



# Approche itérative et analyse du besoin

Robert Bossy

ANF TDM 3 octobre 2024



# Questions

Quels outils choisir pour répondre à mon besoin ?

Comment adapter ces outils à mon problème ?

Comment savoir quelle technique est la plus adaptée :

- Extraction d'Information
- Recherche d'Information
- Classification ou clustering de documents
- Construire un corpus
- Résumé automatique
- Agent conversationnel
- ...



# Questions

Quel outil sera **le plus performant** pour répondre à mon besoin ?

Est-ce que mon besoin peut même être abordé étant donné l'état de l'art ?

Comment **revoir mon besoin** ?

Quels moyens déployer pour mettre en œuvre ces outils dans **ma pratique quotidienne** ?

Quel sera le **coût de mise en œuvre** ?



# Faites-le vous même

Démocratisation des outils permettant d'aborder rapidement les questions.

Possibilité d'obtenir des résultats satisfaisants.

Limites

- Caractérisation de la **qualité**.
- **Irréalisation du potentiel** de l'état de l'art.
- Valorisation, robustesse, passage à l'échelle.



# Réponses

Si, au premier abord, les réponses sont partielles et incomplètes.

→ Processus de co-construction interdisciplinaire entre :

- Spécialiste de Fouille de Texte : expertise dans l'un des domaines de la Fouille de Texte.
- Porteur du besoin : recherche dans domaine d'application qui present un besoin.



# Processus de co-construction interdisciplinaire

## Spécialiste en Fouille de Texte

Dégager de **nouveaux problèmes de recherche** et effectuer des travaux valorisables.

Acquérir des **compétences** thématiques, mobilisables à l'avenir.

## Porteur du besoin

Outils pour explorer et exploiter la littérature, **en grande quantité** avec une **qualité connue**.

**Perspective renouvelée** de sa propre discipline.



# Qui sommes nous ?

**Bibliome** : Extraction d'Information à partir de texte pour des domaines de spécialité d'intérêt pour **Inrae**.

1. **Recherche méthodologique** en Extraction d'Information et Traitement Automatique de la Langue.
2. **Application** dans les domaines de **Inrae** (biologie, agronomie).

Culture de **co-construction** avec des équipes thématiques.



# Qui sommes nous ? (Bibliome)

Position privilégiée pour faire l'**analyse du besoin**.

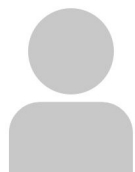
Quelques thèmes abordés :

- Biodiversité microbienne
- Santé des plantes
- Sélection génétique du blé
- Comptes rendus d'imagerie médicale
- Bien-être animal
- ...

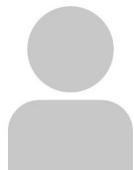




# Qui sommes nous ? (Bibliome)



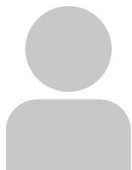
Anne-Sophie  
Foussat  
PhD



Louise  
Deléger  
CR



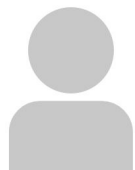
Robert  
Bossy  
IR



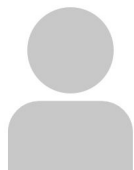
Mariya  
Borovikova  
PhD



Claire  
Nédellec  
DR



Xingyu  
Zhu  
PhD



Arnaud  
Ferré  
CR



Marine  
Courtin  
Post-doc

## Compétences complémentaires

- linguistique computationnelle
- représentation des connaissances
- apprentissage profond
- évaluation et expérimentation
- développement logiciel
- analyse du besoin
- disciplines thématiques



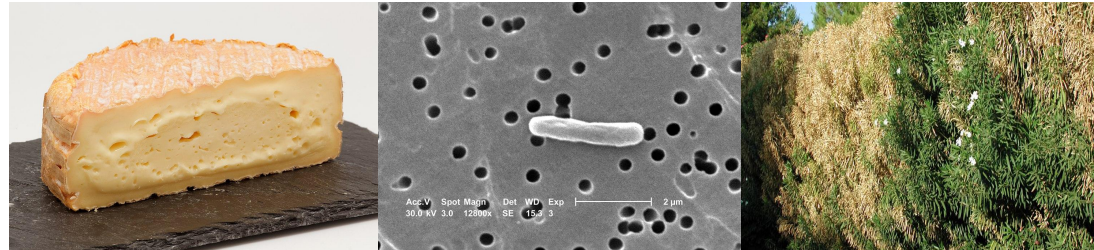
# Fil rouge : biodiversité microbienne

Microbes : petits, nombreux, divers, ubiquitaires.

Interagissent avec tous les organismes vivants : humains, animaux, plantes, autres microorganismes.

Leur présence et diversité est critique pour :

- la santé humaine, animale et végétale
- la production d'aliments
- le traitement des déchets



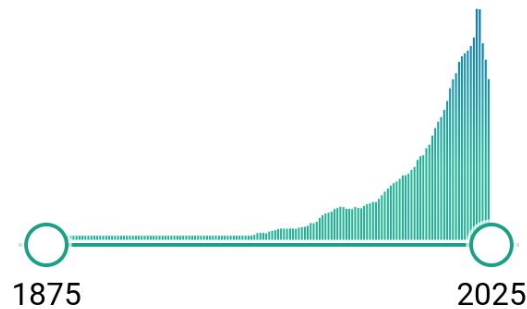
# Évolution de la recherche en microbiologie

Avant : chercheurs spécialisés dans une espèce ou genre.

Maintenant : étude des flores microbiennes.

Incapacité pour un chercheur d'accompagner toute la littérature.

