



**HAL**  
open science

## Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach

Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, 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 and phenotypes using a comprehensive text mining and data fusion approach. PLoS ONE, 2023, 18 (1), pp.e0272473. 10.1371/journal.pone.0272473 . hal-03984703

**HAL Id: hal-03984703**

**<https://hal.inrae.fr/hal-03984703v1>**

Submitted on 13 Feb 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

## RESEARCH ARTICLE

# Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach

Sandra Dérozier<sup>1\*</sup>, Robert Bossy<sup>1</sup>, Louise Deléger<sup>1</sup>, Mouhamadou Ba<sup>1,2</sup>, Estelle Chaix<sup>1</sup>, Olivier Harlé<sup>3</sup>, Valentin Loux<sup>1,2</sup>, Hélène Falentin<sup>3</sup>, Claire Nédellec<sup>1</sup>

**1** Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France, **2** Université Paris-Saclay, INRAE, BioInfOmics, MIGALE Bioinformatics Facility, Jouy-en-Josas, France, **3** INRAE, STLO, Rennes, France

\* [sandra.derozier@inrae.fr](mailto:sandra.derozier@inrae.fr)



## OPEN ACCESS

**Citation:** Dérozier S, Bossy R, Deléger L, Ba M, Chaix E, Harlé O, et al. (2023) Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach. PLoS ONE 18(1): e0272473. <https://doi.org/10.1371/journal.pone.0272473>

**Editor:** Tarunendu Mapper, Bristol-Myers Squibb Company, UNITED STATES

**Received:** July 19, 2022

**Accepted:** January 4, 2023

**Published:** January 20, 2023

**Copyright:** © 2023 Dérozier et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All Omnicrobe data is Open Source and publicly available through a programmatic interface (<https://omnicrobe.migale.inrae.fr/api-doc>) hosted by the Migale institutional facility (<https://migale.inrae.fr>). Ontologies and taxonomies The ontology of microorganism habitats, phenotypes, and uses, OntoBiotope, is available on AgroPortal (<http://doi.org/10.15454/1.4382640528105164E12>). The source code to generate the taxonomy is Open Source and publicly available at

## Abstract

The dramatic increase in the number of microbe descriptions in databases, reports, and papers presents a two-fold challenge for accessing the information: integration of heterogeneous data in a standard ontology-based representation and normalization of the textual descriptions by semantic analysis. Recent text mining methods offer powerful ways to extract textual information and generate ontology-based representation. This paper describes the design of the Omnicrobe application that gathers comprehensive information on habitats, phenotypes, and usages of microbes from scientific sources of high interest to the microbiology community. The Omnicrobe database contains around 1 million descriptions of microbe properties. These descriptions are created by analyzing and combining six information sources of various kinds, i.e. biological resource catalogs, sequence databases and scientific literature. The microbe properties are indexed by the Ontobiotope ontology and their taxa are indexed by an extended version of the taxonomy maintained by the National Center for Biotechnology Information. The Omnicrobe application covers all domains of microbiology. With simple or rich ontology-based queries, it provides easy-to-use support in the resolution of scientific questions related to the habitats, phenotypes, and uses of microbes. We illustrate the potential of Omnicrobe with a use case from the food innovation domain.

## Introduction

This paper describes the Omnicrobe database. It gathers comprehensive information on habitats, phenotypes, and usages of microbes. This information is critical for the development of a large range of research, economic and social activities among which microbial ecosystem services and healthcare are two significant topics. Microbial communities render important functions to their ecosystem [1]. These functions are a result of biotic and abiotic interactions dependent on the habitat where microbes live. These last decades, communities have been

<https://forgemia.inrae.fr/omnicrobe/extended-microorganisms-taxonomy>. Information sources All data sources processed in Omnicrobe are Open Source and publicly accessible on line: PubMed (<https://pubmed.ncbi.nlm.nih.gov>), GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>), BacDive (<https://bacdive.dsmz.de>), CIRM BIA (<https://collection-cirmbia.fr>), CIRM-CFBP (<https://cirm-cfbp.fr>), CIRM-Levures (<https://cirm-levures.bio-aware.com>) and QPS information from EFSA (<https://www.efsa.europa.eu/en/topics/topic/qualified-presumption-safety-qps>). Information system source code Workflows and source code of the Omnicrobe application are distributed on an institutional Git repository (<https://forgemia.inrae.fr/omnicrobe>) under an Apache license. The source code of the text-mining processing workflow is available at <https://forgemia.inrae.fr/omnicrobe/text-mining-workflow>. The source code of the database integration workflow is available at <https://forgemia.inrae.fr/omnicrobe/omnicrobe-database>. The source code of the web interface and the programmatic interface is available at [https://forgemia.inrae.fr/omnicrobe/omnicrobe\\_web](https://forgemia.inrae.fr/omnicrobe/omnicrobe_web).

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

inventoried by high-throughput culturomics and metagenomic analyses revealing the numerous ecosystems' microbial richness and their ecological importance [2]. Notable examples include the ocean (e.g., the Tara Ocean project [3]) and the human gut (Human Microbiome projects [4]). Deciphering microbial traits and phenotypes enables researchers to identify possible beneficial uses for humans in several domains ranging from healthcare or pollution remediation to food [5]. The quantity and the diversity of the information publicly available on focused microbial habitats and phenotypes exponentially increase as a consequence of the growing diversity of microbial ecology research and the parallel evolution of high-throughput sequencing technologies and novel modeling approaches.

However, methods to assess how microbes are distributed across environments remain limited [6]. The information about microbe habitats and phenotypes is indeed scattered among many different sources, ranging from structured databases (e.g. genetic banks, biological resources, biodiversity databases) to document collections (e.g. scientific literature, reports). Despite the increase in volume and openness of these sources, this wealth of information remains largely underexploited for many reasons: it is distributed in a wide range of sources; the information is described by a broad variety of features and metadata that prevents semantic interoperability (e.g. habitat is also called source, isolation, location or host); a large part of the information is expressed in free texts even in databases, which makes it difficult to find.

Centralization of microbe habitat and phenotype information by information providers and aggregators (e.g. PubMed Central; <https://www.ncbi.nlm.nih.gov/pmc/>, GOLD; <https://gold.jgi.doe.gov/>), standardization of database metadata and increasing use of controlled vocabulary and nomenclatures (e.g. NCBI species taxonomy; <https://www.ncbi.nlm.nih.gov/taxonomy>) are heading in the right direction for better access, interoperability and reuse following FAIR principles [7]. The Omnicrobe database project falls fully in line with this effort. Its purpose is to offer a powerful way to retrieve information on microbe taxa, phenotypes, habitats, and uses (we define a "use" as any microbial property that can be targeted for human purposes, such as food aromatization) across environments and usages through a single application.

