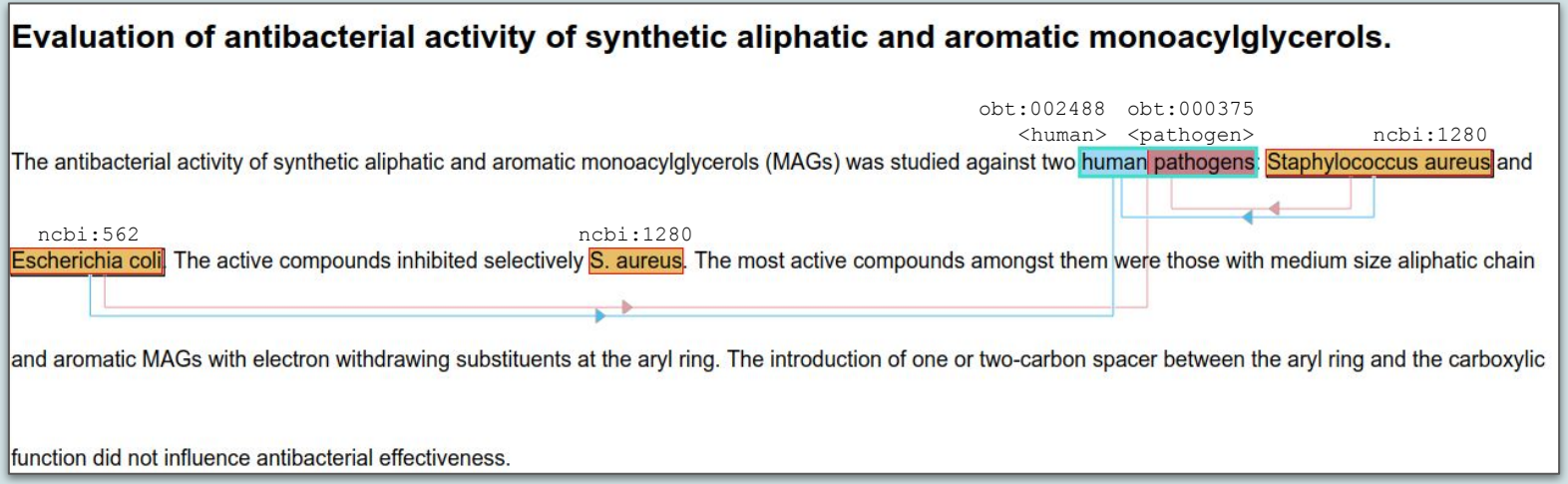
Extraction d'Informations pour la biodiversité microbienne

Robert Bossy - DigitBio 2023-01-12

Informations extraites sur la biodiversité microbienne



Entités Nommées (REN)

- Taxons microbiens
- Habitats
- Phénotypes

Normalisation (NEN)

- Taxonomie
- Ontologie des habitats et phénotypes

Relations (ER)

- *<Vit dans>*
- *<Présente le phénotype>*

Motivations

- Recherche en microbiologie : taxon singulier → études transversales.
- L'expertise sur tous les taxons et tous les milieux de vie est impossible.
- L'information est accessible exclusivement dans des articles ou des champs de bases de données en texte libre.

Image: NCBI 2023

Questions

- Spectre d'*espèces* connu dans un milieu donné.
- Éventail d'habitats colonisés par un *taxon*.
- Ensemble d'organismes qui vivent dans des *milieux similaires* à ceux d'une *souche* donnée.

Contenu de cette présentation

- Cheminement qui *nous* a amené à investir ce sujet.
- Ressources qui ont été nécessaires pour lever les verrous.

1. Analyse du besoin
2. Acquisition des ressources
3. Organisation de challenges
4. Développement de services
5. Conclusion

Analyse du besoin

Demande initiale

Corrélation entre les protéines de surface et les conditions environnementales chez les bactéries vivant dans l'intestin de mammifères.

Marteen 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.

Généralisation & délimitation

Levures impliquées dans la *fermentation* de *jus végétaux*.

Évolution de la diversité des *champignons* présents dans un *sol*.

Symbiotes de la *rhizosphère*.

Insectes vecteurs de *pathogènes* (*humains* ou *plantes*).

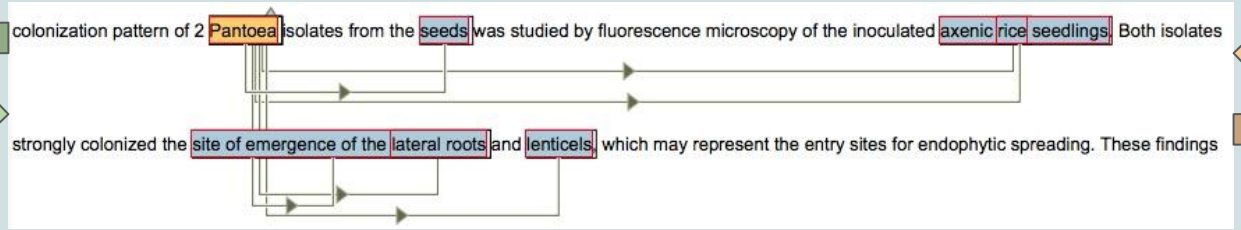
- Intestin → Tous les habitats possibles + phénotypes.
- Bactéries → Tous les microorganismes.
- Protéines de surface → information déjà disponible dans des bases de données.

Méthodologie

Itérations successives d'analyse partagée de textes.

Le texte comme objet de frontière (*boundary object*).

Microbiologistes



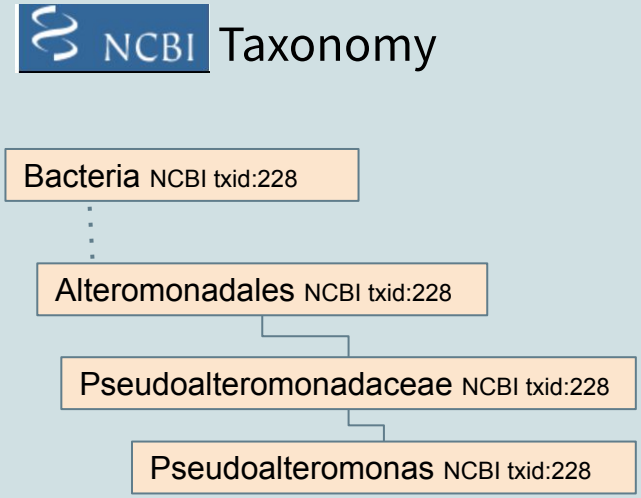
"Nous"

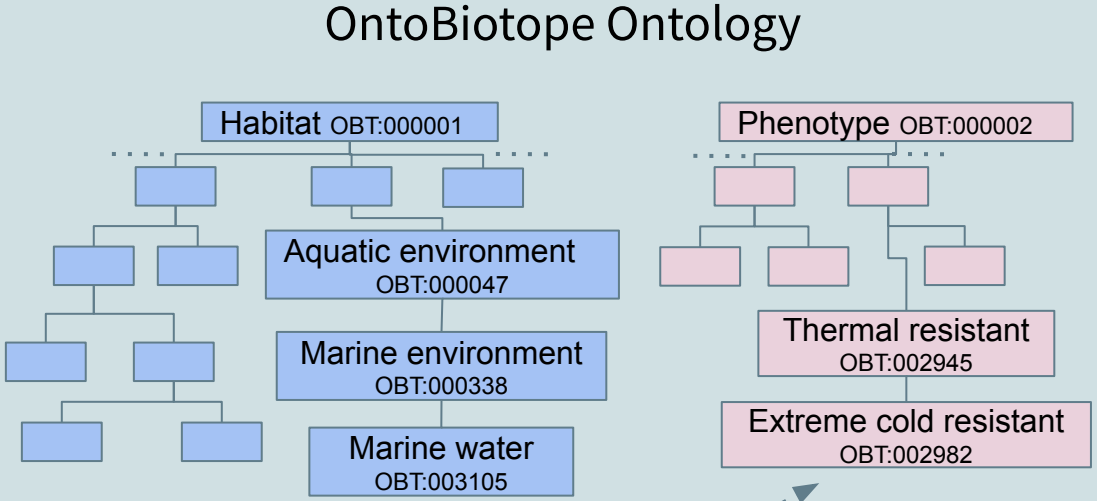
- Formalisation de la définition des entités et relations.
- Affinement de la portée des entités.
- Identifier les verrous en extraction d'information.

