ressources (Input)	Activité (Activity)	Production (Output)	Effet (Outcome)	Impact (Impact)
données centralisées et curées données expérimentales locales et partagées textes ontologies et vocabulaires partagés modèles de connaissance e-infra text-mining et bioinformatiques pour l'accès et le traitement des données outils d'analyse et de visualisation des données standards	composition de chaînes de traitement adaptées alignement d'ontologies connexion d'e-infras développement d'environnements de travail intégrés recherche en bioinformatique et intelligence artificielle accompagnement, formation, transfert de compétences promotion de bonnes pratiques et de standards	outils d'analyse d'informations et de données hétérogènes répondant à des questions scientifiques outils d'analyse intégrés à l'environnement de travail du chercheur chaînes de traitement complexes, modulaires, adaptables et réutilisables ressources de connaissances (ontologies, vocabulaires) nouvelles compétences en fouille de textes et en intégration de données hétérogènes pour le développement d'applications avancées	nouvelles questions de recherche en biologie et en fouille de textes élargissement et enrichissement des ensembles de données pour la "data-driven science" valorisation des résultats de la recherche en fouille de textes augmentation du potentiel d'innovation des acteurs de la bioinformatique française et internationale réduction du coût de développement des chaînes de traitement complexes	contribution à une alimentation plus saine et plus sûre ainsi qu'au développement d'une agriculture avec de meilleures performances environnementales, sociales et sanitaires augmentation de la connaissance de l'origine et des effets des aliments sur la santé contrôle et conception de produit alimentaires innovants réduction des produits phytosanitaires dans les produits alimentaires par la sélection et le contrôle pathogène-hôte (plante-animaux)

Tableau 8. Chaîne d'impact de la généralisation de l'application pilote en science de la vie et agriculture

.

Références

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, Claire Nédellec. Bacteria Biotope at BioNLP Open Shared Tasks 2019 workshop, EMNLP-IJCNLP 2019, nov 2019.

Robert Bossy, Claire Nédellec, Julien Jourde, Mouhamadou Ba, Estelle Chaix, Louise Deléger. Bacteria Biotope Annotation Guidelines May 29, 2019 Bacteria Biotope Task at BioNLP-OST 2019

Robert Bossy, Claire Nédellec, Julien Jourde, Mouhamadou Ba, Estelle Chaix, Louise Deléger. Bacteria Biotope Annotation Guidelines May 29, 2019 Bacteria Biotope Task at BioNLP-OST 2019 https://drive.google.com/file/d/1G0po_xlRjQCZ-qxuA_4PLdipXU6rtYTp/view

Estelle Chaix, Louise Deléger, Robert Bossy, Claire Nédellec "Text mining tools for extracting information about microbial biodiversity in food" Food Microbiology, 2018. https://doi.org/10.1016/j.fm.2018.04.011

http://www.sciencedirect.com/science/article/pii/S0740002017310638

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, Claire Nédellec, Overview of the Bacteria Biotope Task at BioNLP Shared Task, In *Proceedings of the BioNLP Shared Task 2016 Workshop*, Association for Computational Linguistics, Berlin, Allemagne 2016. http://www.aclweb.org/anthology/W16-3002

Hélène Falentin, Stéphanie-Marie Deutsch,Valérie Gagnaire, Anne Thierry, Sandra Dérozier, Claire Nédellec. "Bioinformatics tools as a way to select microbial strains for fermented food products", *15th Symposium on Bacterial Genetics and Ecology "Ecosystem drivers in a changing planet", (BAGECO),* Lisbonne 27 mai 2019.

Nédellec C., Bossy R., Chaix E., Deléger L. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. In *Proceedings of the 4th International Microbial Diversity Conference*. pp. 221-227, ed. Marco Gobetti. Baris, Pub. Simtra. ISBN 978-88-943010-0-7, Bari, October 2017.

Nédellec C., "Le TDM par l'exemple : des microbes dans mon fromage ! "La loi numérique, et après ? Colloque de Meudon, 10 nov. 2016. https://docplayer.fr/64547867-Le-tdm-par-l-exemple-des-microbes-dans-mon-fromage.html

Index des figures

Index des tableaux