To achieve this goal, the Omnicrobe application automatically aggregates and structures the information that it collects from various public sources (section Information sources). Our work on Omnicrobe focuses on a neglected yet critical aspect: to be easily searched and processed at a large scale, the habitat, phenotype, and use information gathered on microbes should be organized along standard classifications [8]. Information in Omnicrobe is indexed by a comprehensive controlled and hierarchically organized vocabulary so that the information can be searched in a concise and fast way, regardless of the diversity of terms in source texts. We use text mining methods for parsing and indexing taxon, habitat, phenotype and use mentions and their relationships [9]. The taxa are indexed by the Omnicrobe taxonomic reference, an extended version of the NCBI taxonomy [10, 11], and the habitat, phenotype, and use information is indexed by the OntoBiotope ontology [12] (section Ontologies and taxonomies). The Omnicrobe database is publicly available at <https://omnicrobe.migale.inrae.fr/>.

## Background and state of the art

Microbiology covers a wide range of research domains, e.g., molecular biology, ecology, systems biology, evolution, and epidemiology, which all deal with microbe environments and phenotypes. Advances in high-throughput technologies generate a tremendous quantity of open access data thanks to the historical investment in shared information systems since the nineties. As a result, data on microbe environments and phenotypes can be found in a diverse range of sources. The most prominent information sources are metagenomic experiment

datasets, although the lack of species identification remains an obstacle. The *Genome online database* (GOLD) of the Joint Genome Institute aggregates isolation information from metagenomics experiments on thousands of identified strains [13]. Genetic sequences are also often published with information on isolation samples, as in the GenBank or Biosample databases; this is notably true for complete genome sequences and sequences used for taxon identification (e.g., 16S rRNA gene) [14]. Numerous Biological Resource Centers (BRC) also publish catalogs of their microbial resources with detailed and curated information on the isolation places and phenotypes. The *WFCC Global Catalogue of Microorganisms* aggregates information from 133 collections on almost 500,000 strains with their isolation sources [15]. The *Bacterial Diversity Metadatabase* (BacDive) [16] provides information on more than 80,000 strains of the DSMZ collection, one of the largest microbe collections in the world. Biodiversity inventories such as the *Global Biodiversity Information Facility* (GBIF) and the *Encyclopedia of Life* (EOL) include microorganisms with geographical information but lack information on observation places. *FoodMicrobionet* gathered metagenomic food sample metadata from public databases and linked them with the FoodEx2 vocabulary [17]. Though limited to food, FoodMicrobionet was one of the first attempts to link genomic and environmental data; it is no longer available, and only its archive can be accessed. These information sources can all be searched through open access web pages. The most advanced of them offer Application Programming Interface (API) access (e.g. BacDive). Among these sources, we selected a first core set of databases to populate the Omnicrobe knowledge base, to be extended in the future. These are GenBank, BacDive, and INRAE Biological Resource catalogs from the *International Center for Microbial Resources* (CIRM). These sources have been selected for either the quality of their data, their easy access, or their large coverage (section Information sources).

The standardization effort invested in these databases concentrates on taxa information rather than on habitats or phenotypes. The main microorganism references used are *The List of Prokaryotic names with Standing in Nomenclature* (LPSN; <https://www.bacterio.net/> [18] (e.g. in BacDive)), the *NCBI taxonomy* (e.g. in all NCBI databases and GOLD), and the *Integrated Taxonomic Information System* (ITIS) in EOL and GBIF. The taxonomic reference of Omnicrobe has a backbone based on the NCBI Taxonomy, with additional strains from BacDive and CIRM. We selected the NCBI taxonomy as a microbial taxonomic reference because it offers decent coverage of the living, taxa of a wide range of ranks including subspecific ranks and strains, and links to many sequence databases, which opens the way to genomic studies [19].

The information on the microbe isolation sites in databases and articles is expressed in free text fields in various languages. Habitat classification is rarely used to index them. Most habitat classifications used to index microbe isolation sites are non-aligned in-house classifications, notably those used by GOLD and BacDive. Isolation sources can be searched in BacDive with the *Microbial Isolation Source Ontology* (MISO), a three-level-controlled vocabulary of 376 terms that is employed to manually index BacDive isolation sources. GOLD classifies the ecosystems of organisms and samples using a five-level controlled vocabulary of 800 terms [13]. The *Earth Microbiome Project Ontology*, EMPO, defines 27 general classes that are relevant for studying correlation patterns between microbial sequences, environment, and environmental gradients at a very large scale [6], but are not sufficient to record the diversity of the microbial habitats for finer-grained studies. The EnvO ontology offers a larger coverage and a deeper structure for the controlled description of environment types but it is not dedicated to microbes; it is used in projects ranging from plants (e.g. Gramene data resource) to environmental features of some marine species of Tara Oceans expedition [20]. It evolved from the initial objective and is more generally concerned with environments as encountered in

ecological applications [21, 22]. Its generality and versatility may be an advantage for biodiversity and ecology research at a planetary scale, but it makes EnvO not as well-suited to focused domains such as microbiology research.

We chose the OntoBiotope ontology [12] for indexing Omnicrobe habitat information because of its focus on microbe habitats, its richness, and its previous uses for text indexing [9]. For phenotype indexing, we also preferred the OntoBiotope classification of microbial phenotypes over other controlled vocabularies, namely the Ontology of Microbial Phenotypes, OMP [23]. OMP focuses on phenotype change (presence, absence, alteration) but lacks some major phenotypes such as morphology phenotypes (e.g. colony color), energy sources beyond carbon and oxygen, or other environmental factors (e.g. response to various temperature scales).

In the next section, we describe the organization of the database, the controlled vocabulary, the information sources, and the text mining process. In the following sections, we present the architecture of the Omnicrobe application and the analysis workflow. We also present the user interface, the content of the current version, and the use of Omnicrobe information in the food innovation domain.

## Materials and methods

We present here the schema of the Omnicrobe database, the current data sources, and the controlled vocabulary that we use as references to index the data. We also present the text processing method that is applied to index the data sources with the controlled vocabulary.

### Omnicrobe database schema

The Omnicrobe schema is composed of entities of biological interest linked by specific relationships. Four types of entities are defined: *microorganisms*, *habitats*, *phenotypes*, and *uses*. They are linked by three types of relations: (i) the *lives\_in* relation links a microorganism to its habitat; (ii) the *exhibits* relation links a microorganism to its phenotype; and (iii) the *studied\_for* relation links a microorganism to its use. This formal schema structures the information in the Omnicrobe database, defines the relevant types of information to be extracted from the data sources, and guides the extraction process.

