



**HAL**  
open science

## **OntoBiotope: une ontologie pour croiser les habitats microbiens avec les analyses de génomes.**

Robert Bossy, Estelle Chaix, Louise Deleger, Arnaud Ferré, Mouhamadou Ba, Philippe Bessieres, Claire Nédellec

### ► **To cite this version:**

Robert Bossy, Estelle Chaix, Louise Deleger, Arnaud Ferré, Mouhamadou Ba, et al.. OntoBiotope: une ontologie pour croiser les habitats microbiens avec les analyses de génomes.. Les journées Bioinformatique de l'Inra, Mar 2016, Montpellier, France. pp.1. hal-01455856

**HAL Id: hal-01455856**

**<https://hal.science/hal-01455856v1>**

Submitted on 5 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# OntoBiotope: une ontologie pour croiser les habitats microbiens avec les analyses de génomes

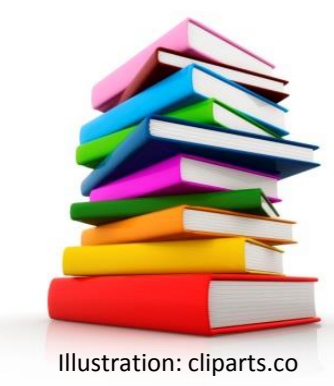
Robert Bossy, Estelle Chaix, Louise Deléger, Arnaud Ferré, Mouhamadou Ba, Philippe Bessières, Claire Nédellec

MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

L'évolution des technologies et des méthodes d'analyse nous permet désormais d'aborder des environnements complexes comme les écosystèmes microbiens: aliment, tube digestif, animal, traitement des eaux usées, sol, océan, etc., et la généralisation de ces approches métagénomiques à grande échelle délivre une énorme quantité d'information en terme d'espèces de microorganismes identifiées dans les environnements étudiés. Il est alors envisageable de considérer globalement les propriétés des écosystèmes dans lesquels vivent ces microorganismes, ainsi que les interactions et les associations qu'ils développent avec leur environnement et entre eux-mêmes. Cependant, le croisement systématique des espèces microbiennes et de leurs habitats pour trouver des corrélations significatives avec phylogénies, fonctions connues, contenus en gènes ou autre se heurte à une difficulté majeure : les habitats microbiens sont décrits sous des formes très variables dans le texte des publications scientifiques ou des bases de données génomiques de référence comme GenBank, GOLD (Genomes Online Database) ou SRA (Sequence Read Archive).

Le projet OntoBiotope a développé des méthodes et des outils d'analyse automatisée pour exploiter ces ressources et d'en extraire des relations entre organismes et milieux. Ainsi, OntoBiotope devrait permettre d'appréhender plus finement le rôle des caractéristiques des milieux sur l'évolution des microorganismes et leurs propriétés physiologiques.

## Habitats microbiens et textes scientifiques



La littérature scientifique regorge de mentions très diverses d'habitats, décrits en langue naturelle.

While the type strain of *C. acetobutylicum* was isolated from soil, *C. acetobutylicum* is ubiquitous. It has been found in ....

*Bifidobacterium longum* is found in adult humans and formula fed infants as a normal component of gut flora.

... strains were selected to represent all *B. cereus* isolates ( $n = 457$ ) with genotypes (RAPD-PCR) that frequently colonised the silo tanks of at least two of the sampled eight dairies.