RESULTS BY YEAR



# Processus de co-construction

1. Élicitation
  - a. Persona
  - b. Objet frontière
  - c. Prototypage
2. Artéfacts
  - a. Corpus annoté
  - b. Référentiels
  - c. Modèles de langue
  - d. Services
3. Deux conseils



# Élicitation



# Demande initiale

Formulation initiale du besoin :

- Tâches et techniques de Fouille de Texte.



# Demande initiale

Formulation initiale du besoin :

- ~~Tâches et techniques de Fouille de Texte.~~
- Verrous méthodologiques de sa propre discipline.



# Demande initiale

Formulation initiale du besoin :

- ~~Tâches et techniques de Fouille de Texte.~~
- Verrous méthodologiques de sa propre discipline.

**Élicitation** : donner les moyens au porteur du besoin de formuler son besoin en termes de Fouille de Texte.





# Élicitation et acculturation

## Spécialiste de Fouille de Texte

- Comprendre les verrous.
- Acquérir le vocabulaire et les concepts.
- Rechercher et montrer l'existant.
- Déterminer des verrous en Fouille de Texte.
- Généraliser.

## Porteur du besoin

- Reformuler son besoin.
- Formaliser les concepts nécessaires.
- Découvrir les possibilités de la Fouille de Texte.
- En comprendre les limites.
- Généraliser.



# Biodiversité microbienne : au début

*“Corrélation entre les protéines de surface et les conditions environnementales chez les bactéries vivant dans l’intestin de mammifères.”*

Maarten van de Guchte (2010)

Constat d’un déficit d’informations disponibles

- Incapacité à lister et distinguer les milieux liés à l’intestin.
- Absence de lexique et de modèle.
- Aucun recensement des bactéries présentes dans l’intestin.
- L’information est accessible exclusivement dans des articles ou des chartes de bases de données en texte libre.

gastro-intestinal  
epithelium  
gut antrum  
intestine  
human gut  
mucosa



# Outils pour l'élicitation

1. Persona
2. Objets-frontières (*boundary objects*)
3. Prototypage



# Persona



Utilisateur fictif (modèle) à partir duquel on construit des scénarios d'interaction avec un produit.

- L'utilisateur est un collaborateur.
- La fiction n'est que très rarement nécessaire.
- L'essentiel : se projeter sur des solutions concrètes et idéales.



# Avantages des *Persona*

Échanger sur des pratiques.

Appréhender l'échelle et la portée d'un besoin.

Recruter des collaborateurs thématiques, **généraliser**.

Orienter vers un domaine de la Fouille de Texte :

- Chercher des documents → Recherche d'Information
- Trier des documents → Classification
- Synthèse de connaissances en fiches ou graphes → Extraction d'Information
- Démarche pédagogique → Résumé automatique, Agent conversationnel



# Personas de la biodiversité microbienne

- Levures impliquées dans la fermentation de jus végétaux.
- Évolution de la diversité des champignons présents dans le sol.
- Symbiotes de la rhizosphère.
- Insectes vecteurs de pathogènes (humains ou plantes).

## Généralisation et délimitation

- Intestin → Tous les habitats possibles + phénotypes (propriétés).
- Bactéries → Tous les microorganismes.
- Protéines de surface → information déjà disponible dans des bases de données.

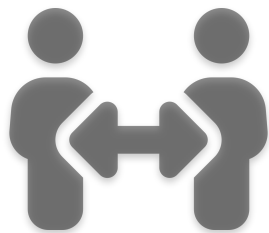


# Généralisation

- **Spectre** d'espèces connu dans un milieu donné.
- **Éventail** d'habitats colonisés par une espèce.
- **Ensemble** des milieux dans lesquels vivent des organismes ayant une propriété (phénotype) donnée.
  
- Requêtes structurées.
- **Démarche encyclopédique.**
- Donc **Extraction d'Information**, avec les entités :
  - Taxons
  - Habitats
  - Phénotypes



# Objet frontière



Objets facilitant la communication et la négociation entre plusieurs communautés.

- Nature : physique (spécimens), numérique, organisationnels (GdT).
- Suffisamment structuré pour être reconnu par l'ensemble des communautés.
- Suffisamment souple pour accommoder divers points de vue.





# Objet frontière : le corpus annoté

L'annotation de corpus consiste à marquer sur quelques textes les informations à extraire (dans l'idéal).

The screenshot displays a web browser window with a URL: [https://biblome.jouy.inra.fr/ydinars/ontobiotope2013/\[49547\]](https://biblome.jouy.inra.fr/ydinars/ontobiotope2013/[49547]). The main content area shows a text document titled "Evaluation of antibacterial activity of synthetic aliphatic and aromatic monoacylglycerols." The text describes the antibacterial activity of synthetic monoacylglycerols (MAGs) against two human pathogens, *Staphylococcus aureus* and *Escherichia coli*. The text is annotated with colored boxes: blue for "human pathogens", orange for "Staphylococcus aureus", and red for "Escherichia coli".

On the left side, there is a hierarchical tree view of the ontology. The selected path is: **human pathogen** (2) > **animal pathogen** (2) > **pathogen** > **human pathogen** (2) > **phytopathogen** (2) > **opportunistic pathogen**.

At the bottom, there is a table titled "Annotations | Text selection" with the following columns: Id, Annotation Set, K, Type, Details, and Vis.

Id	Annotation Set	K	Type	Details	Vis
a8576...	[imported] imported from review by arnaud.ba, claire.diale in campaign 10	📄	Bacteria	<i>Staphylococcus aureus</i>	🔍
424a9...	[imported] imported from review by arnaud.ba, claire.diale in campaign 10	📄	Habitat	human	🔍
9	[imported] imported				



# Objet frontière : le corpus annoté

L'**annotation de corpus** consiste à marquer sur quelques textes les informations à extraire (dans l'idéal).

- Entraînement et évaluation des systèmes d'Extraction d'Information (EI).
- Processus **collaboratif** auquel participent les porteurs du besoin.