The addition of a supplementary structured source of information requires not only textual information processing of the content but also the alignment of the source schemata with the Omnicrobe data schema. Compared to free-text documents, the relationships between microorganism identifiers and their habitat, phenotype, and use descriptions can be more easily extracted from the semi-structured information of the databases. However, the level of information notably varies depending on the databases despite existing efforts on standard schemata for compiling biodiversity data such as the Darwin Core Standard [24] or RDA initiatives (<https://rd-alliance.org>). Some databases separate the habitats and the geographical location fields (e.g., Source and Country in GenBank), they also separate the microbe host and the host part as distinct habitats when appropriate (e.g., CIRM CFBP); some databases differentiate geographical information, between country and location (e.g., BacDive). It also happens that all this information is mixed in a single field and that the actual content of the database fields does not fully comply with the database schemata, mixing languages for instance. We have thus aligned each information source schema with the Omnicrobe data schema and implemented parsers so that the Omnicrobe update is fully automatic.

## Ontologies and taxonomies

The purpose of the Omnicrobe database is data linking and sharing. We have chosen standard controlled reference vocabularies to link microbial data collected from various sources and to make it more findable, accessible, interoperable, and reusable, following the FAIR principles.

The OntoBiotope ontology and the NCBI taxonomy have been selected for their relevance to the microbial domain, their lexical richness that makes text parsing more efficient, and their deep structure. Indeed, the deep hierarchical organization of Omnicrobe data enables both smoother browsing by non-specialist users and microbial distribution analysis at various scales.

**Habitats, phenotypes, and uses in OntoBiotope.** The OntoBiotope ontology (<http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE>) [12] contains 4,219 classes split into three branches, Habitat, Phenotype, and Use (Table 1). A wide range of microbial isolation sources can be found in the Habitat branch divided into 11 domains, and distributed along 13 levels reflecting the diversity of the microbial studies. The Phenotype branch mainly covers physiology, morphology, community behavior, and environmental interaction of various kinds of microbes, e.g., bacteria, fungi, and algae. The Use branch describes the uses and applications of microorganisms in 8 subtrees, i.e. antimicrobial activity, pathogenic activity, and metabolic activity. It focuses on food, food product quality, sensory quality, mixture transformation, and health properties. The food branch of OntoBiotope, inspired by the FoodEx2 product classification [25], has involved a community effort of INRAE microbiologists to extend it with fermented animal and plant products, cheese, cereal, and vegetable juices. These products are the subject of a new and very strong interest in research and innovation [26].

**Microbial taxa.** In order to detect and index microbial “Taxon” entities, we selected the NCBI Taxonomy (<https://www.ncbi.nlm.nih.gov/taxonomy>) as a reference. The NCBI Taxonomy organizes taxa according to state-of-the-art phylogeny. The NCBI Taxonomy consists of a classification of taxa in a taxonomic tree, and a nomenclature including valid scientific names, synonyms, and vernacular names, as needed to standardize taxa in publications even if their name has changed. The NCBI Taxonomy is regularly updated according to new requirements of NCBI databases. The NCBI Taxonomy encompasses all taxa of rank *species* or above, however, it only partially covers strains.

Our ambition is to gather information about microorganisms, which is a purely phenotypic notion covering organisms that require a microscope to see them. Therefore, there is no single common ancestor to all microorganisms. We selected a set of 23 high-level taxa (listed in S1 Table) that includes predominantly microscopic individuals, unicellular organisms (bacteria, archaea, and viruses), and pluricellular organisms (e.g. fungi, algae, nematodes). Omnicrobe data is indexed with all sub-taxa of this selection, which contains more than 730,000 species. In order to increase the strain coverage in Omnicrobe, we expand the NCBI Taxonomy with strains from the DSMZ catalog that are publicly available through the BacDive service.

To link each BacDive strain entry to an NCBI Taxonomy node, we automatically match the species and strain names provided by BacDive to taxon names in the NCBI Taxonomy. The objective is not only to gather more strains but also to place them correctly in the taxonomy and to record synonyms of strain names. The matching process considers several common variations of strain names. The matching has three outcomes:

**Table 1. Number of classes per branch in the OntoBiotope ontology.**

Habitat	Phenotype	Use
3,731	434	66

<https://doi.org/10.1371/journal.pone.0272473.t001>

- if one of the names of the BacDive entry or one of its variations is equal to an NCBI Taxonomy strain name, then we add new synonyms to the NCBI strain, and consider the identifiers of the NCBI Taxonomy strain and of the BacDive entry to be equivalent;
- if the BacDive entry does not match any NCBI Taxonomy strain, but we can identify the species or genus that the strain belongs to, then we add a new node to the taxonomy;
- if the BacDive entry does not match any NCBI Taxonomy strain, nor can we identify which species or genus it belongs to, then we leave this entry out of the reference.

In the current version, the matching process provides 79,041 additional nodes and 796,919 additional synonyms. The Omnicrobe taxonomic reference includes 1,102,673 nodes and 5,862,677 names, including scientific names, vernacular names, and catalog names for strains. The matching code is Open Source and publicly available at <https://forgemia.inrae.fr/omnicrobe/extended-microorganisms-taxonomy>.

### Information sources

We designed Omnicrobe database to aggregate information extracted from various sources. The current version includes information from the bibliographic database PubMed (<https://pubmed.ncbi.nlm.nih.gov>), from the nucleotide database GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>), and from four microbial resource center catalogs: BacDive (<https://bacdive.dsmz.de>) and three catalogs of CIRM (*International Center for Microbial Resources*) on food bacteria, plant pathogens, and yeast. CIRM BIA (for food bacteria) is available at <https://collection-cirmbia.fr>, CIRM-CFBP (for plant pathogens) at <https://cirm-cfbp.fr>, and CIRM-Levures (for yeasts) at <https://cirm-levures.bio-aware.com>. As stressed above, we have chosen these sources according to their richness, quality, popularity, open access license, and accessibility. Omnicrobe also records the Qualified Presumption of Safety (QPS) of biological agents added to food as maintained by the EFSA (<https://www.efsa.europa.eu/en/topics/topic/qualified-presumption-safety-qps>).

The information on microorganisms and habitats is extracted by the Omnicrobe application from all the sources, while the information on phenotypes and uses is extracted from Pubmed only.