Star & Griesemer, 1989.

Acquisition de ressources

Ressource sémantique





Pseudoalteromonas is known to have many cold-adapted enzymes that function in the polar seawater

Image: Deléger, 2021

Référentiel partagé

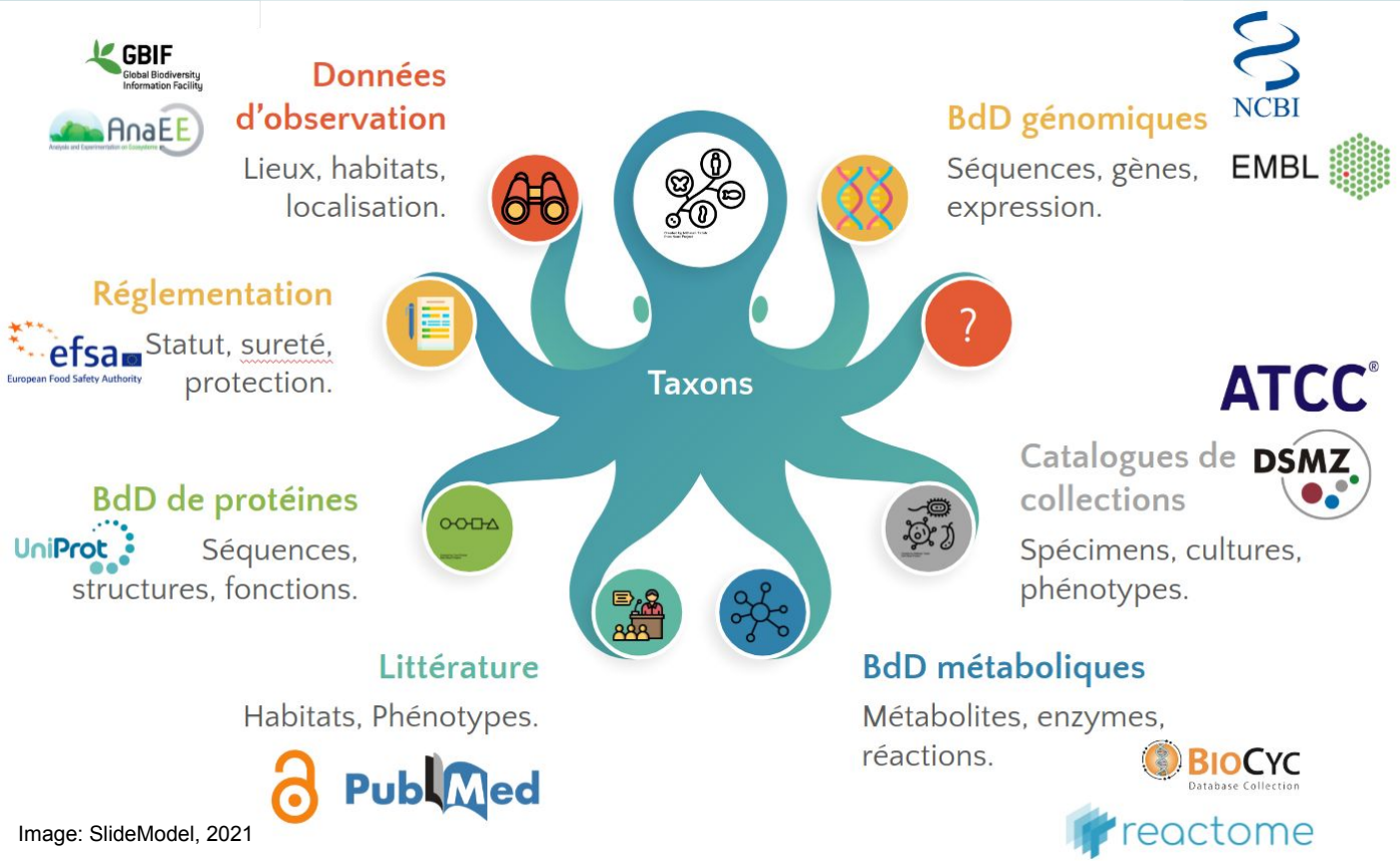


Image: SlideModel, 2021

Critères de réutilisation de ressources existantes

- Exhaustivité sémantique
  - Est-ce que toutes les branches taxonomiques sont représentées ?
- Précision
  - Quels rangs taxonomiques (espèce, souche) ?
- Exhaustivité lexicale
  - Synonymes et noms vernaculaires inclus ?
- Niveau de formalisme
  - Présence de lignées taxonomiques ? Distinction des types de synonymie ?
- Disponibilité technique et légale
  - Téléchargement ou API ? Licence ?
- Adoption et autorité
  - Liens établis avec des BdD ? Fréquence de mise à jour ?

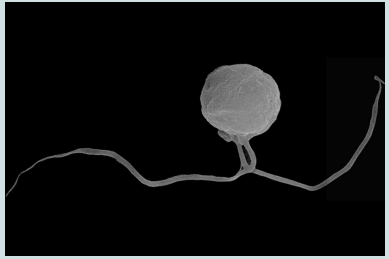
Taxonomie

- NCBI Taxonomy, taxonomie de référence pour :
  - les bases de données du NCBI (GenBank),
  - mais aussi de l'EBI,
  - et UniProt.
- Alternatives
  - GBIF Taxonomic Backbone : référence taxonomique du GBIF.
  - LPSN : référence taxonomique pour *Bacteria* et *Archaea*.
  - ICTV : référence taxonomique pour les virus.
  - …

https://www.ncbi.nlm.nih.gov/taxonomy

Schoch *et al.*, 2020.

Image: Tikhonenkov *et al.*, 2022



ncbi:3004206 <Provora>

Réutilisation de NCBI Taxonomy

- Exhaustivité sémantique
  - Toutes les branches microbiennes sont représentées.
- Précision
  - Rangs sub-spécifiques, incluant souches, pathovars, etc. *Exhaustivité limitée*.
- Exhaustivité lexicale
  - Synonymes inclus, avec et sans autorité. Quelques noms vernaculaires.
- Niveau de formalisme
  - Taxons hiérarchisés. Distinction de différents types de synonymes.
- Disponibilité technique et légale
  - Téléchargement. Licence libre.
- Adoption et autorité
  - Source secondaire, politique de curation peu claire. Largement liée dans différentes BdD.

NCBI Taxonomy : sélection des microorganismes

- Aucune définition intrinsèque satisfaisante de “microorganismes”.
- Approche communautaire
  - Quels chercheurs se reconnaissent dans “microbiologiste” ?
  - OntoBiotope (Réseau métaprogramme MEM).
- Volume
  - ~ 1M taxons
  - ~ 2,5M noms

Taxon	ID
Alveolata	33630
Amoebozoa	554915
Archaea	2157
Bacteria	2
Chlamydomonadales	3042
Chlorella	3071
Choanoflagellida	28009
Cryptophyta	3027
Desmidiales	131210
Diplomonadida	5738
Euglenozoa	33682
Fungi	4751
Glaucocystophyceae	38254
Haptophyta	2830
Ichthyosporea	127916
Nematoda	6231
Oxymonadida	66288
Parabasalia	5719
Prototheca	3110
Retortamonadidae	193075
Rhizaria	543769
Stramenopiles	33634
Viruses	10239

Combinaison de ressources

- Souches dans NCBI Taxonomy : complétude très inégale.
- L'information sur les habitats et phénotypes au niveau de la souche est essentielle.
- Aligner le catalogue du DSMZ :
  - Collection de cultures de référence.
  - Catalogue disponible en ligne.
- Incrément
  - 80k taxons
  - 700k synonymes

Raimer *et al.*, 2022.