## Avantages pour l'élicitation

- Support visuel.
- Permet de discuter à la fois du texte et des concepts.



# Objet frontière : le corpus annoté

presence of an opportunistic **psychrotrophic** LAB named ***Camobacterium maltaromaticum*** has been demonstrated in numerous **French cheeses**



- "opportunistic psychrotrophic" vs "psychrotrophic"
- "LAB" ? (= Lactic Acid Bacteria)
- "French cheeses" vs "cheeses"

- Familiarisation avec le vocabulaire.
- Expliciter les connaissances implicites.
- Préciser le besoin.
- Détecter les verrous en EI.



# Objet frontière : le référentiel

Un système **normatif** dans lequel les objets ou concepts d'une certaine catégorie sont **recensés**.

Exemples : nomenclature, taxonomie, terminologie, ontologie, catalogue, base de données.

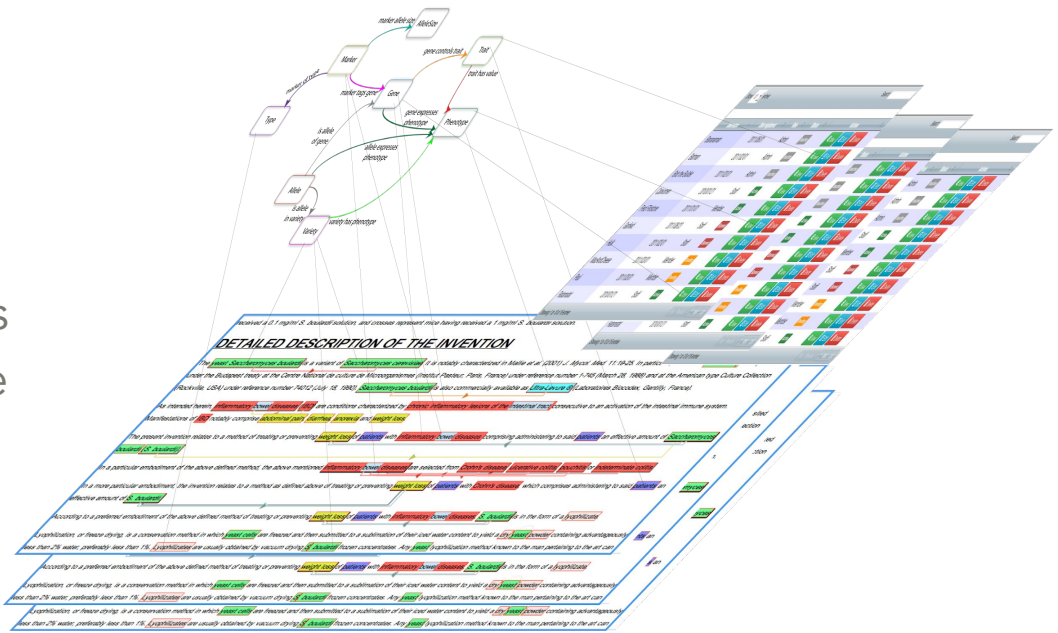
- Exposent le vocabulaire spécialisé.
- Maintenus par la communauté du porteur du besoin par un effort collectif.



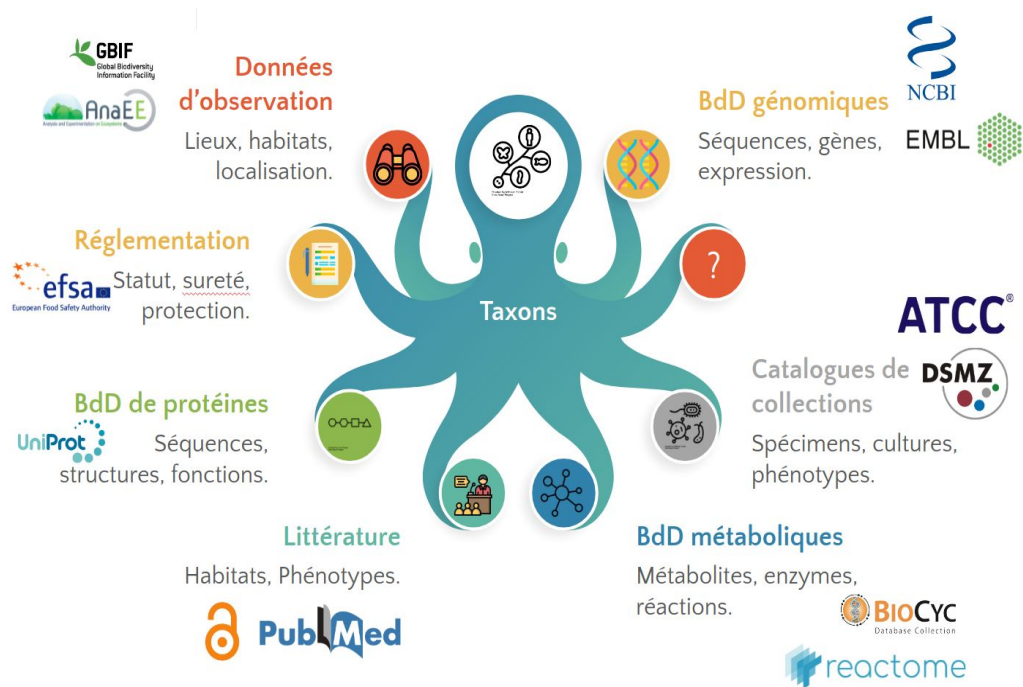
# Objet frontière : le référentiel

Un des objectifs de l'Extraction d'Information est l'**ancrage du texte aux référentiels**.

L'utilisation de référentiels partagés permet le croisement d'informations extraites du texte avec des bases de données.



