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 pressent un besoin.

Processus de co-construction interdisciplinaire

Spécialiste en Fouille de Texte

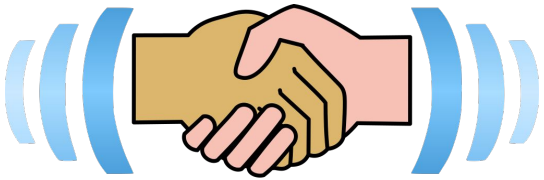
Dégager de nouveaux problèmes de recherche et effectuer des travaux valorisables.

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.

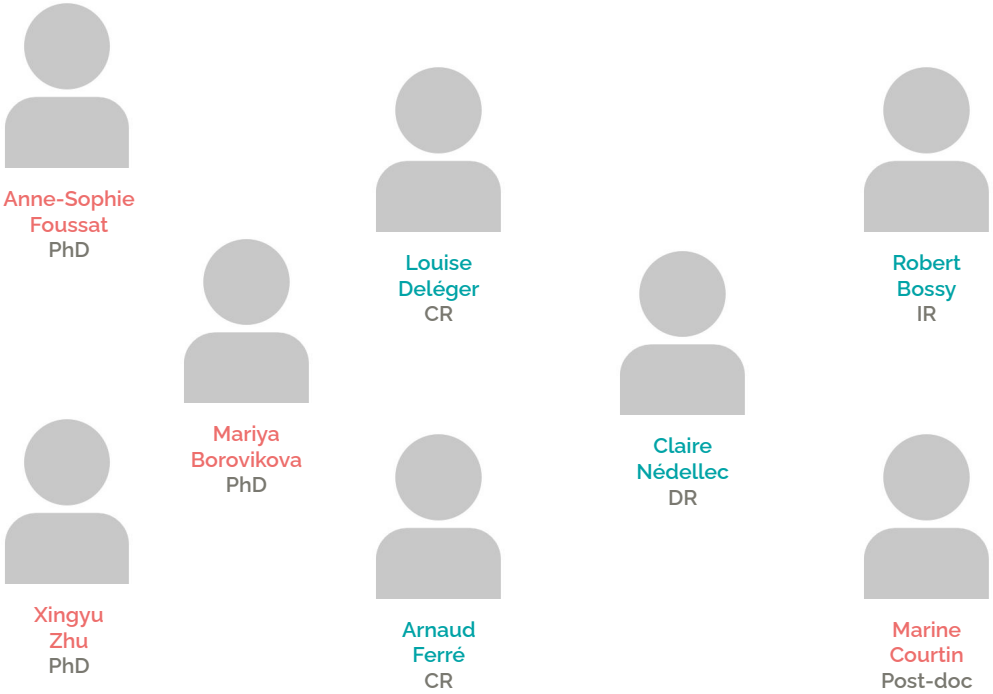
Qui sommes nous ? (Bibliome)

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)



Compétences complémentaires

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

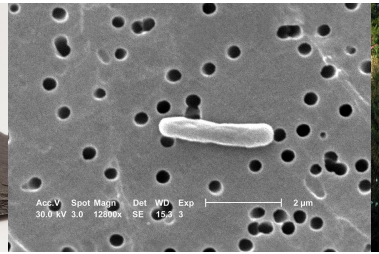
Fil rouge : biodiversité microbienne

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

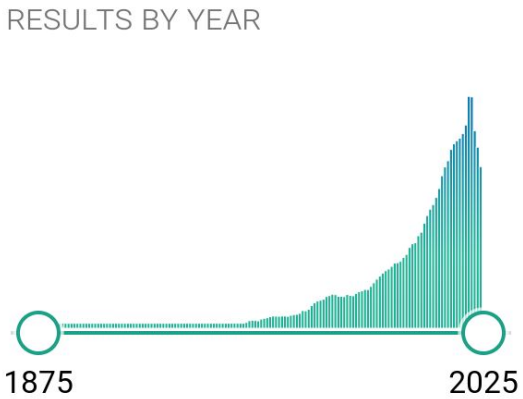


Évolution de la recherche en microbiologie

Avant : chercheurs spécialisés dans une espèce ou genre.

Maintenant : étude des flores microbiennes.

Incapacité pour un chercheur d'accompagner toute la littérature.



Capture : NCBI

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 char de bases de données en texte libre.