Alignement NCBI et DSMZ

```
<strains>
  <list-item>
    <domain>Bacteria</domain>
    <phylum>Proteobacteria</phylum>
    <class>Deltaproteobacteria</class>
    <ordo>Myxococcales</ordo>
    <family>Anaeromyxobacteraceae</family>
    <genus>Anaeromyxobacter</genus>
    <species>Anaeromyxobacter dehalogenans</species>
    <species_epithet>dehalogenans</species_epithet>
    <subspecies_epithet/>
    <full_scientific_name>Anaeromyxobacter
dehalogenans Sanford et al.
2002</full_scientific_name>
    <designation>FRC-D3</designation>
    <variant/>
    <is_type_strain>False</is_type_strain>
    <ID_reference>16578</ID_reference>
  </list-item>
</strains>
```

Image: Joint Genome Institute

```
161493  | Anaeromyxobacter dehalogenans Sanford et al. 2002 | authority |
161493 | Anaeromyxobacter dehalogenans | scientific name |
```

- Nouveau noeud `bd:16578`
- Parent : `ncbi:161493`
- Nom : “*Anaeromyxobacter dehalogenans FRC-D3*”

Alignement NCBI et DSMZ

```
<strains>
  <list-item>
    <domain>Bacteria</domain>
    <phylum>Actinobacteria</phylum>
    <class>Actinobacteria</class>
    <ordo>Actinomycetales</ordo>
    <family>Corynebacteriaceae</family>
    <genus>Corynebacterium</genus>
    <species>Corynebacterium timonense</species>
    <species_epithet>timonense</species_epithet>
    <subspecies_epithet/>
    <full_scientific_name>Corynebacterium
timonense Merhej et al.
2009</full_scientific_name>
    <designation>5401744, CSUR P20</designation>
    <variant/>
    <is_type_strain>True</is_type_strain>
    <ID_reference>16812</ID_reference>
  </list-item>
</strains>
```

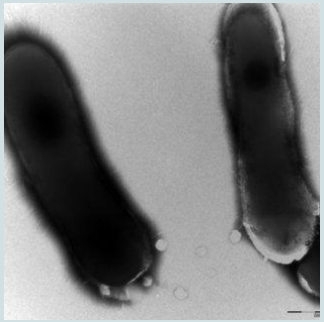


Image: Roux, *et al.*, 2014

```
1203190 | Corynebacterium timonense 5401744        | scientific name |
1203190 | Corynebacterium timonense str. 5401744   | equivalent name |
1203190 | Corynebacterium timonense strain 5401744 | equivalent name |
```

- Équivalence entre `bd:16812` et `ncbi:1203190`
- Nouveau synonyme : “*Corynebacterium timonense CSUR P20*”

Ressource sémantique : habitats et phénotypes

- EnvO : ontologie d'écosystèmes et environnements.
- "*The Environment Ontology (ENVO) is an expressive, machine-actionable knowledge representation of environmental entities.*"

https://sites.google.com/site/environmentontology

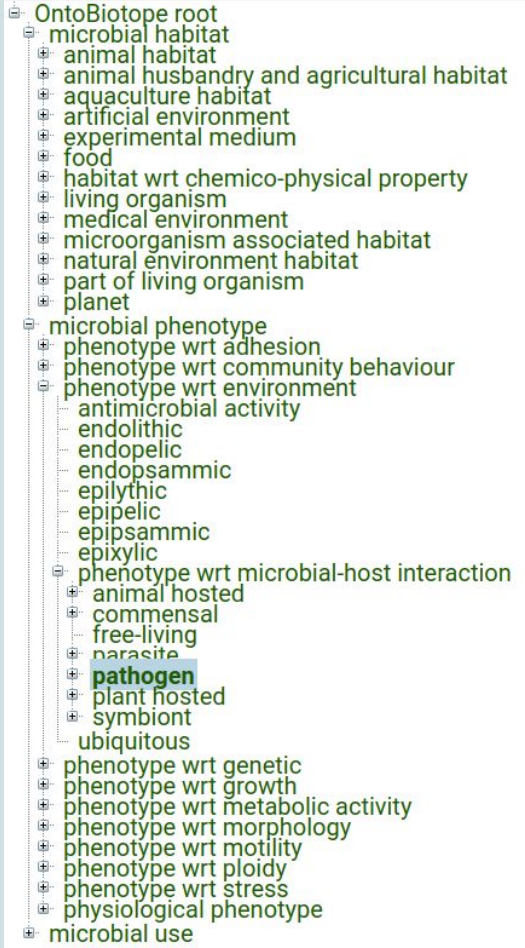
Buttigieg *et al.*, 2013.

Réutilisation de EnvO

- Exhaustivité sémantique
  - Habitats peu adaptés aux microorganismes. Aucun phénotype.
- Précision
  - Très variable.
- Exhaustivité lexicale
  - Étiquettes auto-suffisantes, mais peu de lexicalisation.
- Niveau de formalisme
  - Ontologie formelle.
- Disponibilité technique et légale
  - Téléchargeable. CC0.
- Adoption et autorité
  - Vocabulaire contrôlé de MIxS du Genomic Standards Consortium (adoption effective limitée).

Habitats et phénotypes : OntoBiotope

- Ontologie des habitats et phénotypes microbiens.
- Organisation des habitats selon les propriétés physiques, chimiques et physiologiques d'intérêt pour l'adaptation des microorganismes.
- Adoption:
  - Challenge Bacteria Biotopes.
  - Base de données Omnicrobe.
  - Vocabulaire contrôlé pour Open16S (projet pilote MICA).
- Volume:
  -



https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/1.4382640528105164E12

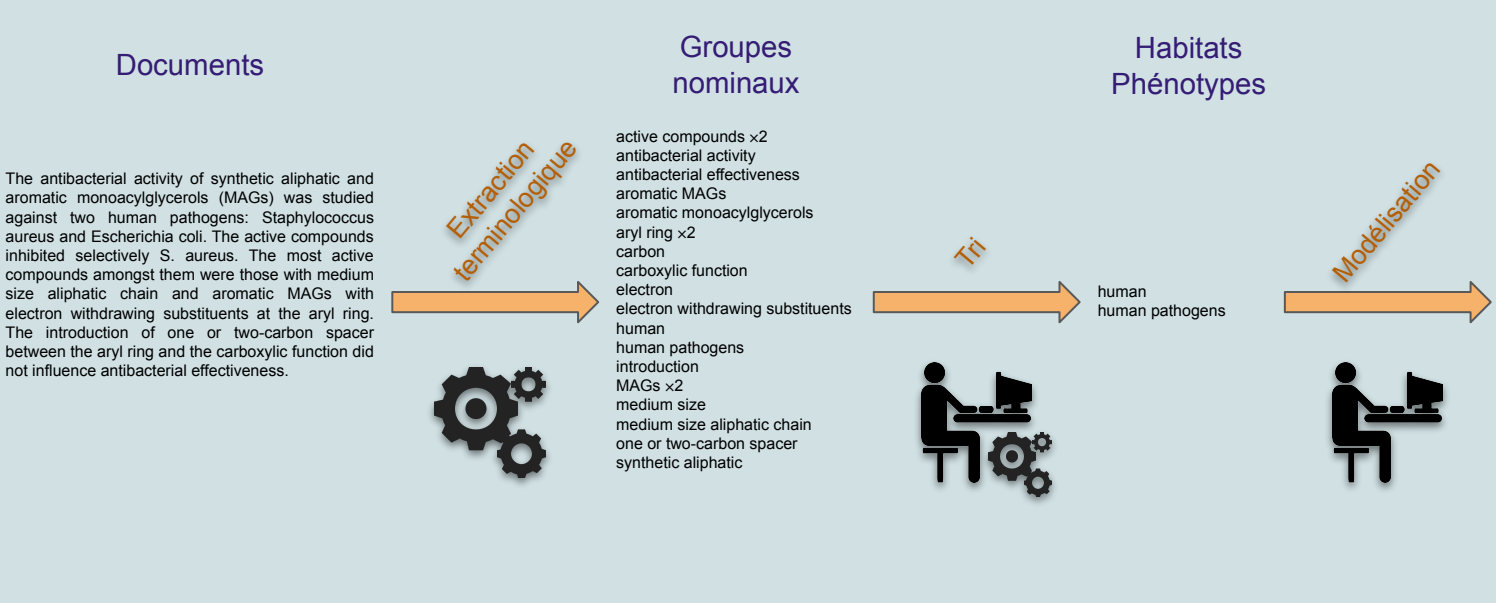
Nédellec *et al.*, 2017.