# Objet frontière : le référentiel taxonomique



Un référentiel taxonomique recense et organise la diversité du vivant, d'après des hypothèses sur l'évolution des espèces.



# Objet frontière : le référentiel taxonomique

Entrez PubMed Nucleotide

Search for  as complet

Display 1 levels using filter: none

Nucleotide  Protein  Structure  Genome  Popset  
 Gene  HomoloGene  SRA Experiments  LinkOut  BLAST  
 BioProject  BioSample  Assembly  dbVar  Genetic Testing Register

[Lineage](#) (full): [cellular organisms](#)

◦ **Bacteria** (eubacteria) *Click on organism name to get more information.*

- [Acidobacteriota](#)
- [Aquificota](#)
- [Atribacterota](#)
- [Bdellovibrionota](#)
- [Caldisericota/Cryosericota group](#)
- [Calditrichota](#)
- [Campylobacterota](#)
- [Candidatus Deferrimicrobiota](#)
- [Candidatus Hinthialibacterota](#)
- [Candidatus Krumholzibacteriota](#)
- [Candidatus Lernaellota](#)
- [Candidatus Moduliflexota](#)
- [Candidatus Tharpellota](#)
- [Chrysiogenota](#)

[Browse by rank](#)  
[Advanced search](#)  
[Subscribe](#)  
[Main](#)  
[Introduction](#)  
[Navigation](#)  
[Nomenclature](#)  
[Etymology](#)  
[Collections](#)  
[Acknowledgements](#)

Search taxonomy

**Domain**

Close all phylums lists

**"Archaea"** [Hide phylum list](#)

- "Candidatus Aenigmataarchaeota"
- "Candidatus Aigarchaeota"
- "Candidatus Asgardarchaeota"
- "Candidatus Augarchaeota"
- "Candidatus Baldrarchaeota"
- "Candidatus Borrarchaeota"
- "Candidatus Diapherotrites"
- "Candidatus Freyrarchaeota"
- "Candidatus Hadarchaeota"
- "Candidatus Hadesarchaeota"
- "Candidatus Hermodarchaeota"
- "Candidatus Hodarchaeota"
- "Candidatus Huberarchaea"
- "Candidatus Huberarchaeota"
- "Candidatus Hydrothermarchaeota"
- "Candidatus Iainarchaeota"
- "Candidatus Kariarchaeota"

**"Bacteria"** [Hide phylum list](#)

- Abditibacteriota**
- "Candidatus Acetithermota"
- Acidobacteriota**
- Actinomycetota**
- "Candidatus Aerophobetes"
- "Candidatus Aerophobota"
- "Candidatus Altimarinita"
- "Candidatus Aminicenantes"
- "Candidatus Aminicenantota"
- Aquificota**

**TaxRef**

RÉFÉRENTIEL TAXONOMIQUE : FAUNE, FLORE ET FONGE DE FRANCE  
MÉTROPOLITAINE ET D'OUTRE-MER  
TAXREF V16.0

Données

- > Recherche de données
- > Données de synthèse sur les espèces
- > Données de synthèse sur les espaces

Référentiels

- > Référentiel taxonomique (TaxRef)
- > Base de connaissance « Statuts »
- > Référentiel habitats (HabRef)
- > Référentiel des organismes

**Méthodologie, sources et guide d'utilisation**

[Téléchargez la méthodologie de TAXREF v16.0 au format PDF](#)

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P. & Poncet, L. 2022. TAXREF, référentiel taxonomique, mise en œuvre et diffusion. Rapport PatrinNat (OFB-CNRS-MNHN), Muséum national d'Histoire naturelle, Paris. 47 pp.

**Télécharger l'intégralité de TAXREF**

Le fichier archive Zip comporte huit fichiers :

NCBI Taxonomy	NCBI
TaxRef	GBIF France
LPSN	DSMZ

# Objet frontière : le référentiel taxonomique

Critères pour choisir un référentiel taxonomique :

- Exhaustivité et précision
- Exhaustivité lexicale
- **Disponibilité** technique et légale
- **Adoption** et autorité
- Fréquence de mise à jour
- **Liens** établis avec des bases de données





# Objet frontière : le référentiel

S'approprier le vocabulaire et les conventions lexicales.

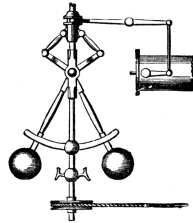
Délimiter les objets à extraire.

Comprendre les pratiques collectives et les enjeux scientifiques.

Prendre une position scientifique commune.



# Prototypage



Déploiement précoce des méthodes sur la base de l'état d'élicitation.

- Concrétiser partiellement les scénarios discutés.
- Détecter les acquis et les verrous en Fouille de Texte.
- Valider les référentiels, les documents et les besoins formulés.
- Faire évoluer le besoin.

# Prototypage : prérequis

- Ébauche d'analyse du besoin.
- Méthodes rapidement déployables
  - simples à implémenter
  - préexistantes
  - non-supervisées
  - frugales