**PubMed.** The open access bibliographic database PubMed contains more than 33 million citations and abstracts from the biomedical literature maintained by the NCBI and the NLM. Omnicrobe uses a thematic subcorpus of the PubMed references that mention at least one microbe taxon from the Omnicrobe taxonomic reference (section Ontologies and taxonomies). To identify those references, we use an alignment of the Omnicrobe taxonomic reference and the Organisms [B] subtree of the MeSH (Medical Subject Headings) thesaurus that indexes PubMed references. MeSH is available at <https://meshb-prev.nlm.nih.gov/search>.

At the time of publication of this paper, the Omnicrobe corpus comprises around 2,870,000 PubMed references.

**GenBank.** The GenBank sequence database maintained by the NCBI is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.

We feed the Omnicrobe taxon field with the parsed content of the “organism” and “strain” fields and the Omnicrobe habitat field with the “isolation\_source” field from GenBank releases. The current release (release 249) contains around 5,100 sequence records in the GenBank flat file format. Among these sequences, we select only 16S gene rRNA sequences because they are generally used to identify bacterial species [27]. The isolation source of the sample is often described by the optional GenBank isolation\_source field [14]. We set the minimum

sequence size to 800 base pairs to ensure the quality of species identification and we use the NCBI Taxonomy part of the Omnicrobe taxonomic reference to filter the relevant GenBank taxa.

**BacDive DSMZ.** BacDive is a web service that gives access to the DSMZ catalog. In 2022, the DSMZ collection maintains more than 80,000 bacterial and archeal strains, including type strains, and is growing fast. The BacDive API returns detailed metadata for each entry such as taxonomy, morphology, physiology, environment, and molecular biology. Most of the BacDive data is manually curated. The current version of Omnicrobe covers BacDive taxon and habitat information. We feed the Omnicrobe taxon field with the parsed content of the BacDive fields “Full Scientific Name”, “Strain Designation” and “Culture col. no.”, and the habitat field with “Sample type/isolated from” information. BacDive also offers rich information on phenotypes that we plan to process in future work.

**CIRM.** The *International Center for Microbial Resources* (CIRM), managed by INRAE, preserves more than 15,000 strains of bacteria and yeasts among which a subpart is publicly available. Omnicrobe integrates data from three catalogs: CIRM-BIA, dedicated to bacteria of food interest, CIRM-Levures, dedicated to traditional French ferments and yeasts involved in biotechnologies, and CIRM-CFBP on plant-associated bacteria.

Alignments of CIRM database fields to Omnicrobe fields are given in the [S2 Table](#). The upcoming availability of CIRM information through the Microbial Resource Research Infrastructure (MIRRI; [https://catalog.mirri.org/page/Strains\\_catalog\\_query](https://catalog.mirri.org/page/Strains_catalog_query)) will provide a single-point and unified access in the near future.

**Volume of Omnicrobe sources.** [Table 2](#) gives the figures of the current version of Omnicrobe sources, with the last revision dates in brackets. [S1 Text](#) gives the queries that are used to gather data from PubMed and GenBank sources.

These seven sources are the first to be processed and integrated into Omnicrobe but the architecture allows for the inclusion of new sources.

## Text mining process

A significant part of the information contained in the above-mentioned sources is unstructured and expressed in natural language, either in scientific abstracts (PubMed) or in free-text fields of the databases (GenBank, BacDive DSMZ, CIRM). The text mining process aims to automatically extract this information and structure it according to the Omnicrobe data schema to populate the Omnicrobe database.

The text-mining process relies on Natural Language Processing (NLP) techniques and combines rules, lexical resources, and linguistic analysis to detect relevant information in texts [28]. The general process consists of three main steps, namely entity recognition, entity

**Table 2. Omnicrobe sources, data volumes, types, and extraction dates in the May 2022 version.**

Source	Volume and types of data [date]
PubMed	2.8 million abstracts from articles concerning microbes [2022–04]
GenBank	492,031 entries extracted from the Genbank database. The entries record information about species, strains, isolation sources, and hosts [2022–04]
BacDive DSMZ	28,779 entries recording taxa and habitats [2018–01]
CIRM-BIA	1,726 entries recording taxa, strains, and habitats [2022–04]
CIRM-Levures	2,158 entries recording taxa, strains, and habitats [2021–08]
CIRM-CFBP	6,988 entries recording taxa, strains, and habitats [2022–04]
QPS species of the EFSA	788 bacteria with Qualified Presumption of Safety (QPS) status [2022–01]

<https://doi.org/10.1371/journal.pone.0272473.t002>



**Entity recognition and normalization methods.** We designed methods that combine entity recognition and normalization. Compared to other types of entities, microorganism names in scientific text are more consistent and exhibit less variation. We handle them with a classic lexicon-and-rule-based approach. That is, we directly map textual entity mentions to entries of our extended taxonomy. The taxonomy includes both canonical names and synonyms of taxa, as well as common variations generated using rules. For instance, the first word of *Bacillus subtilis* is abbreviated and the two following variants are generated: *B. subtilis* and *B subtilis*. We also designed additional rules to recognize strain names, as the strain level is highly relevant for many studies [29]. Recognized microorganism mentions are assigned the identifier of the taxon to which they were mapped.

Habitats and phenotypes show more diversity in the way they are expressed in text. According to [9], respectively 72.5% and 91.2% of text mentions do not directly match a concept label of the OntoBiotope ontology. Thus, we use an approach that relies on deeper linguistic analysis and a more flexible matching process. We start by extracting all noun and adjectival phrases using an existing tool, the YaTeA term extractor [30]. YaTeA also analyzes the syntactic structure of the extracted terms, detecting the “head” of each phrase (*i.e.*, the main word, such as the word “milk” in the phrase “dairy cow milk”). Then, our mapping method ToMap tries to match candidate terms with concepts from the OntoBiotope ontology, based on structural and lexical similarity [31]. ToMap performs a first mapping between terms and concept labels that have the same syntactic head, then it computes a similarity score (the Jaccard index) for each pair based on the number of words they have in common. Finally, the term-concept pair with the highest score is selected. If a candidate term cannot be mapped in the first step (*i.e.*, it does not share the same head as any concept label), then the term is ignored. This process is further refined with hand-designed heuristics to handle a number of recurring ambiguous cases and systematic errors.

Recognition of the last type of entities—use entities—relies on a simple lexicon-based method that directly maps text with concept labels from the microbial use branch of the OntoBiotope ontology. This part of the text-mining process is still a work in progress and we plan to refine it in future work.