OntoBiotope : méthodologie top-down

- Approche épistémique pour les concepts les plus élevés :
  - Organismes vivants
  - Parties d’organismes vivants
  - Nourriture
  - Habitats liés à l’agriculture (et aquaculture)
  - Environnements médicalisés
  - Environnements naturels
  - …

OntoBiotope : methode bottom-up

- Approche ascendante pour les concepts les plus profonds.



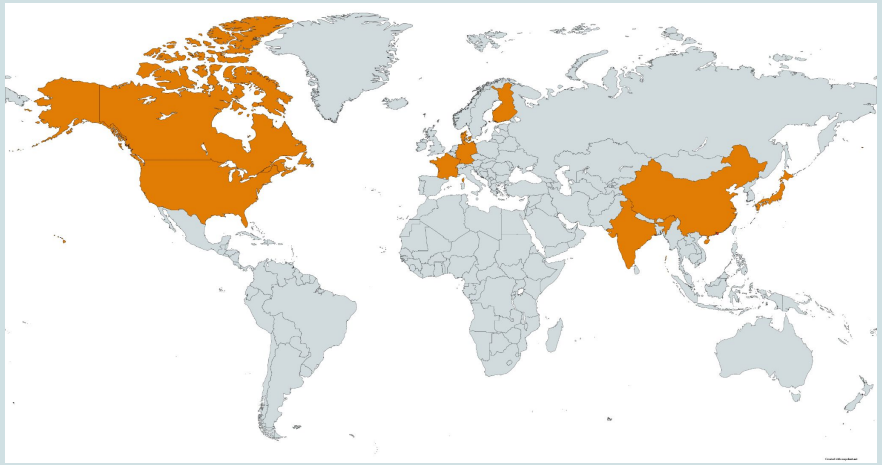
- Exhaustivité sémantique et lexicale.

Images: Shutterstock, 2022

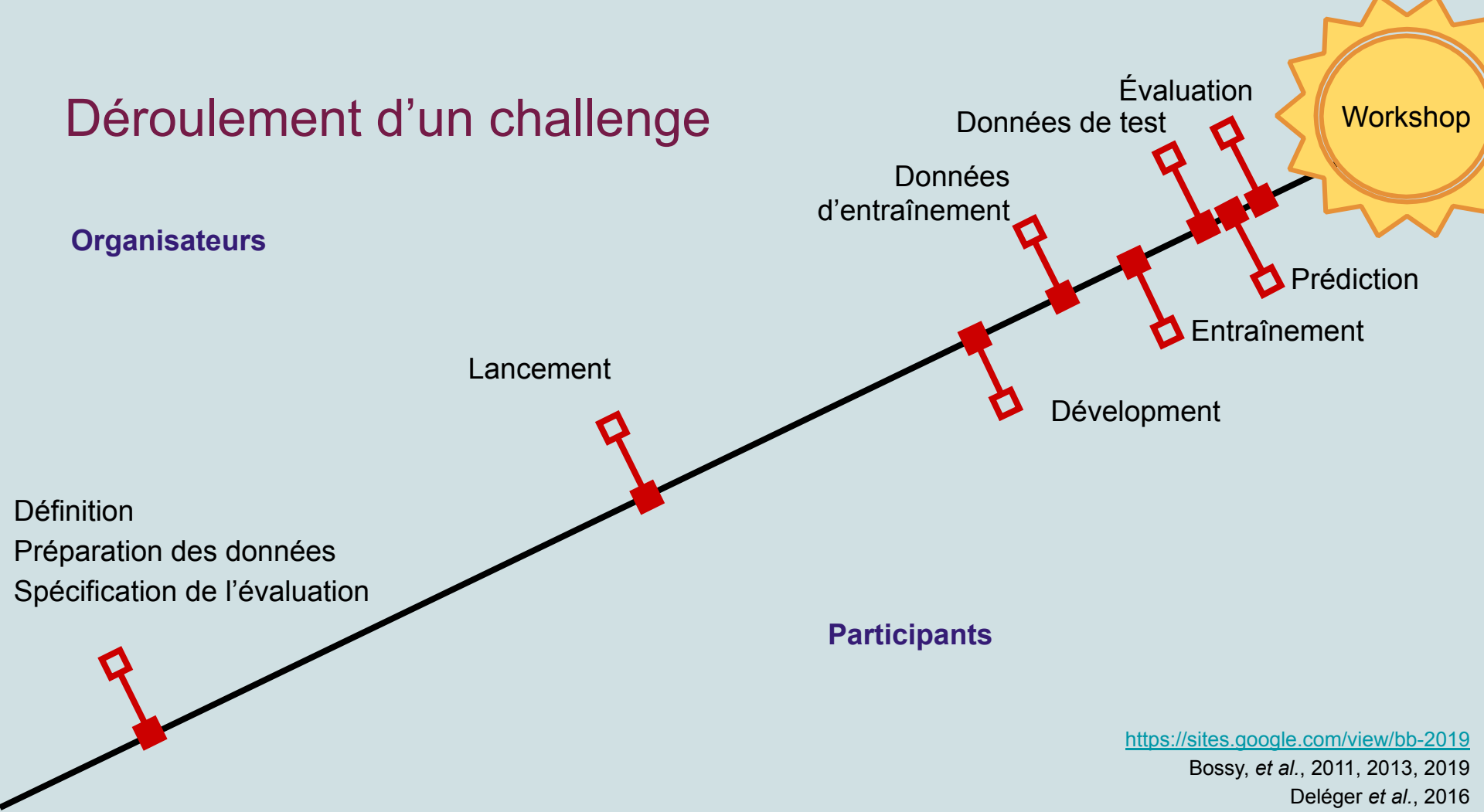
Challenge

Bacteria Biotopes

- Motivation
  - Mettre en avant les problématiques de INRAE
    - Biodiversité, Microbiologie, Agronomie et Alimentation
    - Entités nommées désignateurs non-rigides
  - Faire un point sur l'état de l'art méthodologique
- Bilan
  - BioNLP Shared Tasks
  - 4 éditions (2011, 2013, 2016, 2019)
  - 63 soumissions de 24 participants
  - Plus de 500 citations



Déroulement d'un challenge



Organisateurs

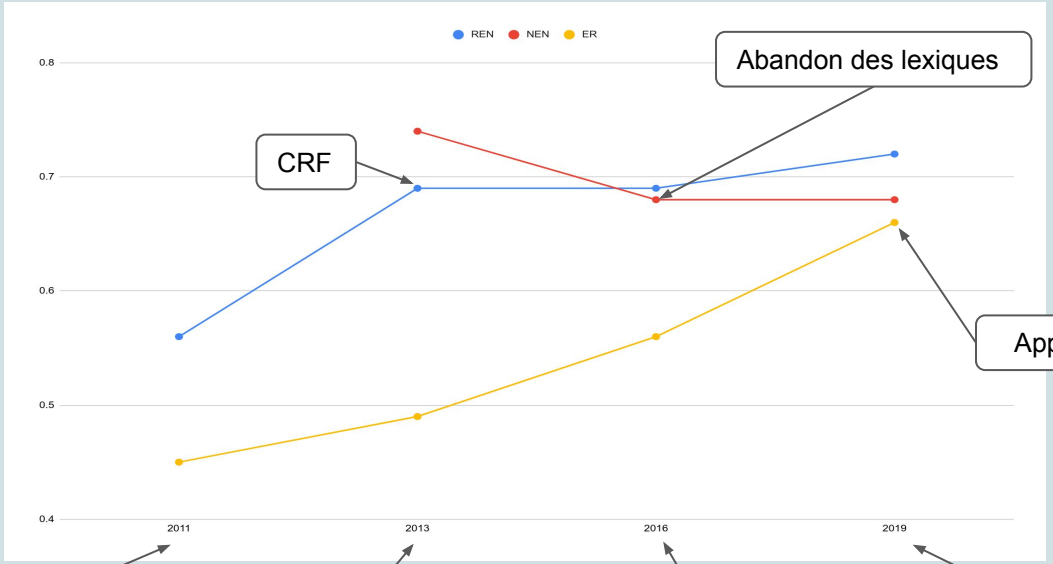
Participants

https://sites.google.com/view/bb-2019

Bossy, *et al.*, 2011, 2013, 2019

Deléger *et al.*, 2016

Évolution des méthodes et performances



Lexiques et patrons
Peu de ML (SVM, MEMM)