Outils pour l'élicitation

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

Kim Goodwin. 2009. Designing for the Digital Age: How to Create Human-Centered Products and Services. ISBN:978-0470229101.

Persona

Image : PUBLIC DOMAIN i4p4n

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

Susan Leigh Star and James R. Griesemer. 2023. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology. DOI:10.1177/030631289019003001

Objet frontière

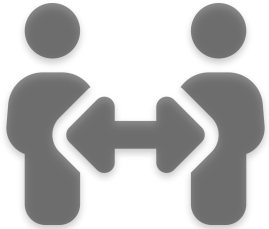


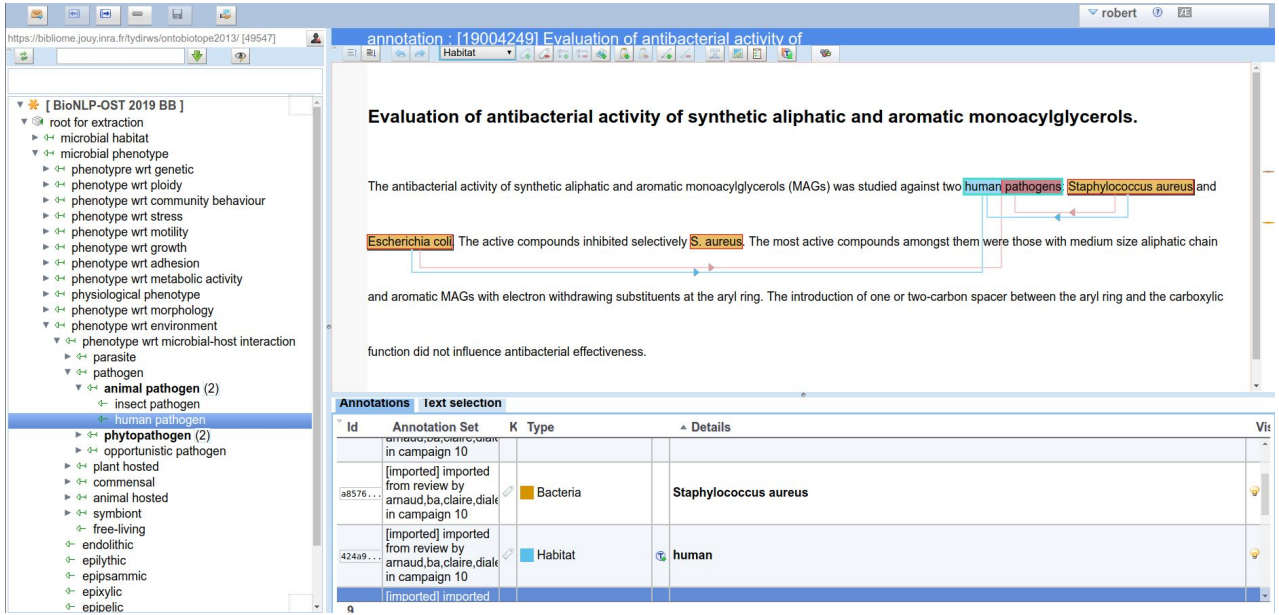
Image : Font Awesome

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).



Objet frontière : le corpus annoté

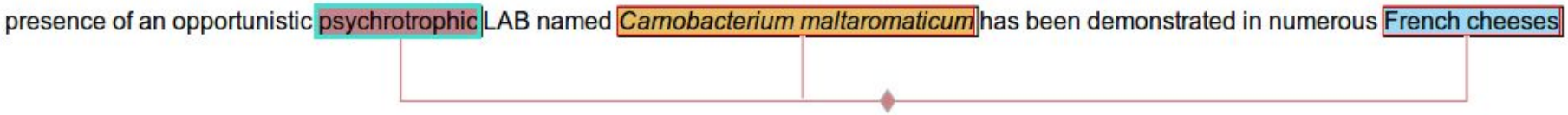
L'annotation de corpus consiste à marquer sur quelques textes les informations à extraire (dans l'idéal).

- 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



- “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 recensées.