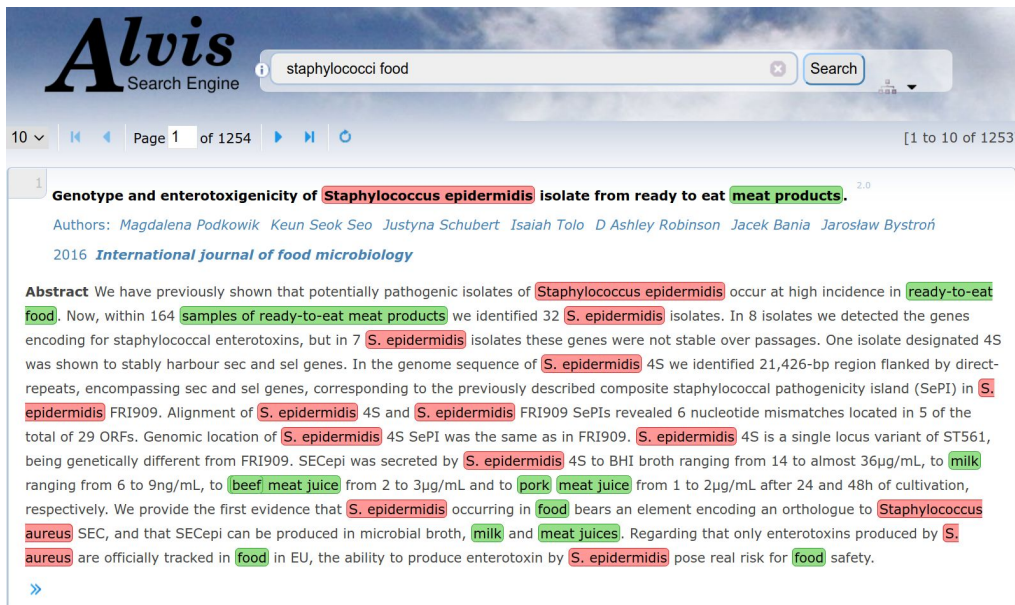


# Prototypage : tableau

	A	B	C	D	E	F	G
1	PMID	Offset	Microorganism	NCBI Tax	Offset	Habitat	OntoBioto
2	7524515	378-407	<i>Moloney murine leukemia virus</i>	ncbi:11801	278-292	several yeasts	OBT:002657
3	7524515	378-407	<i>Moloney murine leukemia virus</i>	ncbi:11801	331-336	avian	OBT:000010
4	7544909	122-129	viruses	ncbi:10239	0-10	Macrophage	OBT:002995
5	7544909	122-129	viruses	ncbi:10239	141-149	protozoa	OBT:000400
6	7544909	131-139	bacteria	ncbi:2	0-10	Macrophage	OBT:002995
7	7544909	131-139	bacteria	ncbi:2	141-149	protozoa	OBT:000400
8	7544909	154-159	fungi	ncbi:4751	0-10	Macrophage	OBT:002995
9	7544909	154-159	fungi	ncbi:4751	141-149	protozoa	OBT:000400
10	7504623	987-994	<i>S.pombe</i>	ncbi:4896	987-1000	<i>S.pombe</i> cells	OBT:000061
11	7504623	987-994	<i>S.pombe</i>	ncbi:4896	1092-1118	various cellular membranes	OBT:000061
12	7508994	415-418	GLV	ncbi:29255	481-497	propagated cells	OBT:000061
13	7508994	443-446	GLV	ncbi:29255	481-497	propagated cells	OBT:000061
14	7508994	621-631	<i>G. lamblia</i>	ncbi:5741	544-549	cells	OBT:000061
15	7508994	621-631	<i>G. lamblia</i>	ncbi:5741	559-573	culture medium	OBT:000007
16	7508994	681-684	GLV	ncbi:29255	652-660	WB cells	OBT:000061
17	7508994	681-684	GLV	ncbi:29255	826-831	cells	OBT:000061
18	7508994	788-791	GLV	ncbi:29255	652-660	WB cells	OBT:000061
19	7508994	788-791	GLV	ncbi:29255	826-831	cells	OBT:000061
20	7508994	985-992	viruses	ncbi:10239	946-951	cells	OBT:000061
21	7508994	1192-1213	<i>Tritrichomonas foetus</i>	ncbi:1144522	1164-1190	related parasitic protozoa	OBT:000400
22	7508994	1218-1239	<i>Trichomonas vaginalis</i>	ncbi:5722	1164-1190	related parasitic protozoa	OBT:000400
23	7508994	1263-1266	GLV	ncbi:29255	1164-1190	related parasitic protozoa	OBT:000400
24	7508994	681-684	GLV	ncbi:29255	1139-1144	cells	OBT:000061



# Prototypage : moteur de recherche sémantique



The screenshot shows the Alvis Search Engine interface. The search bar contains the text "staphylococci food". Below the search bar, there are navigation controls including "Page 1 of 1254" and "[1 to 10 of 1253]". The search results display a single entry with the following details:

- Genotype and enterotoxigenicity of *Staphylococcus epidermidis* isolate from ready to eat meat products.**
- Authors: Magdalena Podkowik, Keun Seok Seo, Justyna Schubert, Isaiah Tolo, D Ashley Robinson, Jacek Bania, Jaroslaw Bystroń
- 2016 *International journal of food microbiology*
- Abstract:** We have previously shown that potentially pathogenic isolates of *Staphylococcus epidermidis* occur at high incidence in ready-to-eat food. Now, within 164 samples of ready-to-eat meat products we identified 32 *S. epidermidis* isolates. In 8 isolates we detected the genes encoding for staphylococcal enterotoxins, but in 7 *S. epidermidis* isolates these genes were not stable over passages. One isolate designated 4S was shown to stably harbour sec and sel genes. In the genome sequence of *S. epidermidis* 4S we identified 21,426-bp region flanked by direct-repeats, encompassing sec and sel genes, corresponding to the previously described composite staphylococcal pathogenicity island (SePI) in *S. epidermidis* FRI909. Alignment of *S. epidermidis* 4S and *S. epidermidis* FRI909 SePIs revealed 6 nucleotide mismatches located in 5 of the total of 29 ORFs. Genomic location of *S. epidermidis* 4S SePI was the same as in FRI909. *S. epidermidis* 4S is a single locus variant of ST561, being genetically different from FRI909. SECEpi was secreted by *S. epidermidis* 4S to BHI broth ranging from 14 to almost 36µg/mL, to milk ranging from 6 to 9ng/mL, to beef meat juice from 2 to 3µg/mL and to pork meat juice from 1 to 2µg/mL after 24 and 48h of cultivation, respectively. We provide the first evidence that *S. epidermidis* occurring in food bears an element encoding an orthologue to *Staphylococcus aureus* SEC, and that SECEpi can be produced in microbial broth, milk and meat juices. Regarding that only enterotoxins produced by *S. aureus* are officially tracked in food in EU, the ability to produce enterotoxin by *S. epidermidis* pose real risk for food safety.