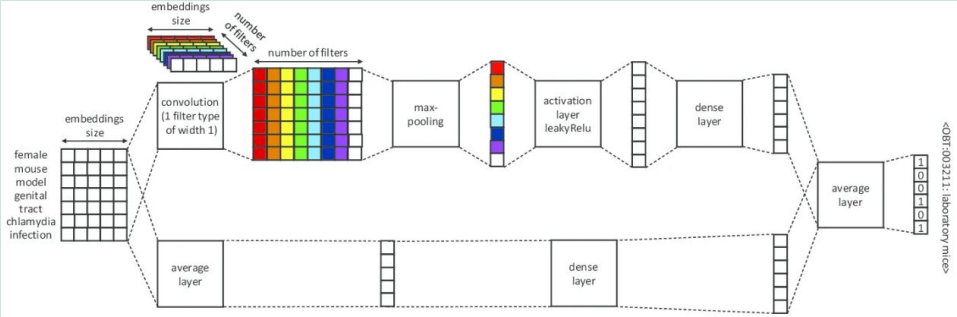
Analyse grammaticale
kNN, CRF, SVM

Abandon des lexiques
CRF, **SVM**, NN

SVM+LR, biLSTM, CNN,
Transformers

C-Norm

Méthode de normalisation qui combine supervision standard et faible.



	Similarité (Habitats)
PADIA (meilleur 2019)	0.68
C-Norm	0.78

Ferré, *et al.*, 2020.

Préparation des données

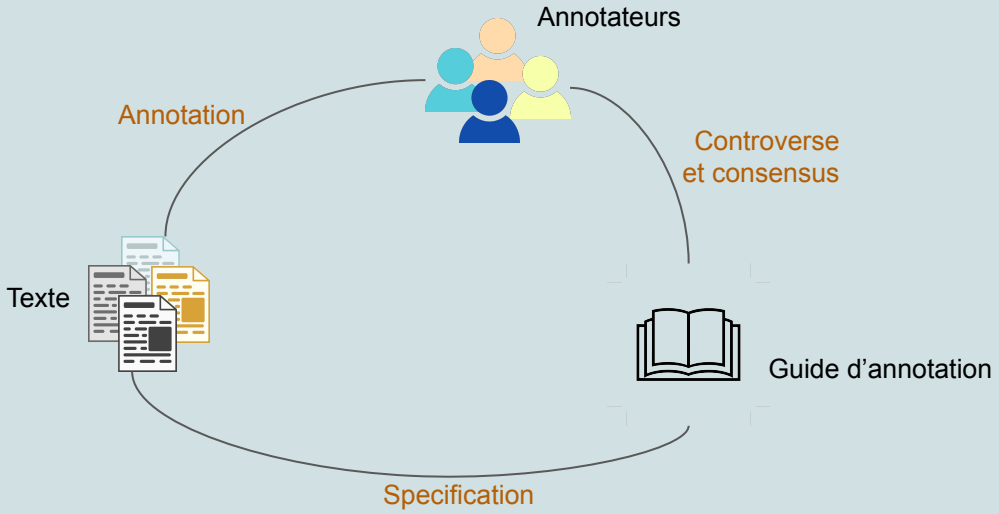
- *Gold standard* : annotation manuelle de documents.
- La “réalité du terrain” du texte est difficile à obtenir :
  - Le langage est sujet à interprétation.
  - Sa compréhension mobilise des connaissances extérieures.
  - La langue écrite est très dense en information.
  - Une annotation manuelle peut varier d’une personne à l’autre.

Differences in the in vitro adhesion of Staphylococcus strains on rabbit tissues were evaluated by viable unit counts and radio-labeling.

- “*in vitro*” disqualifie-t-elle la relation ?
- Ne s’agit-il pas d’une hypothèse ?
- Voudrait-on retrouver cette relation dans une base de données ?

Méthodologie d’annotation

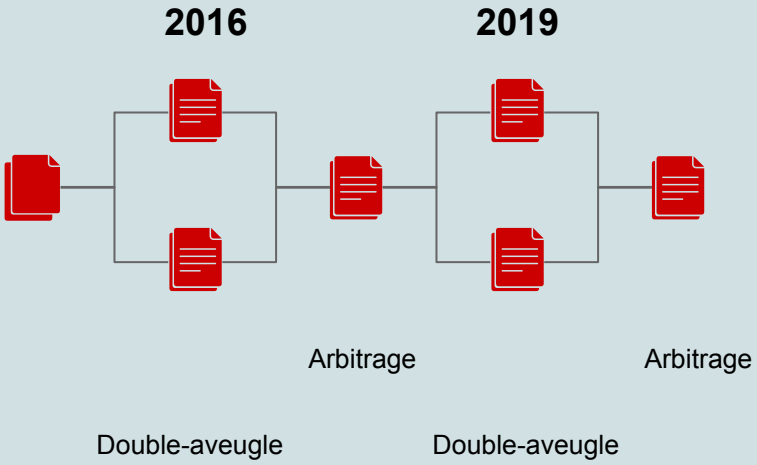
- Chaque texte doit être annoté plusieurs fois.
- Les annotateurs partagent les interprétations lors de l’annotation.
- Les décisions sont enregistrées dans le guide d’annotation (“guidelines”, “consignes”).



Extraits du guide Bacteria Biotopes

Le guide d’annotation est le seul élément de reproductibilité d’un texte annoté.

Annotation en double aveugle



	REN (F_1)	NEN (S)	ER (F_1)
Accord inter-annotateur	.89	.97	.79
Meilleur système	.72	.68	.66

Volume de Bacteria Biotopes

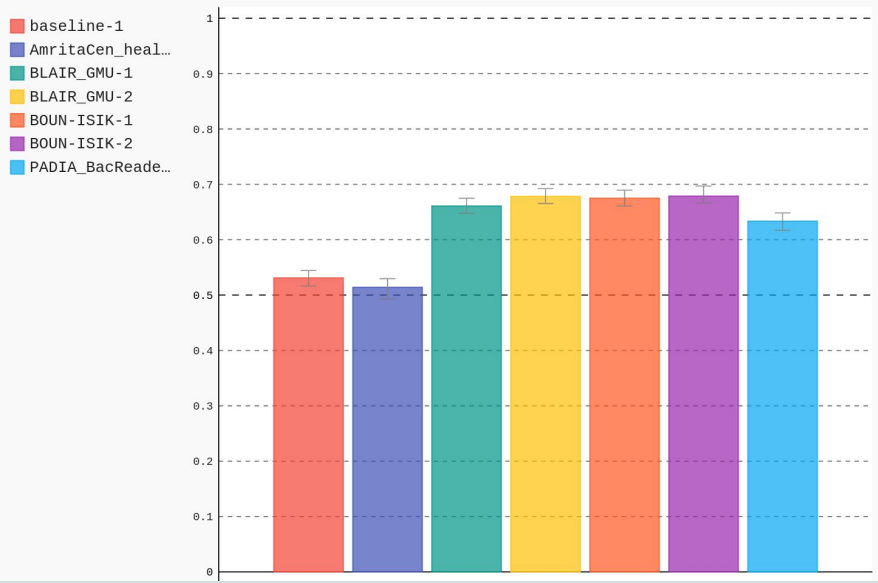
Documents	392	Abstracts PubMed + extraits full-text
Tokens	60.402	
Entités	7.232	
Relations	3.578	
Annotateurs	**18**	