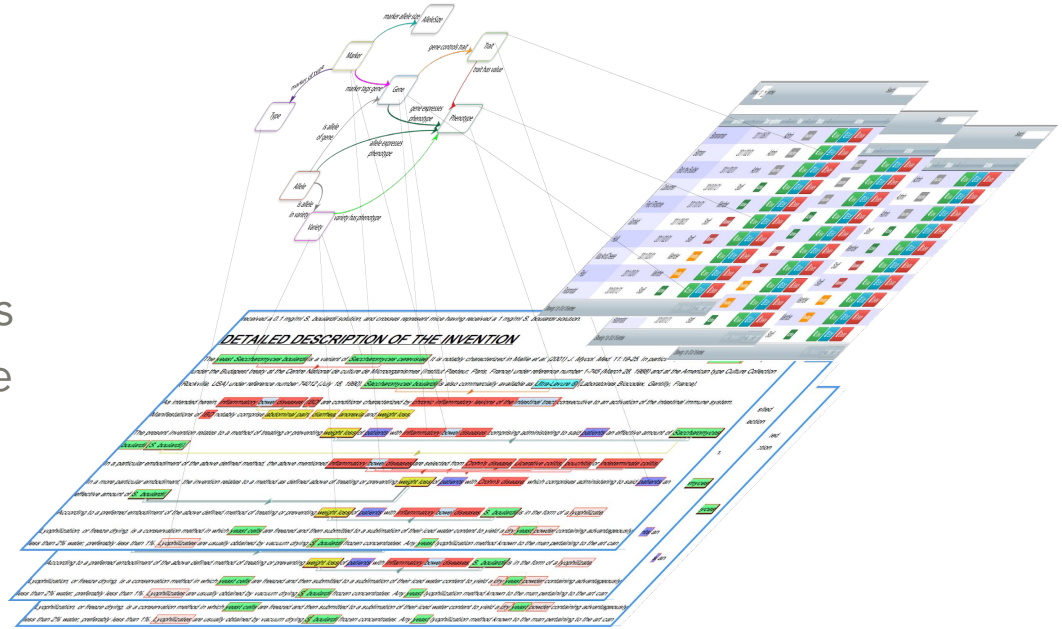
Exemples : nomenclature, taxonomie, terminologie, ontologie, catalogue, base de données.

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

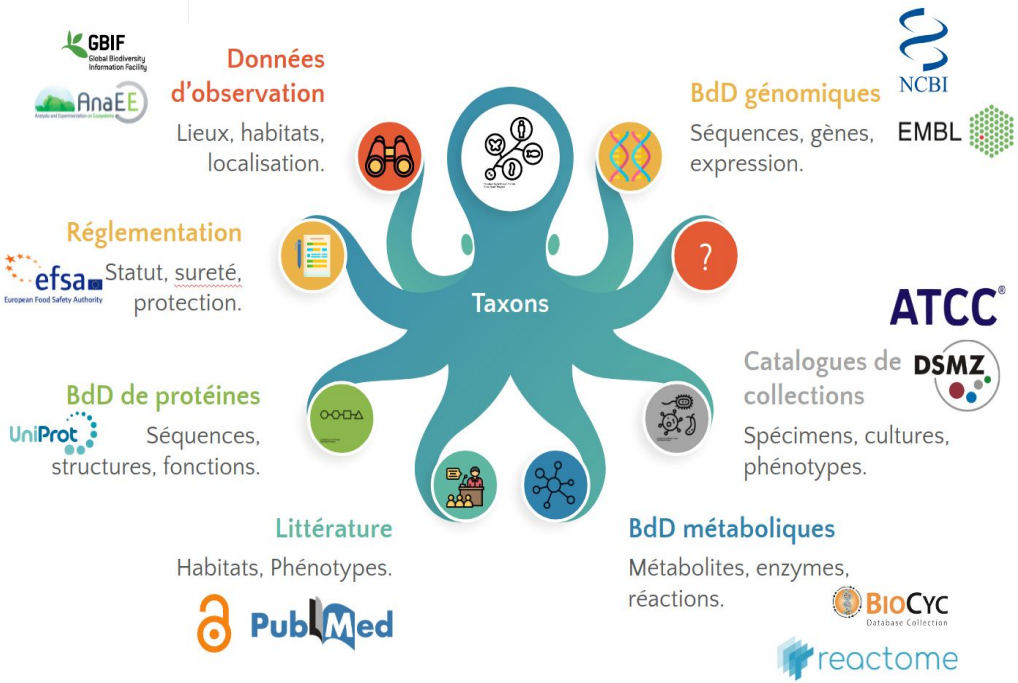
Objet frontière : le référentiel

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.