<https://bibliome.migale.inrae.fr/omicrobe/alvisir/webapi/search>

- Recherche dans le texte et dans les résultats de l'Extraction d'Information.
- Service familial.
- Prise de conscience des verrous en Fouille de Texte.
- Observation des améliorations lors de l'avancement du projet.



# Issues possibles de l'élicitation

1. Le besoin correspond à un état de l'art établi
  - Le porteur du besoin peut se débrouiller avec des moyens d'ingénierie.
2. Le besoin nécessite une adaptation ou une composition d'existants :
  - Ingénierie plus importante, nécessitant éventuellement une expertise.
3. Le besoin soulève des problématiques en Fouille de Texte :
  - Programme de recherche interdisciplinaire.
  - Quelques spécificités de la recherche dans les domaines de la Fouille de Texte.



# Corpus annoté



# Résultat de recherche : le corpus annoté

L'annotation de corpus consiste à marquer sur quelques textes les informations à extraire (dans l'idéal).

The screenshot shows a web browser window with a URL `https://biblome.jouy.inra.fr/ydinars/ontobiotope2013/ [49547]`. The main content area displays a text document titled "Evaluation of antibacterial activity of synthetic aliphatic and aromatic monoacylglycerols." The text contains several highlighted entities: "human pathogens", "Staphylococcus aureus", and "Escherichia coli". Colored arrows point from these highlights to a table of annotations at the bottom of the page.

**Annotations | Text selection**

Id	Annotation Set	K	Type	Details	Vis
	on micrococci, clostridia, streptococci in campaign 10				
a8576...	[imported] imported from review by arnaud.ba, claire.diale in campaign 10		Bacteria	Staphylococcus aureus	
424a9...	[imported] imported from review by arnaud.ba, claire.diale in campaign 10		Habitat	human	
9	[imported] imported				





# Résultat de recherche : le corpus annoté

L'**annotation de corpus** consiste à marquer sur quelques textes les informations à extraire (dans l'idéal).

- **Entraînement** et **évaluation** des systèmes d'Extraction d'Information (EI).
- Travail nécessaire si le besoin est nouveau.
- Processus **collaboratif**, intense.
- Méthodologies spécifiques.

Pour la biodiversité microbienne : corpus *Bacteria Biotopes*.



# Annotation de corpus : sélection de documents

## Représentativité

### Langue

- monolinguisme
- multilinguisme

### Genre

- articles scientifiques
- abstracts
- rapports médicaux
- news
- réseaux sociaux

### Temporalité

- documents récents
- passés
- historiques.

## Accès

### Technique

- requêtable
- téléchargeable
- reproductible

### Juridique

- droit de rediffusion
- propriété
- confidentialité

L'annotation est toujours la propriété des annotateurs, mais elle n'a que peu de sens sans le contenu.



# Annotation de corpus : guide d'annotation

Document de référence qui explicite ce qui doit être annoté :

- Rappel et explicitation du besoin.
- Définitions générales : délimitation et intentions.
- Cas particuliers et exemples.
- Document rédigé avant et amendé pendant l'annotation.

Élément essentiel de **reproductibilité**.



# Guide d'annotation : Bacteria Biotopes

presence of an opportunistic **psychrotrophic** LAB named ***Camobacterium maltaromaticum*** has been demonstrated in numerous **French cheeses**

## Bacteria Biotope Annotation Guidelines

Authors : Robert Bossy, Claire Nédellec, Julien Jourde, Mouhammadou Ba, Estelle Chaix, Louise Deléger

May 29, 2019

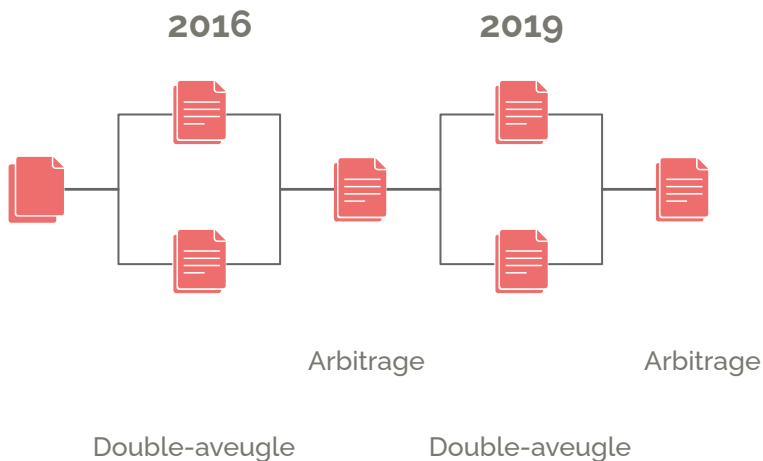
*Bacteria Biotope Task at BioNLP-OST 2019*

### Contents

<b>0 Note</b>	<b>3</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Copyright and License	3
1.2 Conventions	3
<b>2 Microbial taxon names</b>	<b>4</b>
2.1 Entity domain	4
2.1.0 Microorganism definition	4
2.1.1 Gram staining	4
2.1.2 Abbreviations	5
2.1.3 Lactic acid bacteria	5
2.1.4 Too general	5
2.2 Boundaries	6
2.2.1 Phenotype acronyms designating microorganisms	7
2.2.2 Strain specification	7
2.2.3 Nomenclatural suffixes: sp., spp., gen. nov., sp.nov.	8
2.3 Taxon ID	9
2.3.1 Unknown taxon identifier	9
2.3.2 Partial coreference	9
<b>3 Habitat mentions</b>	<b>10</b>
2.1 Entity domain	10



# Annotation de corpus : double annotation



Chaque document est annoté par deux experts + troisième passe d'arbitrage.