Entraînement	133	⅓
Développement	66	⅙
Test	193	½

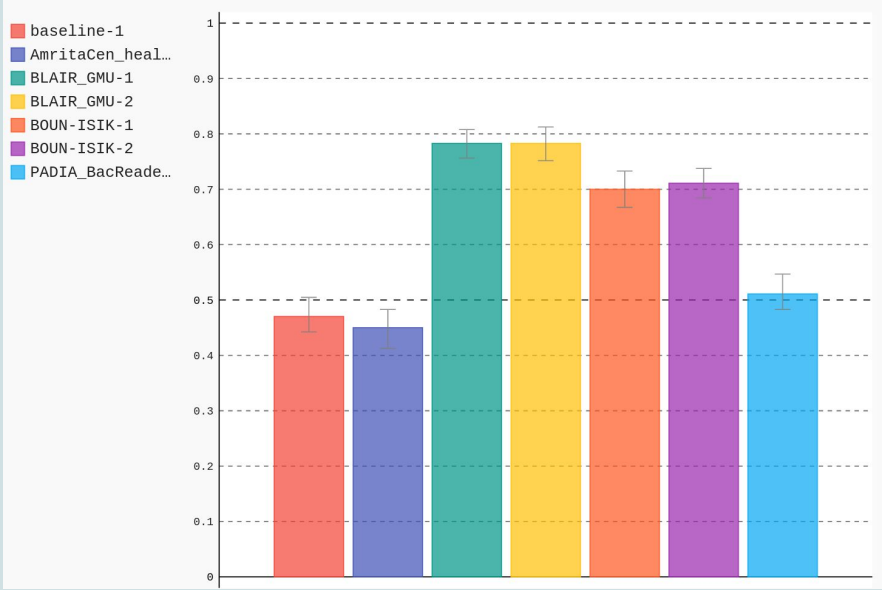
Métriques

- Métriques classiques de classification : Rappel, Précision, F-score
- Inconvénients
  - Information pauvre sur les forces et les faiblesses de chaque méthode.
  - Couverture partiellement les besoins.
  - Biais métriques.
- Solutions
  - Scores multiples : différencier les types d’annotations, ou des phénomènes linguistiques.
  - Métriques spécialisées : s’approcher des besoins ou pallier les biais.

Scores multiples (Normalisation)



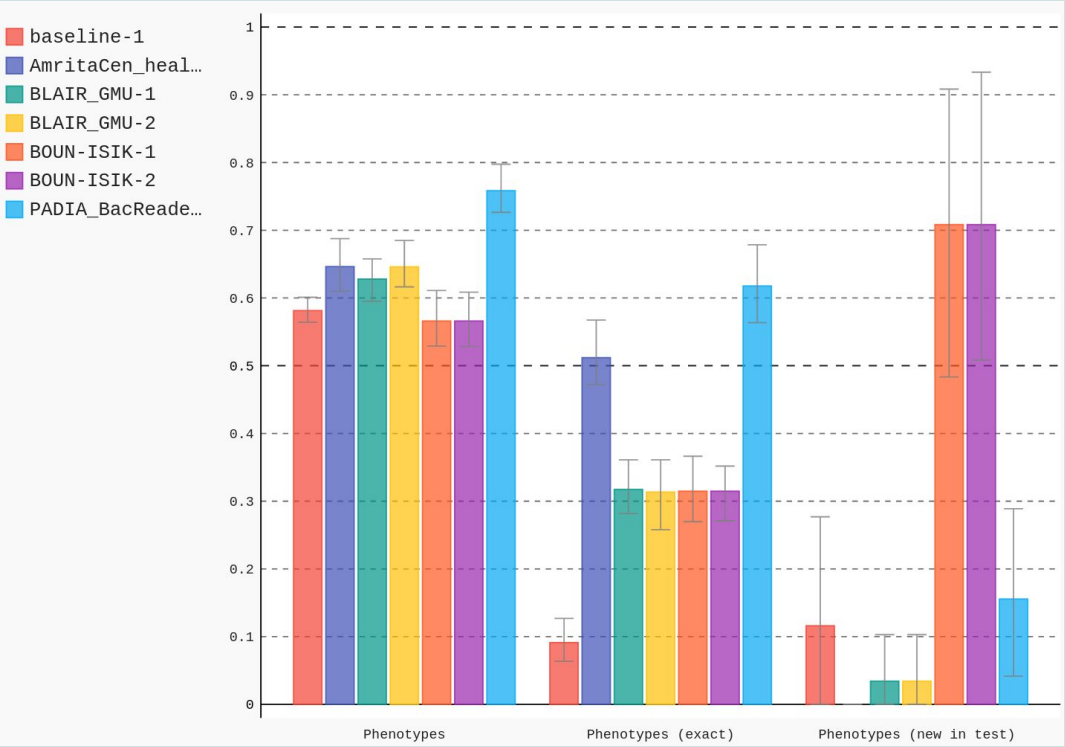
Score global



Microorganismes

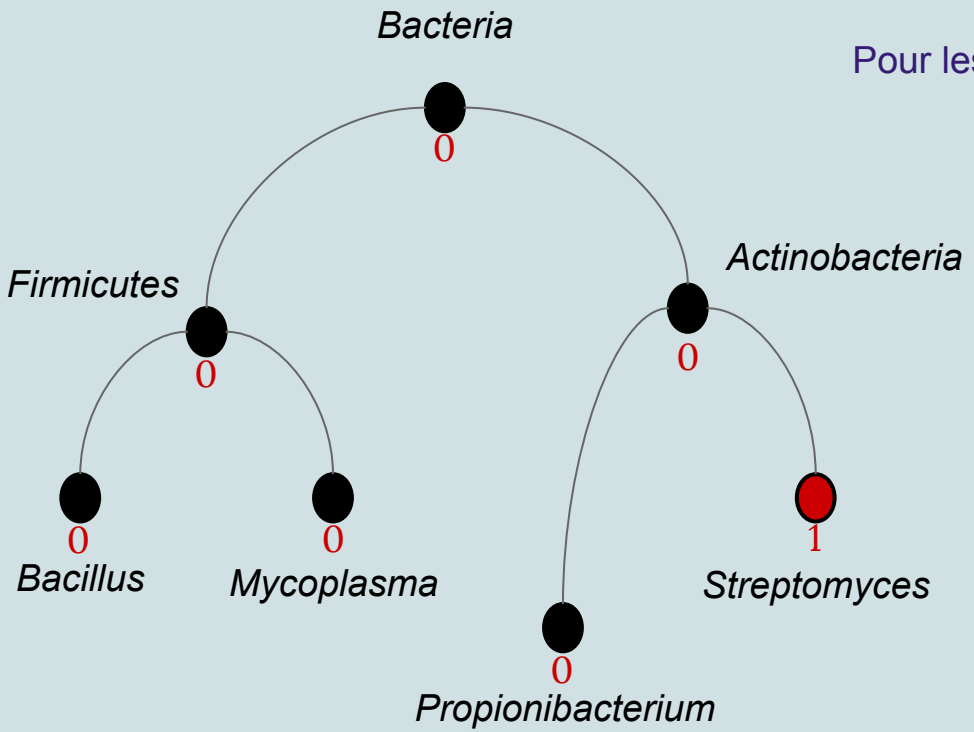
Métrique : similarité.