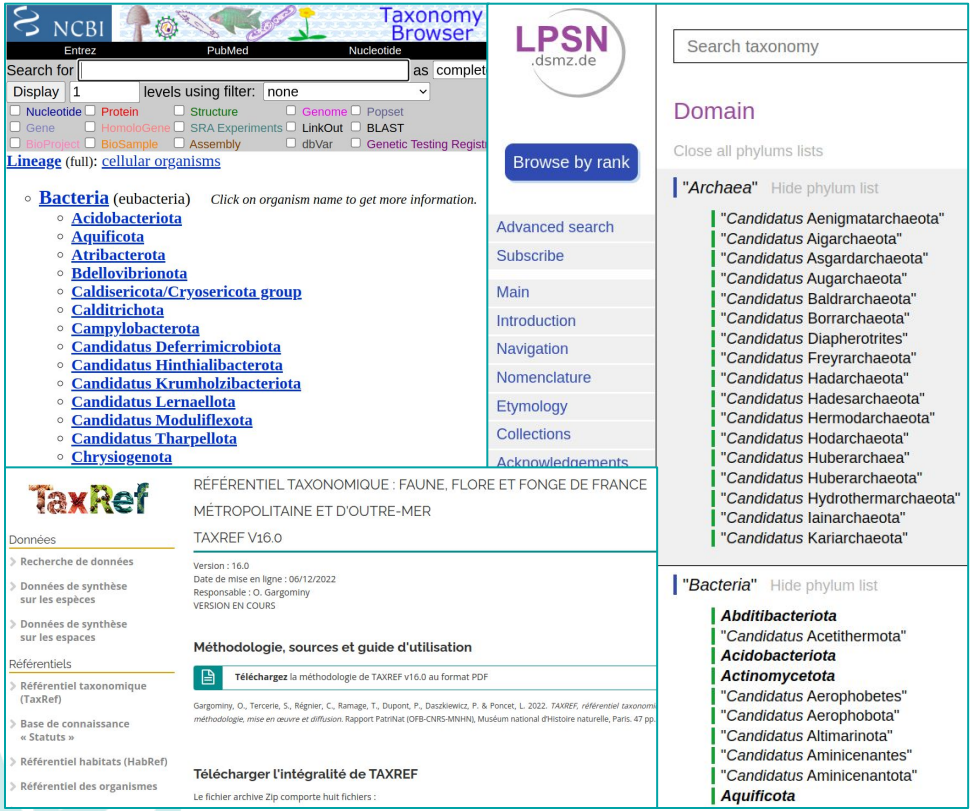
Objet frontière : le référentiel taxonomique



Un référentiel taxonomique recense et organise la diversité du vivant,

d'après des hypothèses sur l'évolution des espèces.

Objet frontière : le référentiel taxonomique



NCBI Taxonomy	NCBI
TaxRef	GBIF France
LPSN	DSMZ

Captures : NCBI, INPN, 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

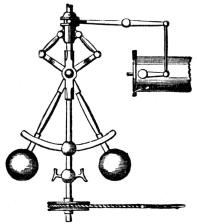


Image : Store Norske Leksikon

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.