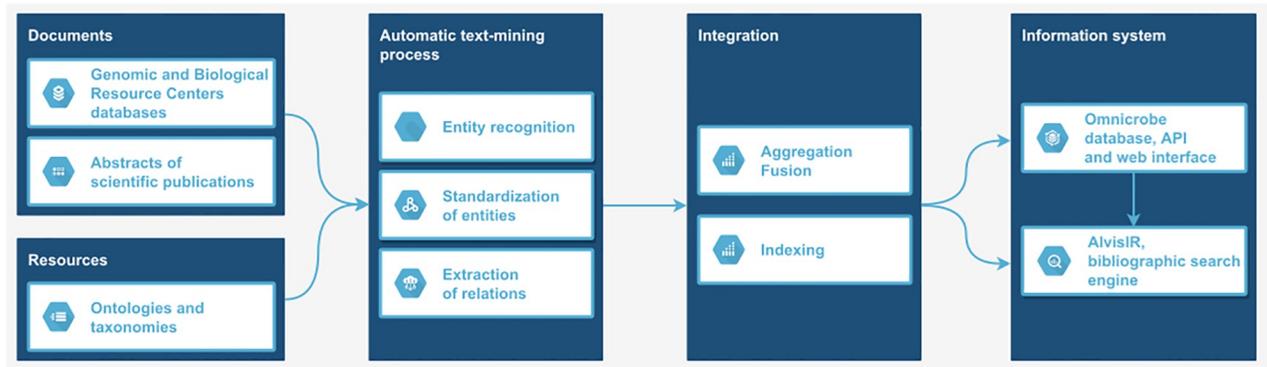
**Relation extraction method.** We extract relations between recognized entities based on textual proximity and lexical clues. More specifically, we link two entities when (1) they occur in the same sentence and (2) the sentence contains trigger words, *i.e.* expressions that are potential indicators of a relation between two entities (*e.g.*, “isolated from” is a trigger for the `lives_in` relation between a microorganism and a habitat). In addition, we also perform anaphora detection for microorganisms to handle cases where entities occur in different sentences. Anaphora are textual expressions used to refer to an entity previously mentioned in the text, *e.g.* pronouns (“it”) or general terms (“this bacteria”).

## Results

We designed workflows (section Workflows) to automatically gather, analyze and combine the microbial information in the Omnicrobe Information System (sections User web interface and Application programming interface). Section Omnicrobe content details its current content. The use of Omnicrobe information for food innovation (section Fermentation of soy milk use case) illustrates how relevant data can be retrieved in a few queries.

## Information system

As shown in Fig 2, the data from documents are mapped to the Omnicrobe schema and processed by the automatic text-mining workflow. Then this data is integrated into an



**Fig 2. Information system flowchart.**

<https://doi.org/10.1371/journal.pone.0272473.g002>

information system and accessed through a web interface and an application programming interface.

## Workflows

The different processes to generate and integrate data into the Omnicrobe application are implemented into dedicated workflows. These workflows facilitate the management and orchestration of the processing, update, and evolution of the Omnicrobe data.

For integrating the heterogeneous processing steps and automating their execution we use Snakemake [32], a workflow management system popular in bioinformatics. Snakemake manages the tool dependencies and orchestrates the execution of the full workflows in cluster-based environments.

As depicted in Fig 2, several main steps are required to consume the source data and to automatically extract and integrate the results into the information system (i.e. annotations about *microorganisms*, *habitats*, *phenotypes*, and *uses* and relations between them). These steps are materialized into the following workflows that process and integrate data from each source (<https://forgemia.inrae.fr/omnicrobe/text-mining-workflow>):

- **Data collection and preprocessing workflow:** abstracts from PubMed and data from the Genetic and Biological resources centers (GenBank, BacDive, and CIRM) are collected and pre-processed by separate workflows. The ontologies and taxonomies are also preprocessed by selecting and filtering the useful parts.
- **Text mining workflow:** this workflow <https://forgemia.inrae.fr/omnicrobe/text-mining-workflow> implements the text mining process described in the Materials and Methods section (section Text mining process). It ensures the automatic extraction of information from the data sources (PubMed, GenBank, BacDive, and CIRMs). The main text mining steps are implemented using the AlvisNLP text-mining pipeline (<https://github.com/Bibliome/alvisnlp>). External and third-party tools are used for data pre- and post-processing, i.e., input and output formatting, and indexing. The last step of the workflow merges information occurring more than once in the same information source.
- **Data integration workflow:** the data integration workflow (<https://forgemia.inrae.fr/omnicrobe/omnicrobe-database>) integrates the data generated by the text-mining workflows into a PostgreSQL relational database. The database indexes the data by the controlled

reference vocabulary through ‘hierarchical paths’. It means that each entity is indexed by its specific class, and all more general classes of the controlled reference vocabulary.

The Omnicrobe database is regularly updated in order to continuously improve the quality of the data. We continuously enhance the text mining methods according to feedback on the quality of the results and errors observed. We regularly update the information sources and integrate any changes from ancillary resources (e.g., OntoBiotope, Omnicrobe taxonomic reference) when revised versions of these resources are released.

The Omnicrobe database has been updated more than 10 times since 2018. Initially the project covered microorganisms and habitats, and then was expanded to cover phenotypes and uses. The input data collected from the different sources has also been notably enriched since 2018 with newly published articles from PubMed and the additional BRC.

## User web interface

The Omnicrobe data can be queried through a public web interface (<https://omnicrobe.migale.inrae.fr/>). It is designed to help microbiologists find information related to their scientific interests. We consulted microbiologists to choose the navigation tools in order to help them make the most of this interface.

Fig 3 shows the main frames of the Omnicrobe web interface. The tabbed navigation (panel A) makes it quick and easy to select the type of search a user wants. For example, users looking for places where a microbe or a family of microorganisms lives, will use the tab “Taxon lives in habitat”. An advanced search functionality is also available and allows advanced users to perform combined multi-criteria searches, for example searching for specific phenotypes of organisms living in a given habitat.

Searches (panel B) are carried out using the terms of habitats, phenotypes, and uses of the classes and synonyms of the OntoBiotope ontology, and the name of taxon classes of the Omnicrobe taxonomic reference. According to the needs, queries can be expressed at different levels of generality, from the very specific (e.g., strain, given local specific cheese) to the very general (taxon order, food). In order to optimize searches and avoid any input errors, input

The screenshot displays the Omnicrobe web interface. On the left, the 'OntoBiotope navigation' menu is visible, with 'cheese' selected under the 'food' category. The main content area shows the 'Habitat contains Taxon' search results for 'cheese'. The interface includes a search bar, filters for sources and taxon, and a table of results. The table columns are Source text, OntoBiotope class, Relation type, Scientific name of taxon, QPS, and Source. The results show entries for 'queso fresco' and 'blue veined cheese' containing '[Brevibacterium] frigoritolerans' and '[Candida] cabralensis'.