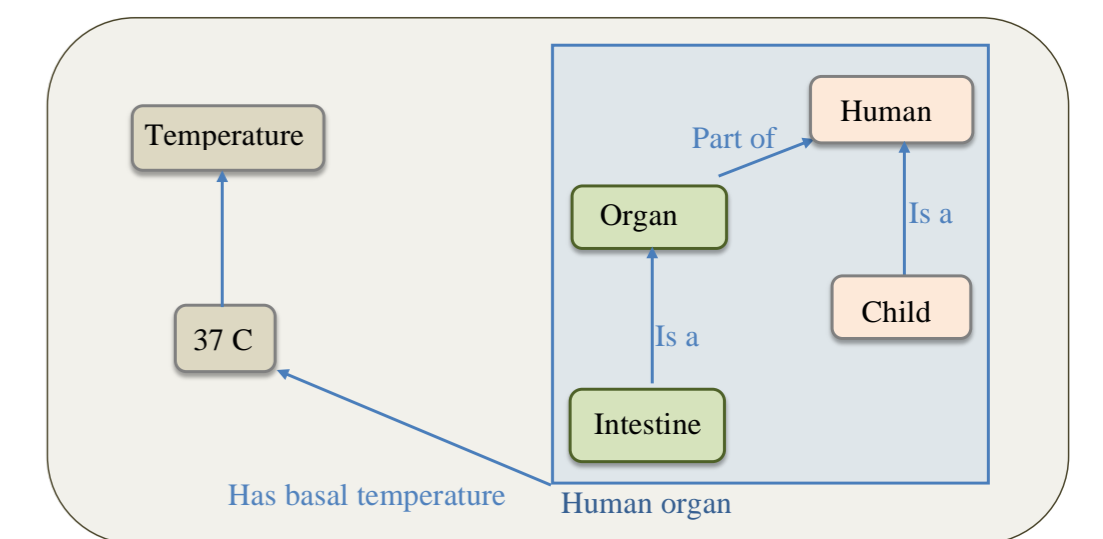
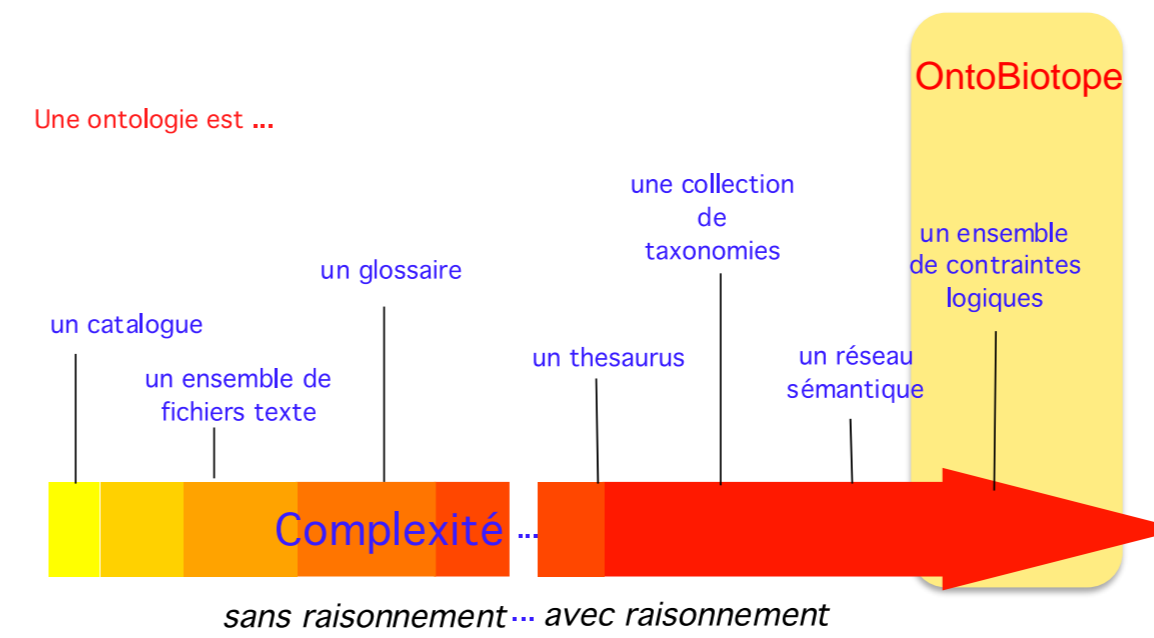
Les habitats microbiens de tous types sont décrits dans la littérature scientifique et de plus en plus dans les bases de données de référence.

Mais ces informations ne sont ni "requêtables", ni comparables parce que décrites à différents niveaux de précision et non normalisés.

La question : Comment extraire et indexer ces informations sur l'habitat microbien pour permettre des recherches de corrélation avec d'autres données ?