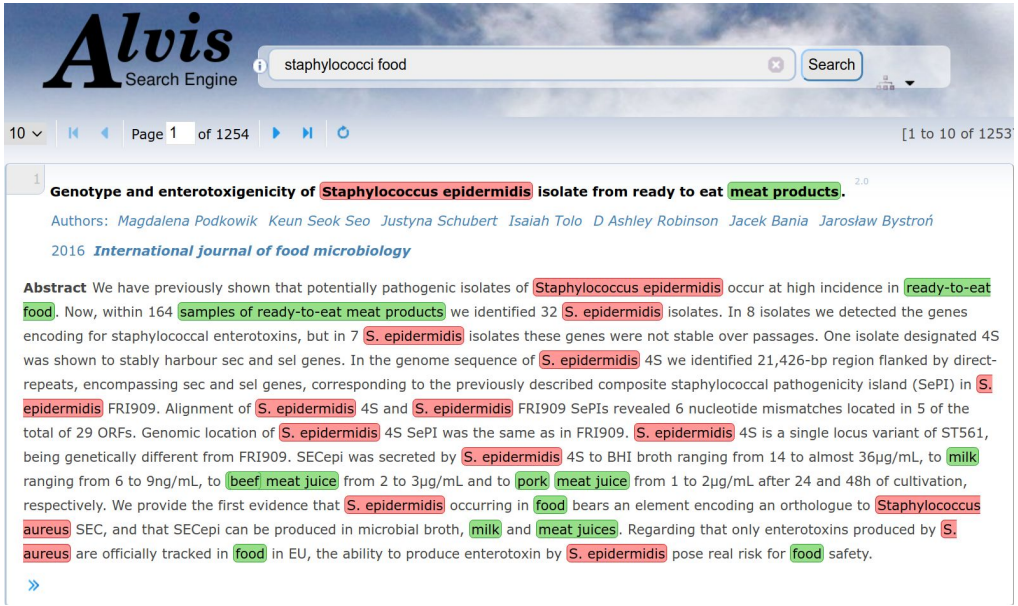
Prototypage : prérequis

- 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	*Moloney murine leukemia virus*	ncbi:11801	278-292	*several yeasts*	OBT:002657
3	7524515	378-407	*Moloney murine leukemia virus*	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	*S.pombe*	ncbi:4896	987-1000	*S.pombe cells*	OBT:000061
11	7504623	987-994	*S.pombe*	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	*G. lamblia*	ncbi:5741	544-549	*cells*	OBT:000061
15	7508994	621-631	*G. lamblia*	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	*Tritrichomonas foetus*	ncbi:1144522	1164-1190	*related parasitic protozoa*	OBT:000400
22	7508994	1218-1239	*Trichomonas vaginalis*	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



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

https://bibliome.migale.inrae.fr/omnicrobe/alvisir/webapi/search

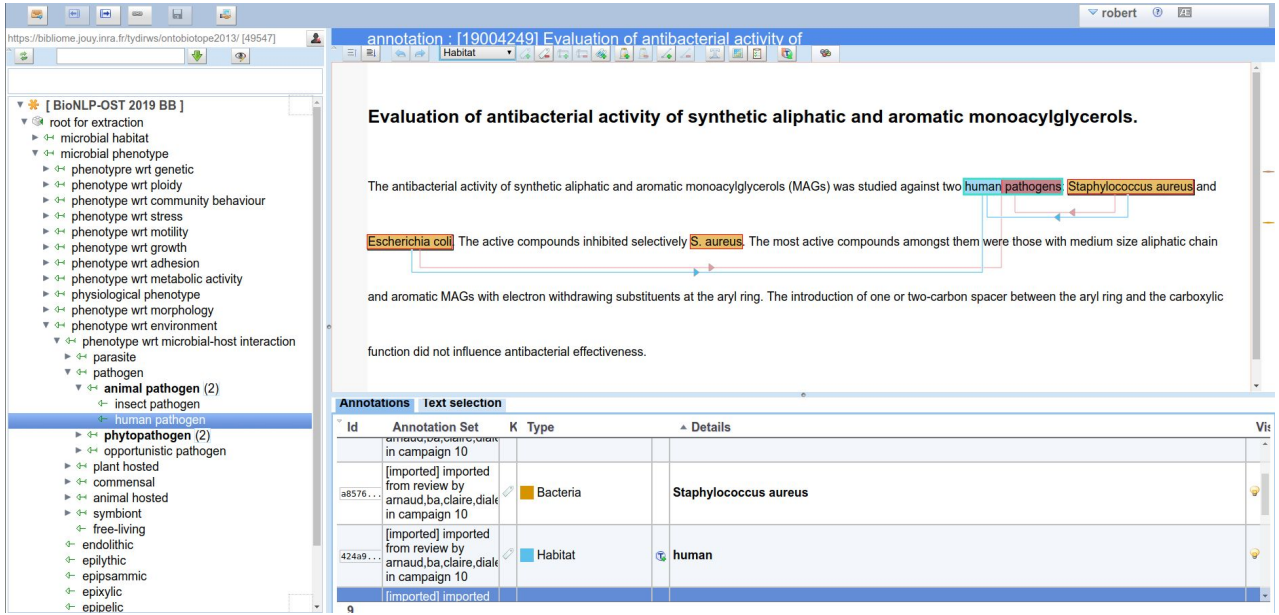
Issues possibles de l'élicitation

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).