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

**Fig 3. Omnicrobe web interface.** Republished from <http://omnicrobe.migale.inrae.fr> under a CC BY license, with permission from INRAE MaIAGE, original copyright 2022.

<https://doi.org/10.1371/journal.pone.0272473.g003>

fields with automatic completion are proposed. Depending on the type of search, several filters are available, for example, filtering according to the source of the information (e.g., CIR-M-BIA) or food safety criteria, i.e., QPS status.

To express his/her query, the user can also navigate in the hierarchical structure of Onto-Biotope from the Omnicrobe interface (panel C) and select the class of interest among habitats, phenotypes, or uses.

The results (panel D) are displayed in columns for better readability. A link to PubMed and Genbank original text is provided. PubMed references are displayed by the AlvisIR semantic search engine. AlvisIR allows a quick reading of the text thanks to the highlighting of the entities and relations to assess the relevance of the information. Chaix et al. [28] give a detailed presentation of the AlvisIR interface and examples of its use.

The Advanced search option of the Search menu is intended for the expression of complex embedded queries that combine “and” and “or” operators. Selected columns of query results obtained from dedicated panels and Advanced search can be exported in various formats (*i.e.*, CSV, Excel, and PDF) for further processing by users who are not familiar with API.

Omnicrobe runs on an Apache web server. It is written in HTML5/Javascript (client-side) and Python with the Flask web framework (server-side). Using AJAX reduces response time by minimizing data transfers: it sends simultaneous server requests and takes advantage of the processing capability of the clients.

## Application programming interface

All the data in the Omnicrobe database can be accessed through an Application Programming Interface (API). The API provides the same capability as the web interface. It allows the search for taxa, habitats, and phenotypes, as well as relations between these entities. The main difference is that responses are computer-readable. Thus, the API is suitable to embed Omnicrobe data into another information system. The API is developed with the Python Flask-RESTX framework. The main entry point is available at <https://omnicrobe.migale.inrae.fr/api> and the documentation at <https://omnicrobe.migale.inrae.fr/api-doc>.

## Omnicrobe content

In this section, we give an overview of the content of Omnicrobe in its current update (May 2022). The Omnicrobe database content reflects the focus of microbial research studies rather than the worldwide distribution of microbes with respect to their habitat, phenotype, or use. The following descriptive statistics reveal the domains where large sets of scientific knowledge have been gathered, and highlight potential gaps that need further attention.

Table 3 shows the number of relations extracted per source after removing duplicates. PubMed is the only source of Omnicrobe for Taxon-Phenotype and Taxon-Use relationships;

**Table 3. Number of distinct relationships extracted per source in the May 2022 version.**

Source	Relationship		
	Taxon-Habitat	Taxon-Phenotype	Taxon-Use
GenBank	259,745	-	-
BacDive	35,696	-	-
CIRM BIA	604	-	-
CIRM Levure	943	-	-
CIRM CFBP	1,109	-	-
PubMed	721,244	50,410	16,007

<https://doi.org/10.1371/journal.pone.0272473.t003>

Integrating Taxon-Phenotype relationships from databases is reserved for future work. PubMed is also the most prolific source for Taxon-Habitat relationships, followed by GenBank. This is expected since PubMed is also the largest source of Omnicrobe.

[S4 Table](#) lists the most frequent taxa, habitats, phenotypes, and uses involved in relationships obtained from PubMed. This ranking reflects the most investigated subjects in papers indexed by PubMed. As expected, they are related to human health: most taxa are pathogens, such as *Staphylococcus aureus*, and a large part of habitats are humans or part of humans, e.g. *patient*, *blood*, or *respiratory tract*. Phenotypes and uses are related to pathogenicity, resistance, and prevention. The *pathogen* phenotype and *health risk* use are both representative examples. There is also a substantial focus on fundamental biology (e.g., *cell*, *mouse*). Nevertheless, many other topics are being investigated, such as environmental microorganisms or food contamination (e.g., *plant*, *spoilage*).

[Fig 4](#) shows the distribution of taxa ranks. The high number of species and subspecific taxa answers critical needs in microbiology research and is a strong point of the Omnicrobe database.

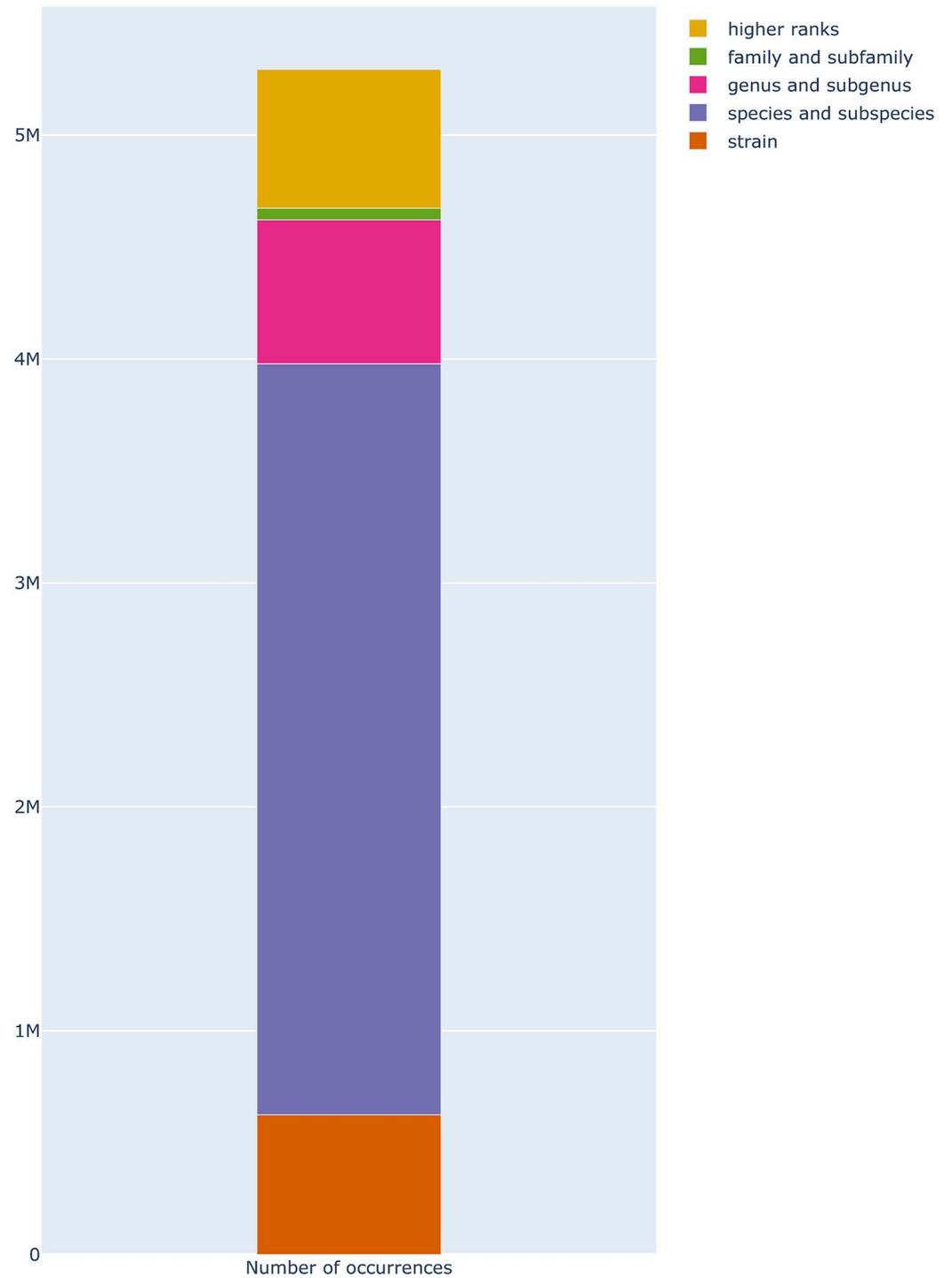
[Fig 5](#) shows the distribution of microbe taxa in *Lives\_In* relations. As expected, *Bacteria*, *Viruses*, and *Fungi* are the most frequent taxa.