Multiplier les points de vues et les interprétations possibles.

Accord inter-annotateur : mesure de convergence des interprétations.

- Difficulté de la tâche.
- Qualité du guide d'annotation.

	REN ( $F_1$ )	NEN (S)	ER ( $F_1$ )
<b>Accord inter-annotateur</b>	.89	.97	.79



# Annotation de corpus : logiciel d'annotation

The screenshot displays the AlvisAE web-based annotation editor. The main window shows a text document titled "Evaluation of antibacterial activity of synthetic aliphatic and aromatic monoacylglycerols." The text contains several annotations: "human pathogens" (blue box), "Staphylococcus aureus" (orange box), and "Escherichia coli" (orange box). A table at the bottom lists these annotations with their IDs, sets, and types.

**Annotations Table:**

Id	Annotation Set	K	Type	Details	Vis
	arnaud,ba,claire,diale in campaign 10				
a8576...	[imported] imported from review by arnaud,ba,claire,diale in campaign 10		Bacteria	Staphylococcus aureus	
424a9...	[imported] imported from review by arnaud,ba,claire,diale in campaign 10		Habitat	human	
	[imported] imported				

# Corpus annoté : effort

Documents	392	Abstracts PubMed + extraits full-text
Tokens	60.402	
Entités	7.232	
Relations	3.578	
Annotateurs	18	
ETP	~2 ans	

**Valorisation ?**



# Corpus annoté : valorisation par un *challenge*

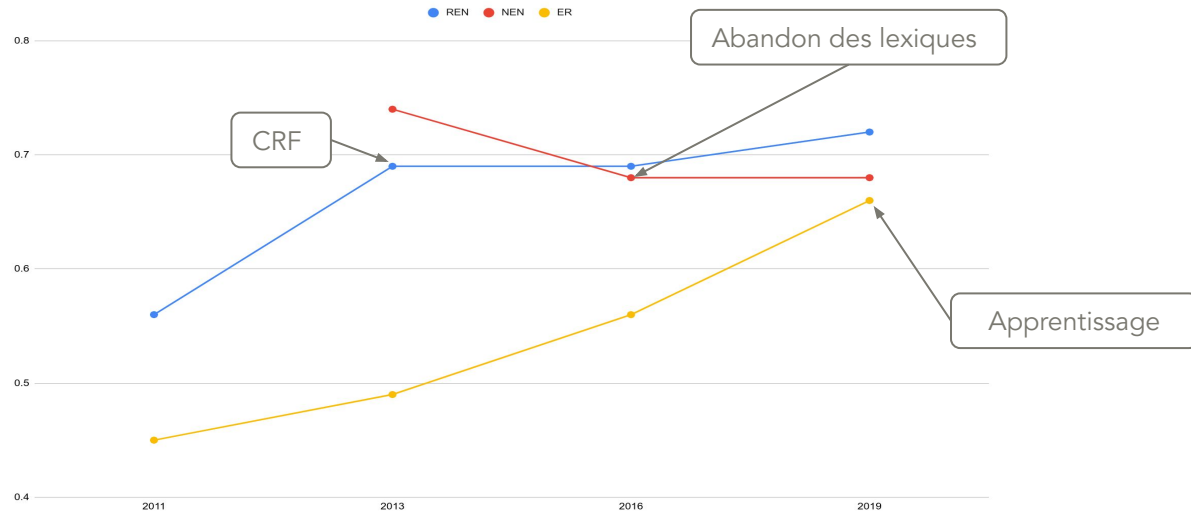
Lors d'un *challenge*, on invite l'ensemble de la communauté d'Extraction d'Information de travailler sur le corpus annoté.

- **Calendrier** de quelques mois : phases d'entraînement, de test et d'évaluation.
- **Restitution** à une conférence ou workshop : reconnu et cité.
- Publication cosignée par les porteurs du besoin.
- Le **besoin devient une tâche reconnue** en Fouille de Texte : x10 participants, x100 citations.





# Évolution des méthodes et performances



Lexiques et patrons  
Peu de ML (SVM, MEMM)

Analyse grammaticale  
kNN, CRF, SVM

Abandon des lexiques  
CRF, SVM, NN

SVM+LR, biLSTM, CNN,  
Transformers

# Autres artéfacts



## Référentiels

Si le besoin est un front de science pour le porteur du besoin, alors on peut être amené à construire son propre référentiel (lexiques, terminologies, ontologies).

- Construction et maintenance **collaborative**.
- Participation à l'effort de normalisation de la communauté du porteur du besoin.
- Valorisation dans la communauté du porteur du besoin.

Biodiversité microbienne : ontologie **OntoBiotope** des habitats et phénotypes microbiens.



# Modèles de langue

Modèle statistique ou probabiliste [réseaux de neurones profonds] du texte.

- Capture des structures linguistiques (et sémantiques).
- Réutilisables dans diverses tâches de Fouille de Texte.
- Résultats d'un calcul assez intensif.
- Exemples : BERT, GPT-4.

La mise à disposition d'un modèle de langue est très facilement valorisable.



# Services



# Services

Portail, interface, base de données permettant l'exploration des résultats de la Fouille de Texte sur une grande quantité de documents.

