



HAL
open science

Omnicrobe, an open-access database of microbial habitats, phenotypes and uses extracted from text

Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Valentin Loux, Hélène Falentin, Claire Nédellec

► To cite this version:

Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, et al.. Omnicrobe, an open-access database of microbial habitats, phenotypes and uses extracted from text. JOBIM 2022, Jul 2022, Rennes, France. , Proceedings Posters Demos, pp.42. hal-04061565

HAL Id: hal-04061565

<https://hal.inrae.fr/hal-04061565>

Submitted on 6 Apr 2023

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.



Distributed under a Creative Commons Attribution 4.0 International License

Omnicrobe, an open-access database of microbial habitats, phenotypes and uses extracted from text

S. Dérozier¹, R. Bossy¹, L. Deléger¹, M. Ba^{1,2}, E. Chaix¹, V. Loux^{1,2}, H. Falentin³, C. Nédellec¹

¹ Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

² Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas 78350, France

³ INRAE, STLO, Rennes, France



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

The drastic increase in microbe descriptions, habitats, phenotypes and uses in databases, reports and papers presents a two-fold challenge for the access to the information: (1) a standard representation to integrate heterogeneous data and (2) the normalization of textual descriptions by semantic analysis. Recent information extraction technologies from the text mining domain offer powerful ways to detect and structure textual information along ontology-based representations.

The Omnicrobe database contains around 1 million descriptions of microbe properties and is populated by an Information Extraction workflow:

a) 6 information sources:

- biological resource catalogues: Inrae **CIRM**, **BacDive** by DSMZ,
- sequence database: **GenBank**,
- scientific literature: **PubMed** abstracts in microbiology.

b) entity recognition: microbe **taxa**, **habitats**, **phenotypes** and **uses**.

c) normalization with taxa from the **NCBI taxonomy** [2] and concepts from the **OntoBiotope ontology** [3].

d) **relation extraction** between entities.

All information is aggregated, merged and indexed to be accessible from the information system. Omnicrobe offers powerful ways to express simple and complex ontology-based queries.

Omnicrobe also exposes an **API** (Application Programming Interface) that allows users to automatically integrate microbe biodiversity knowledge in external information systems.