## Le projet OntoBiotope : construire une ontologie des habitats microbiens

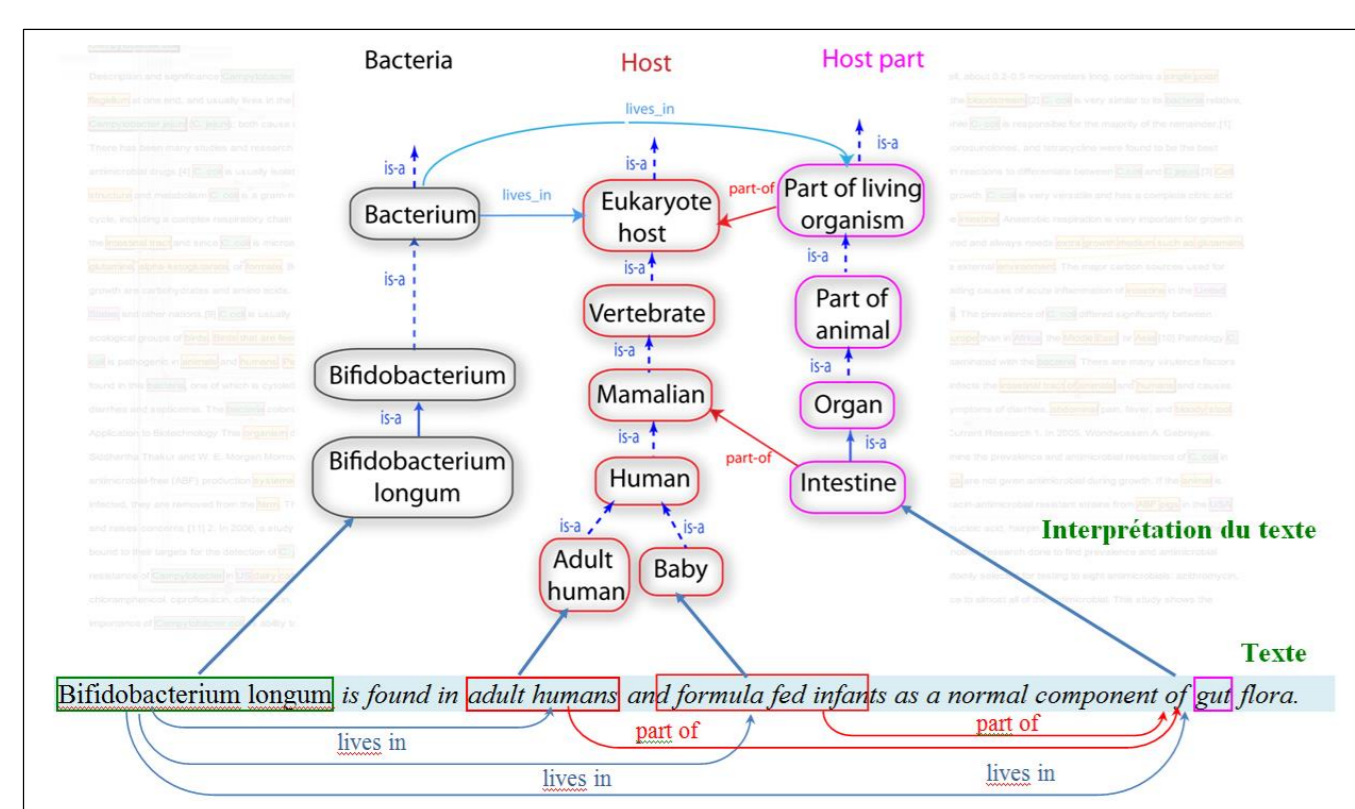
Qu'est-ce une ontologie ? "A formal, explicit specification of a shared conceptualisation" (T. Gruber, 1992)