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

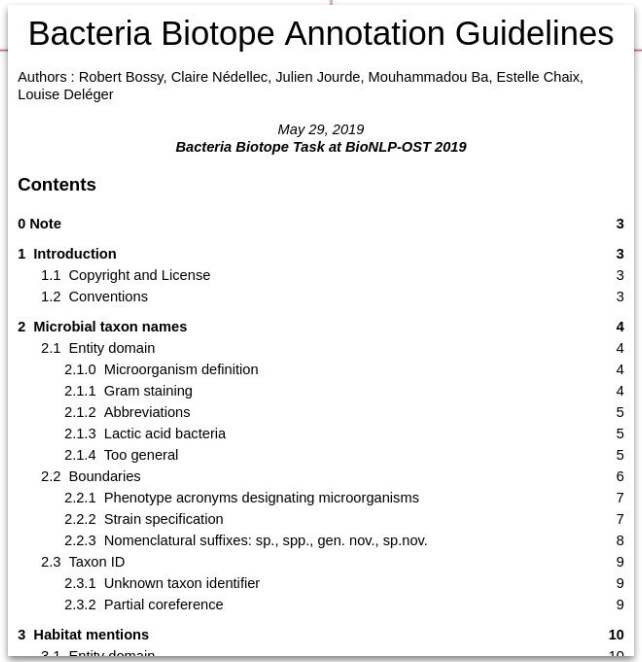
Document de référence qui explicite ce qui doit être annoté :

- 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 *Carnobacterium maltaromaticum* has been demonstrated in numerous French cheeses



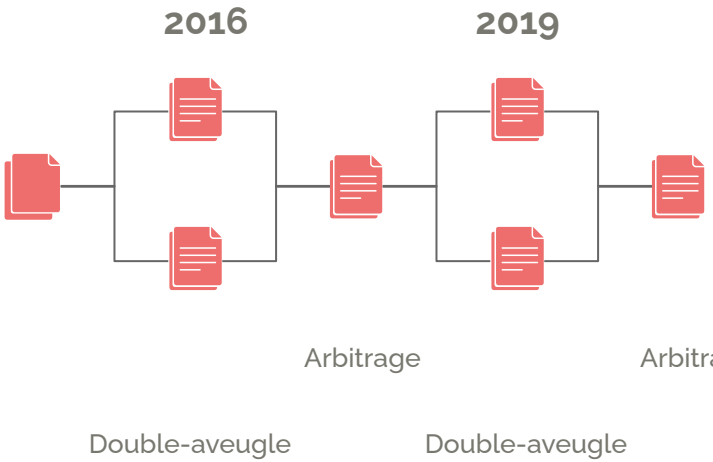
Bacteria Biotope Annotation Guidelines

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

May 29, 2019
Bacteria Biotope Task at BioNLP-OST 2019

Contents

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.