The distribution of habitats at the four highest levels is shown in [Fig 6](#). The most frequent habitats are related to the biomedical domain that represents the bulk of Omnicrobe information. They are followed by engineered (industrial, agricultural, food) environments, non-human hosts, and environmental habitats. The distribution is similar to the one shown in the isolation sites distribution chart of BacDive (<https://bacdive.dsmz.de/dashboard>) [16].

[Fig 7](#) shows the proportion of *Lives\_In* relations from each source that is also extracted from PubMed abstracts. One might think that the sheer volume of information extracted from PubMed would render the other sources redundant, but this figure indicates the contrary. The overall intersection between PubMed and the other sources is rather low. This demonstrates the complementarity between different sources and that the integration of different sources in Omnicrobe provides more comprehensive information on microbe habitats.

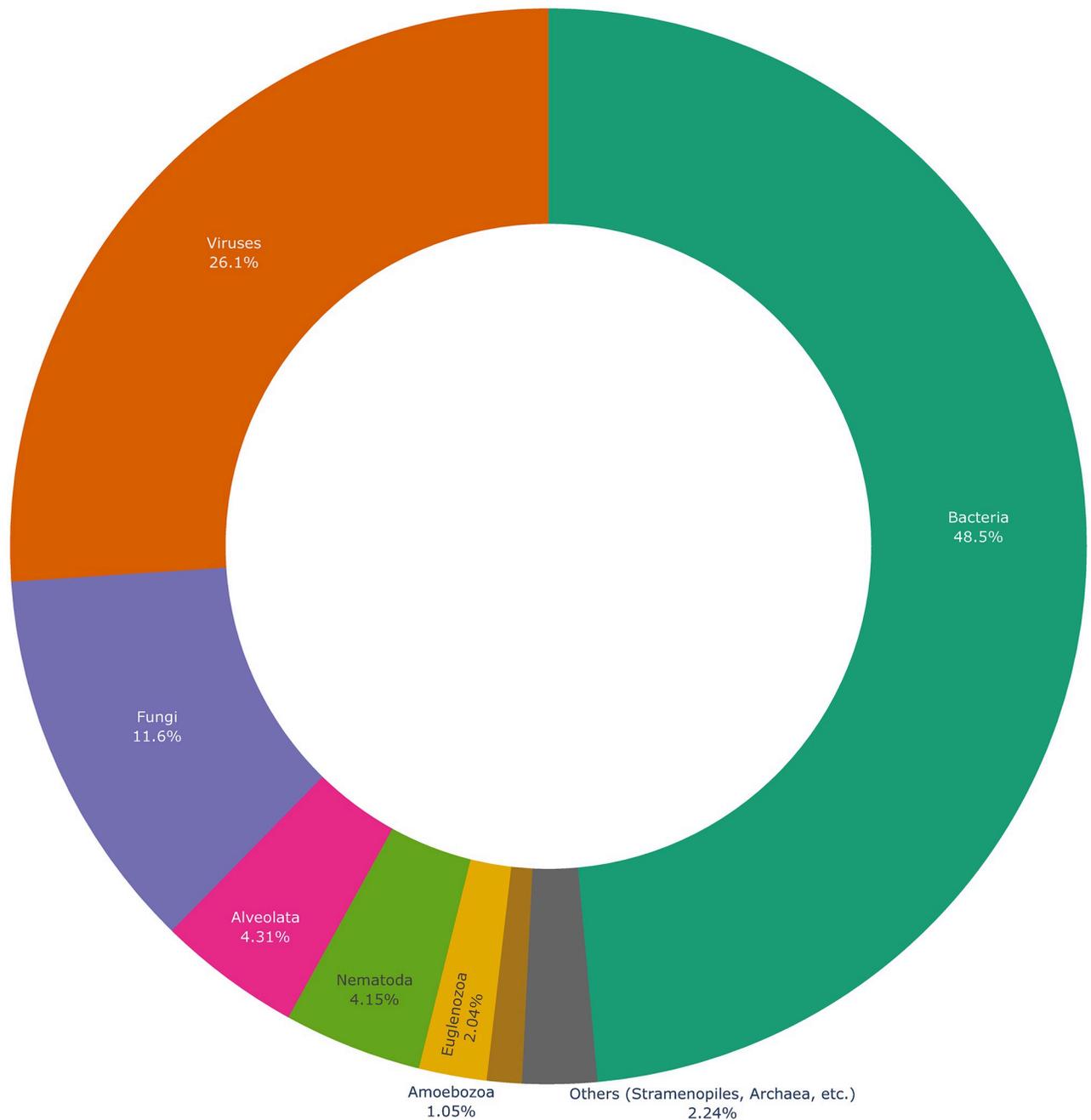
[Fig 8](#) compares, for the most frequent habitats, the number of relations in which they occur and the number of different microorganism taxa to which they are linked by a relation. In other words, the frequency of each habitat is contrasted with the diversity of microorganisms that inhabit them. Diversity is not necessarily correlated to frequency. For instance, environment and plant habitats (e.g. *plant*, *yeast*, *water*, *soil*, *marine environment*) display more relative diversity than human and health-related habitats (e.g. *human*, *patient*, *hospital*). Indeed, studies on humans focus on the narrow range of pathogenic microorganisms, whereas studies on plants and environments focus on the biodiversity hosted by these habitats.