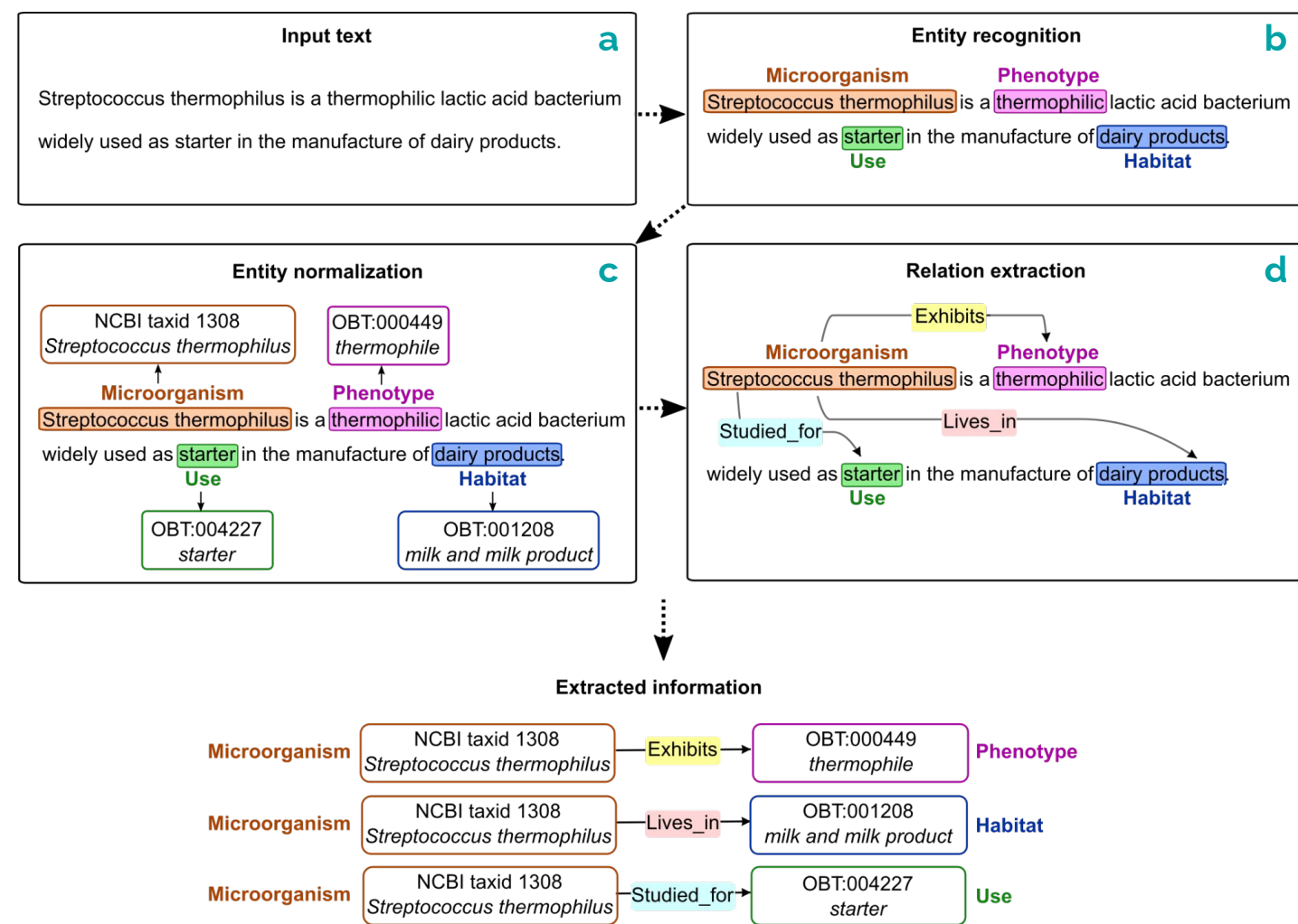


Figure 1. Text mining process.

Search results for 'cheese' in 'Habitat contains Taxon':

Source text	OntoBiotope class	Relation type	Scientific name of taxon	QPS	Source
KC429595, KC429639, ...	queso fresco	Contains	[Brevibacterium] frigoritolerans	no	GenBank
Occurrence in text (habitat) fresco surface					
Occurrence in text (taxon) [Brevibacterium] frigoritolerans					
Full source text KC429595, KC429639, KC429588, KC429641, KC429585					
19965993	cheese	Contains	[Candida] cabralensis	no	PubMed
19965993	blue veined cheese	Contains	[Candida] cabralensis	no	PubMed

Figure 2. Omnicrobe web interface.

Quickly select the desired type of search with tabbed navigation (A). Searches (B) are carried out using habitats, phenotypes, uses and taxa names, synonyms and hierarchy from the OntoBiotope ontology and the NCBI taxonomy. Depending on the type of search, several filters are available: e.g. source of the information, or food safety certification. The user may also navigate in the hierarchical structure of OntoBiotope (C) and select the class of interest, habitats, phenotypes or uses. The results (D) are displayed in columns for better readability. The original PubMed or GenBank entry is accessible through a link.

Omnicrobe has been used to quickly target useful strains in a food innovation application [4]. The Omnicrobe database was used to identify species capable of fermenting soybean juice. Out of 10 species represented by 229 strains tested in vitro, 9 species, i.e. 179 strains, acidified (pH<6) soybean juice within 48h.

In future work, we plan to extend Omnicrobe to include new data sources such as GOLD, GBIF, ...

References

1. Reimer LC, Vetcinina A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, Overmann J. BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.* 47(D1):D631-D636, 2019.
2. Schoch CL, Ciufo S, Domrachev M, Hottel CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Sousoff V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*, 2020.
3. Chaix E, Deléger L, Bossy R., Nédellec C. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology*, 2018.
4. Harlé O., Falentin H., Niay J., Valence F., Courselaud C., Chuat V., Maillard M-B., Guédon E., Deutsch S-M., Thierry A. Diversity of the metabolic profiles of a broad range of lactic acid bacteria in soy juice fermentation. *Food microbiology*, 89, 2020.