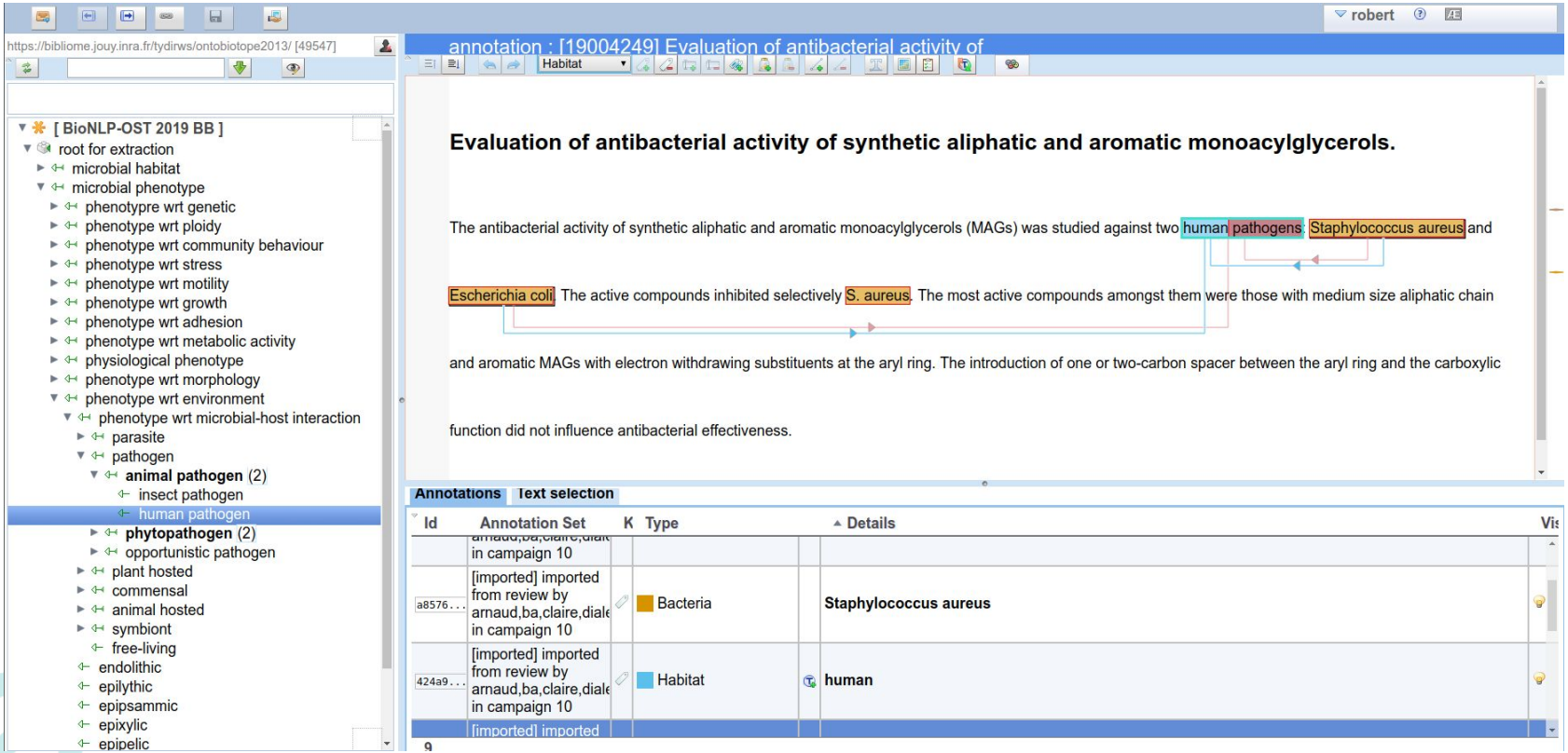
Accord inter-annotateur : mesure de convergence des interprétations.

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

	REN (F_1)	NEN (S)	ER (F_1)
Accord inter-annotateur	.89	.97	.79

Frederic Papazian, Robert Bossy, Claire Nédellec. 2012. AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. https://hal.inrae.fr/hal-02748212v1

Annotation de corpus : logiciel d'annotation



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 ?

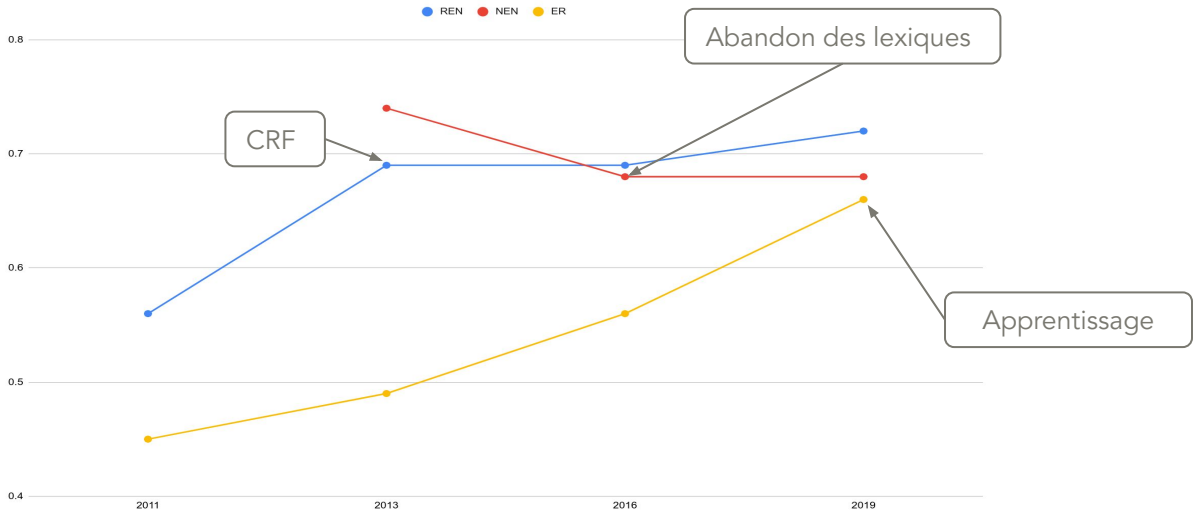
Robert Bossy, Julien Jourde, Philippe Bessières, Maarten Van De Guchte, Claire Nédellec. 2011. BioNLP shared task 2011 - Bacteria Biotope.