Scores multiples (Normalisation)



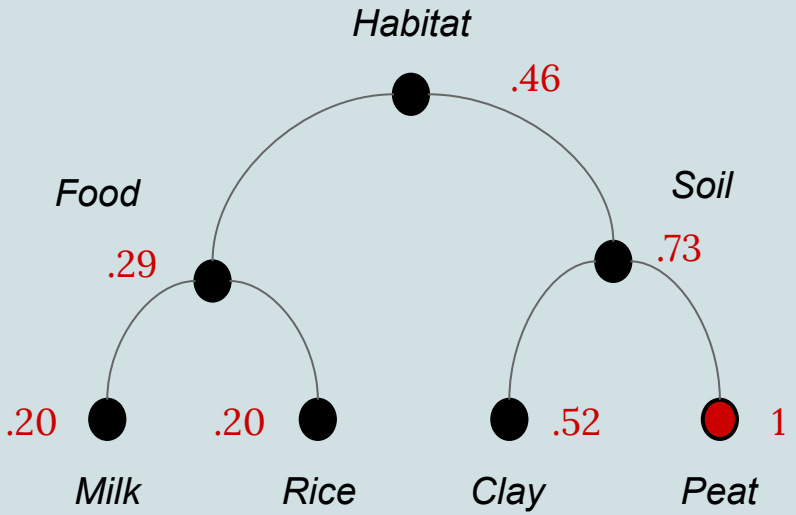
Phénotypes

Métriques spécialisées (Normalisation)



Pour les taxons, la normalisation doit être stricte.

Métriques spécialisées (Normalisation)



Pour les habitats et phénotypes une normalisation proche est acceptable.

Wang *et al.*, 2016.

Services

Extraction d’information et applications

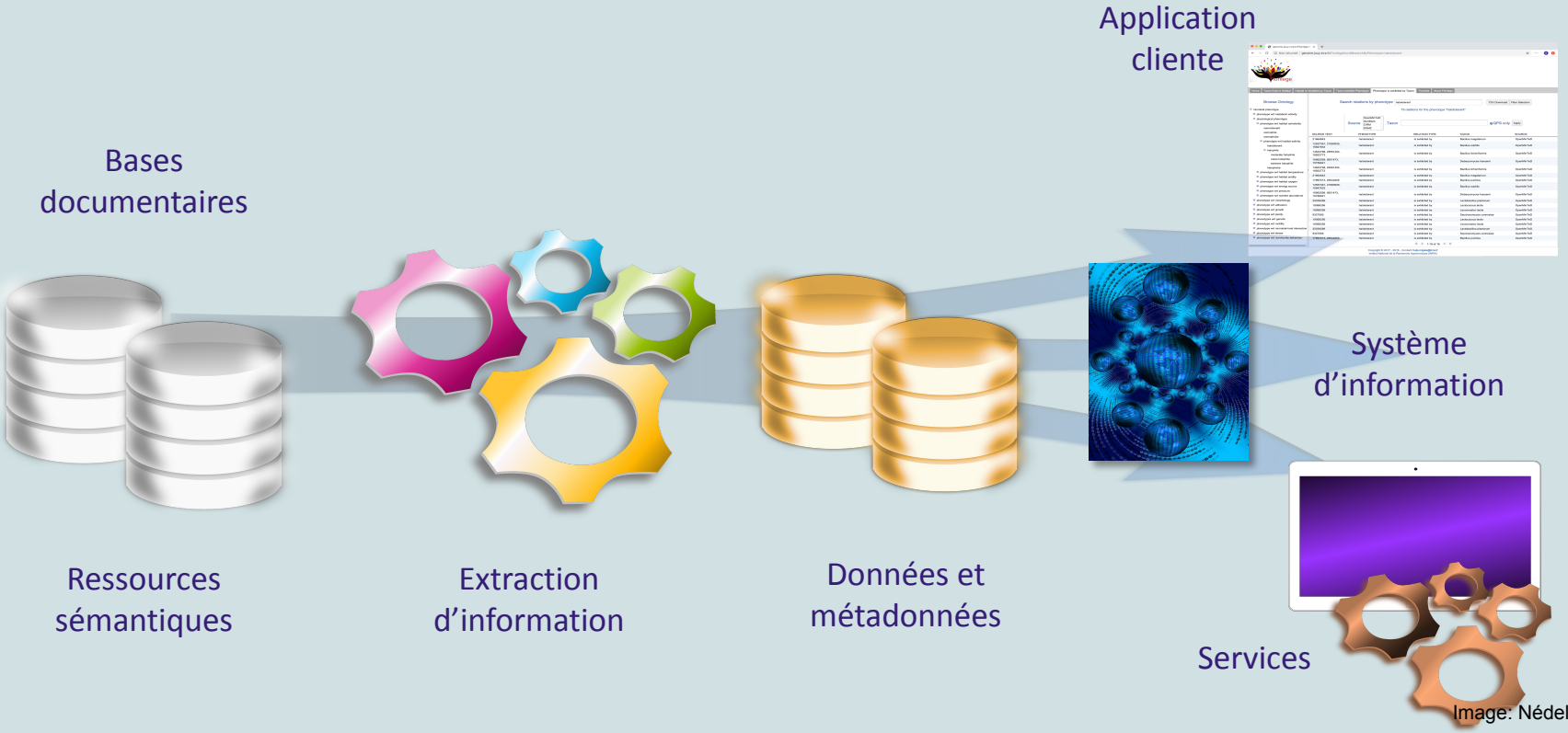


Image: Nédellec, 2020

Zoom sur l'extraction d'information

Les différentes tâches doivent être enchaînées et combinées.

- Passage à l'échelle → Robustesse.
- Mises à jour → Reproductibilité et infrastructure de calcul.
- Collaboration → Pérennité des outils et documentation.

AlvisNLP : moteur de workflows spécialisé dans l'extraction d'information modulaire et paramétrable.

- Workflow déclaratif → reproductibilité.
- Bibliothèque étendue et extensible → capitalisation, continuité entre recherche et service.

AlvisNLP
corpus processing engine

https://bibliome.github.io/alvisnlp/

Nédellec, Nazarenko & Bossy, 2029.

Ba & Bossy, 2016.

Exemples d’applications

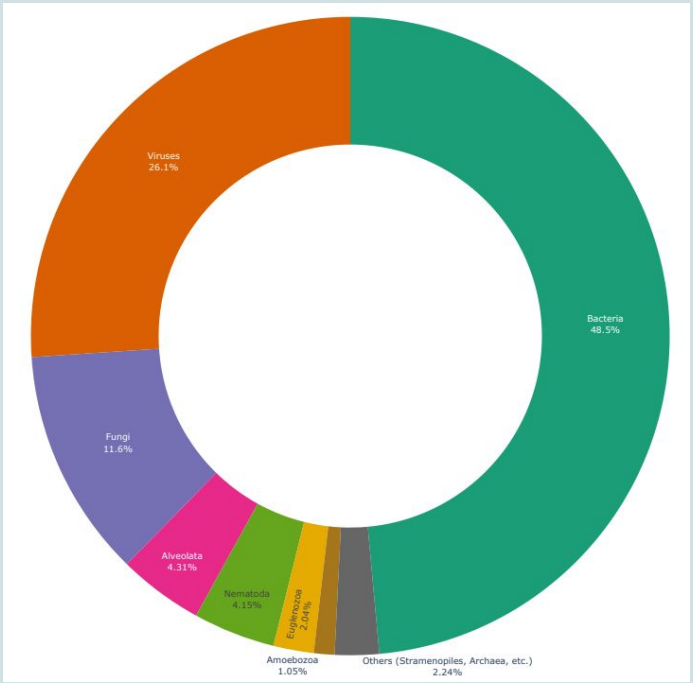
- Recherche d’information
  - Recherche sémantique de documents.
  - Intégration de données de sources diverses : bibliographie, données expérimentales, données de référence…
- Aide à la décision
  - Cartographie thématique (synthèse de corpus, tendances).
  - Profilage (par exemple, sélection de relecteurs).
- Extension de référentiels
  - Recherche de nouveaux concepts ou synonymes.