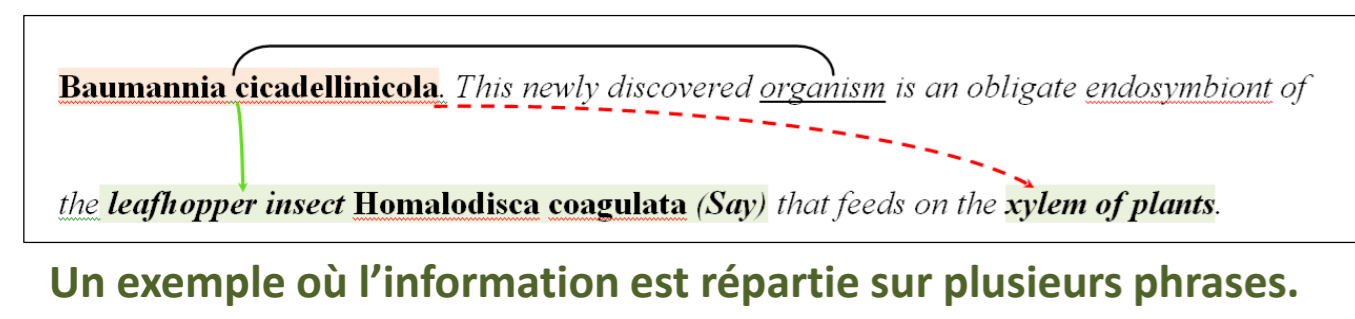
Un exemple des informations contenues dans l'ontologie d'OntoBiotope (dans la version 2013 disponible <http://agportal.lirmm.fr/ontologies/ONTOBIOTOPE>).

Ce modèle permet de faire des déductions qui peuvent enrichir nos connaissances. Si *Clostridium perfringens* vit dans l'intestin humain (donnée d'un article), alors elle vit à 37 °C (inférence avec l'ontologie).

## Le projet OntoBiotope : extraire les relations microbe - habitat des textes



Un exemple d'interprétation de texte.



Un exemple où l'information est répartie sur plusieurs phrases.

... treatment plant for halogenated hydrocarbon waste ... → waste treatment plant

... alkaline anaerobic environment of Mono Lake ... → alkaline lake

Exemples d'association de termes du texte à des concepts de l'ontologie (plant, lake).

Le projet OntoBiotope a permis de développer de nouvelles méthodes automatiques d'extraction d'information du texte, évaluées dans le challenge BioNLP-ST : l'ontologie OntoBiotope a été utilisée pour créer un corpus annoté de descriptions d'habitats bactériens

Meilleurs scores à la tâche Bacteria Biotope de BioNLP-ST 2013 (Bossy et al. 2015):

Reconnaissance et typage des entités : 57% SER  
Extraction des relation : 59% F-mesure

Ces méthodes font appel à des méthodes d'Intelligence Artificielle (Traitement Automatique de la Langue et d'Apprentissage Automatique).

Un nouveau corpus est disponible : <http://2016.bionlp-st.org/tasks/bb2> avec 2902 concepts d'habitat annotés

## Base de données en ligne des relations extraites de PubMed

La base de données OntoBiotope est en ligne : <http://bibliome.jouy.inra.fr/demo/alvisdb/obt/browse2>

L'interface web (navigateur internet) permet

- de chercher par différents niveaux de taxonomie des bactéries et/ou habitats et
- d'obtenir toutes les relations bactérie – habitat correspondantes dans un fichier csv (utilisables dans Excel).

## Moteur de recherche en ligne

Un moteur de recherche bibliographique sémantique intelligent est également disponible en ligne : <http://bibliome.jouy.inra.fr/demo/pubmed-ontobiotope/alvisir/webapi/search>

Ce moteur de recherche permet de chercher dans >1.150.000 résumés PubMed, indexés par OntoBiotope, et de consulter les résumés identifiés (Mis à jour en octobre 2015)

Il est capable de combiner plusieurs informations de l'ontologie pour identifier des nouveaux termes. Dans l'exemple présenté, une recherche a été faite pour la bactérie "Vibrio" dans l'habitat "mariculture farm". Le moteur de recherche a identifié automatiquement que "Danish eel farm", un terme qui n'était pas présent dans l'ontologie au départ, représente un "mariculture farm".

## CONCLUSION et PERSPECTIVES

### Animation

- Identification des membres d'un nouveau réseau intéressés par cette thématique et analyse du besoin
- Recensement des microbiologistes par thème d'application
- Réunions, présentation de l'état de l'art



### Résultats

- Définition de champs d'intérêt/besoins, applications
- Nouvelles méthodes de text-mining
- Développement et mise à disposition de services

### Futur

- Projet H2020 OpenMinTeD
  - Infrastructure Européenne de text-mining
  - Etude de la biodiversité (GBIF, CRB)
  - Services en ligne
- Thèse IDI (IDEX) sur l'analyse de termes complexes avec une ontologie