The screenshot displays the Alvis search engine interface. At the top, there is a search bar with the query 'staphylococci food'. Below the search bar, a table lists search results with columns for PMID, Offset, Microorganism, NCBI Tax, Offset, Habitat, and OntoBioto. The results include entries for *Moloney murine leukemia virus*, *viruses*, *bacteria*, *fungi*, and *S.pombe*.

The main content area shows a search result for 'staphylococci food' from the *International Journal of food microbiology*. The abstract text is as follows:

**Genotype and enterotoxigenicity of *Staphylococcus epidermidis* isolate from ready-to-eat meat products**  
 Authors: Magdalena Podkowiak, Keun Seok Seo, Justyna Schubert, Isaijah Tolo, D Ashley Robinson, Jacek Banik  
 2016 *International Journal of food microbiology*

**Abstract** We have previously shown that potentially pathogenic isolates of *Staphylococcus epidermidis* occur at high incidence in ready-to-eat food. Now, within 164 samples of ready-to-eat meat products we identified 32 *S. epidermidis* isolates. In 8 isolates we detected the genes encoding for staphylococcal enterotoxins, but in 7 *S. epidermidis* isolates these genes were not stable over passages. One isolate designated 4S was shown to stably harbour sec and sel genes. In the genome sequence of *S. epidermidis* 4S we identified 21,426-bp region flanked by direct-repeats, encompassing sec and sel genes, corresponding to the previously described composite staphylococcal pathogenicity island (SePI) in *S. epidermidis* FR1909. Alignment of *S. epidermidis* 4S and *S. epidermidis* FR1909 SePIs revealed 6 nucleotide mismatches located in 5 of the total of 29 ORFs. Genomic location of *S. epidermidis* 4S SePI was the same as in FR1909. *S. epidermidis* 4S is a single locus variant of ST561, being genetically different from FR1909. SECEPI was secreted by *S. epidermidis* 4S to BHI broth ranging from 14 to almost 36µg/mL, to milk ranging from 6 to 9ng/mL, to beef meat juice from 2 to 3µg/mL and to pork meat juice from 1 to 2µg/mL after 24 and 48h of cultivation, respectively. We provide the first evidence that *S. epidermidis* occurring in food bears an element encoding an orthologue to *Staphylococcus aureus* SEC, and that SECEPI can be produced in microbial broth, milk, and meat juice. Regarding that only enterotoxins produced by *S. aureus* are officially tracked in food in EU, the ability to produce enterotoxin by *S. epidermidis* pose real risk for food safety.

On the right side of the interface, there is a 'Habitat contains Taxon' section with a search bar and a table of results. The table has columns for Source text, Habitat, Relation type, Taxon, GPS, and Source. The results include entries for 'Cheedar: Cheedar', 'soft cheese: soft Hispanic type cheese', 'cheese: surface - ripened cheese, soft', 'cottage cheese: cottage cheese', 'semi soft cheese: semi - hard cheese', 'soft cheese: soft cheese', 'Toma: La Toma', and 'ripened cheese: short - ripened acid - curd cheese'.

At the bottom right, there is a heatmap visualization showing the distribution of data across different categories, with a color scale from blue (low) to red (high).

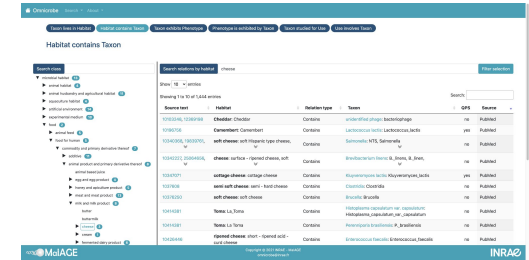


# Services : Omnicrobe

Portail de référence sur la biodiversité microbienne qui permet d'explorer les connaissances sur les habitats et phénotypes microbiens.

Conditions additionnelles :

- Constitution du corpus (PubMed microbio)
- Frugalité numérique (3M de résumés)
- Maintenance et mise à jour (continuité de service, corpus, état de l'art)
- Évaluation centrée sur l'utilisation



<https://omnicrobe.migale.inrae.fr/>



# Services : troisième partenaire

Participation d'un tiers spécialisé dans le déploiement et la maintenance de services.

- Migale (Plateforme INRAE) : [Micro]biologie, bioinformatique.
- IFB (Investissement d'Avenir) : biologie et bioinformatique.
- Huma-Num (TGIR) : SHS
- CorTexT (Plateforme INRAE + IFRIS) : Bibliométrie
- INIST (UAR) : Sciences





# Si vous avez un besoin

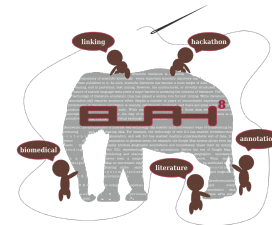
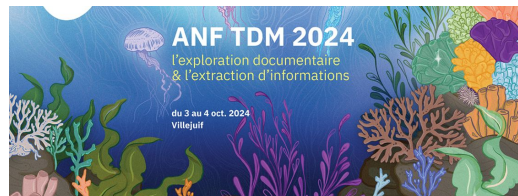




# 1. Débrouillez-vous !



- **Démocratisation des outils** de Fouille de Texte et beaucoup de ressources en Open Access (bibliothèques, tutoriels, modèles, jeux de données).
- **Événements** pour s'initier (formations) ou approfondir et échanger (hackathons) :
  - Ateliers de l'ANF-TDM
  - JDEV (DevLog)
  - BioHackathon
  - BLAH



## 2. Collaborez

- Formulation et révision de votre besoin.
- Orientation vers des ressources existantes.
- Programme de recherche.
  - Ouverture et généralisation du besoin.
  - Création d'artéfacts (corpus annoté, référentiel, modèle de langue).
  - Valorisation : conférences, *challenges*.
- Déploiement de service
  - Partenaire spécialisé.



**Merci de votre attention**