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.

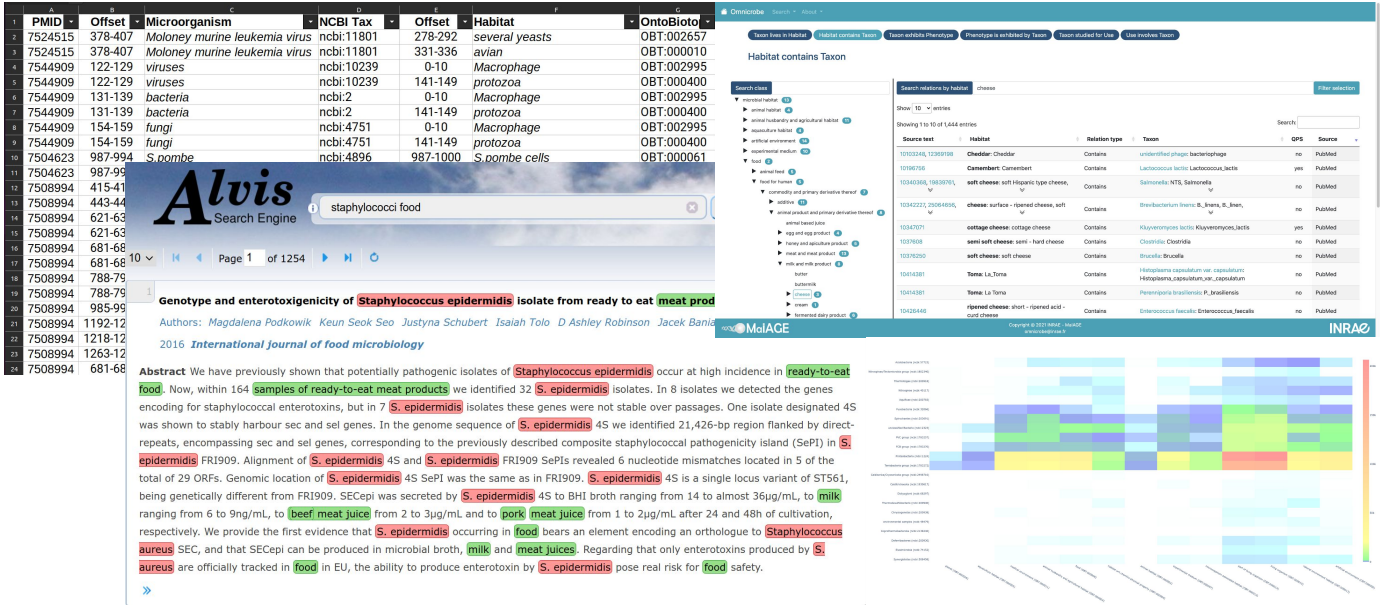
- Capture des structures linguistiques (et sémantiques).
- Réutilisables dans diverses tâches de Fouille de Texte.
- Résultats d'un calcul assez intensif.
- Exemples : BERT, GPT-4.

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

Services

Services

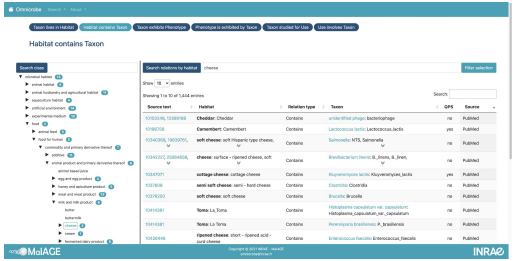
Portail, interface, base de données permettant l'exploration des résultats de la Fouille de Texte sur une grande quantité de documents.



Services : Omnicrobe

Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Valentin Loux, Hélène Falentin, Claire Nédellec. 2023. Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. https://dx.doi.org/10.1371/journal.pone.0272473

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



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

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

Services : troisième partenaire

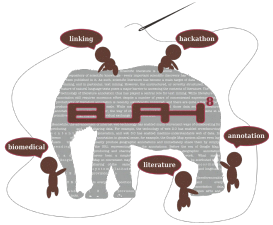
Participation d'un tiers spécialisé dans le déploiement et la maintenance de services.

- 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