Omnicrobe

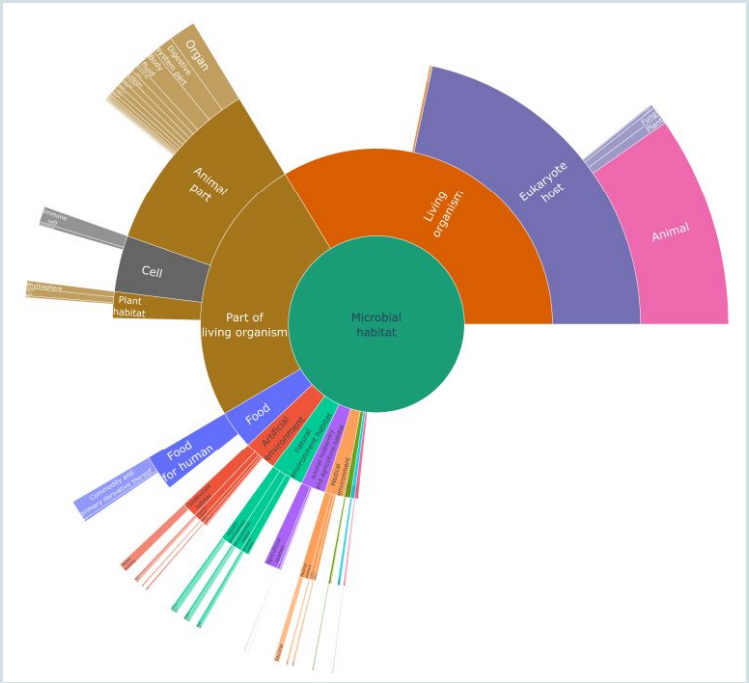
- Base de données de référence sur la biodiversité microbienne
  - Relations entre taxons, habitats et phénotypes.
  - Concepts normalisés par NCBI Taxonomy et OntoBiotope.
- Intégration de données de sources
  - Résumés PubMed.
  - GenBank (champs *organism*, *host*, *isolation_site*).
  - Entrées du catalogue DSMZ (champs *taxonomy*, *denomination*, *environment*).
  - Entrée de catalogues de CIRM.

Derozier *et al.*, 2023.

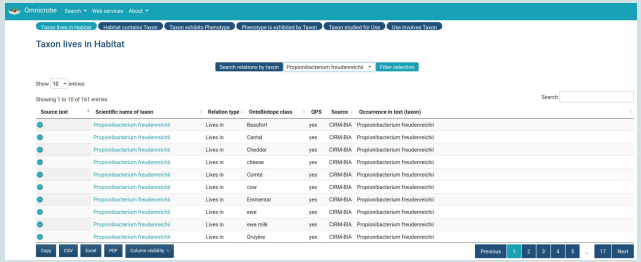
Omnicrobe : contenu



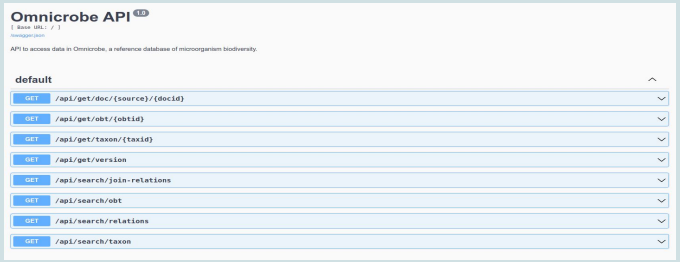
10^6 Relations



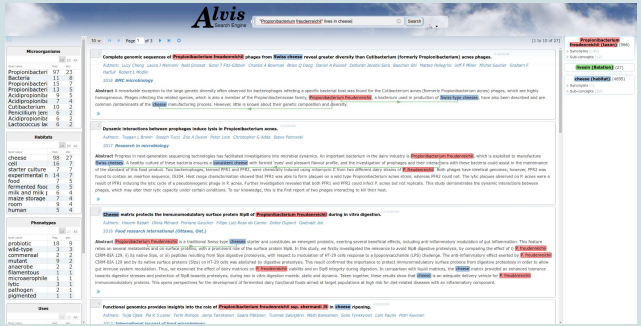
Omnicrobe : accès aux données



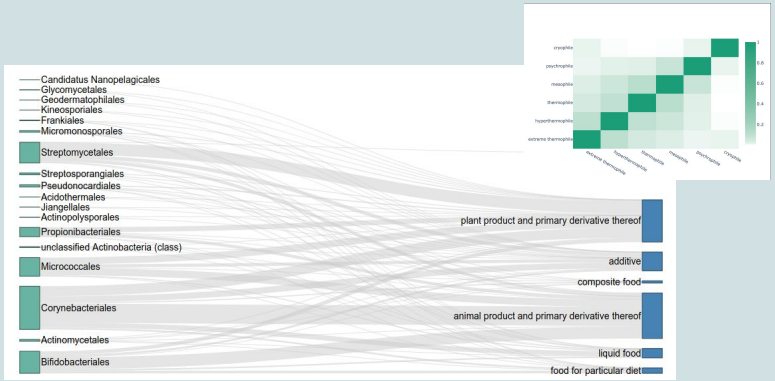
Interface Web



API



Moteur de recherche sémantique



Dump

https://omnicrobe.migale.inrae.fr

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

Omnicrobe : les technologies

Image: Nédellec, 2020

Omnicrobe : utilisateurs

- Florilège (Métaprogramme MEM)
  - Conception d’un ferment pour réaliser des yaourts à base de jus de soja.
- Food Microbiome Transfert (Micalis, MaIAGE, Migale, CNIEL)
  - Sélection de souches de référence pour la métagénomique du fromage.
- Open16S (projet pilote MICA)
  - Curation de métadonnées de projets métagénomiques.

L’équipe (“nous”)


Bedis Dridi


Marteen van de Guchte


Pierre Renault


Hélène Falentin

Micalis

STLO
SCIENCE & TECHNOLOGIE du LAIT & de l'ŒUF

OBT

Réseau OntoBiotope

Florilege

Projet Florilège

Besoin

Omnicrobe


Sandra Derozier

Logiciels


Frédéric Papazian


Robert Bossy


Mouhamadou Ba

Infrastructure

migale

Projet


Philippe Bessières


Claire Nédellec

MaIAGE

Annotation


Dialekti Valsamou


Léonard Zweigenbaum


Julien Jourde

OntoBiotope

ER


Anfu Tang

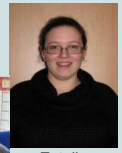
REN/NEN


Louise Deléger


Arnaud Ferré


Zorana Ratkovic


Wiktoria Golik


Estelle Chaix

Conclusion

L’application de l’extraction d’information nécessite

1. Analyse du besoin
   - Co-construction entre spécialistes de l’extraction d’information et les porteurs du besoin.
   - Préciser les contours des informations désirées et la faisabilité.
2. Identification et construction de ressources sémantiques
   - Activité à l’interface de l’ingénierie des connaissances et du domaine d’application.
   - Les ressources sémantiques permettent à la fois d’opérer l’extraction d’information et d’en exploiter le résultat.
3. Évaluer et développer des méthodes
   - La communauté d’extraction d’information est structurée autour de challenges.
4. Infrastructure de calcul et de développement logiciel
   - Porter les résultats aux utilisateurs.

Omnicrobe et la biodiversité microbienne

- Base de données de référence
  - Le service expose des résultats de l'extraction d'information.
  - Le développement et la maintenance mobilise trois équipes de l'unité MaIAGE.
- Perspectives
  - Emploi de méthodes à l'état de l'art (thèses Anfu Tang et Mariya Borovikova).
  - Agrégation de sources supplémentaires (convention avec l'ANSES).
- Ré-emploi dans le domaine de l'épidémiosurveillance végétale
  - Entités : organismes nuisibles, plantes, vecteurs, voies de dissémination.
  - Projets : TIERS-ESV (BioSP), Beyond (ANR).

https://www6.inrae.fr/beyond/

Morris, *et al.*, 2021.

Merci!