In order to build an indicator of Omnicrobe consistency, we focused on the six phenotype classes that define the temperature tropism of microorganisms. The classes form a gradient of preferred temperature that ranges from cryophile to extreme thermophile. In this gradient, two phenotypes next to one another (e.g. *thermophile* and *mesophile*) are conceivably compatible. On the other hand, two phenotypes further apart from each other (e.g., *hyperthermophile* and *psychrophile*) are incompatible with the same microorganism. Indeed, it is unlikely that an organism has two very distant optimal growth temperatures. [Fig 9](#) displays the matrix of correlation between all the pairs of the temperature tropism phenotypes. As expected, the diagonal is the most intense and the intensity decreases with the distance along the temperature range. This shows that incompatible temperature-based phenotypes are less likely predicted than compatible phenotypes.



**Fig 4. Distribution of taxa ranks in Lives\_In relations in Omnicrobe.** “Strain” ranks comprise strains and isolates. “Species and subspecies” ranks include species and ranks below species and above strain (*e.g.*, subspecies, varieties, morph). “Genus and subgenus” ranks include genus and ranks below genus and above species (*e.g.*, subgenus, section, series). “Family and subfamily” ranks consist of family and ranks below family and above genus (*e.g.*, subfamily, tribe). “Higher ranks” include all ranks above the family (*e.g.* order, class, phylum, kingdom). The height of the bars is proportional to the number of *Lives\_In* relations in Omnicrobe.

<https://doi.org/10.1371/journal.pone.0272473.g004>

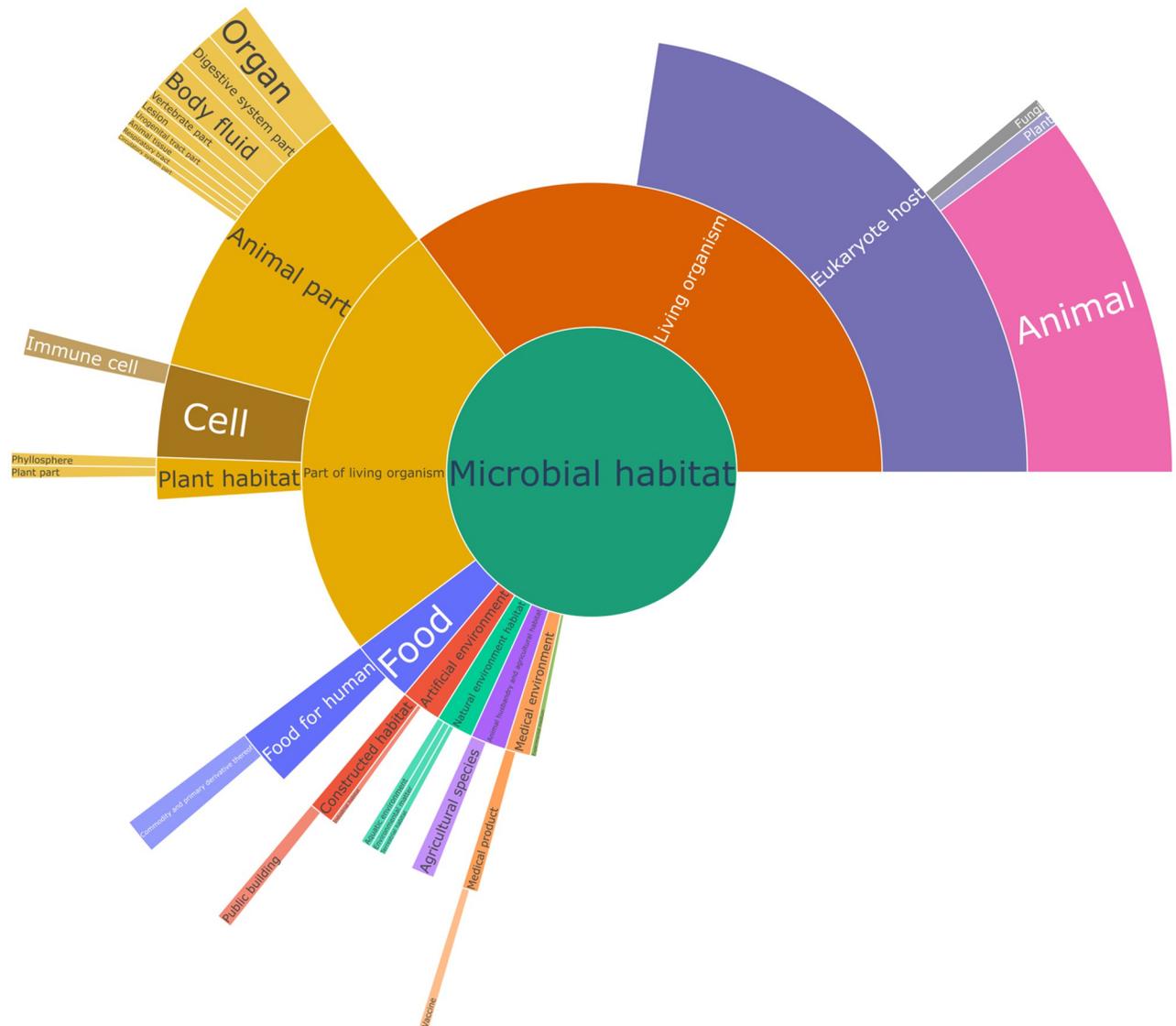


**Fig 5. Distribution of microbe taxa in Lives\_In relations extracted from PubMed in Omnicrobe.** The taxa represented in this chart are taxon roots selected as microorganisms in Omnicrobe (see section Ontologies and taxonomies). The arc is proportional to the number of Lives\_In relations that involve the taxon or any descendant. “Others” include taxa that account for less than 1% of relations: *Archae*, *Chlamydomonadales*, *Chlorella*, *Choanoflagellida*, *Cryptophyta*, *Desmidiiales*, *Diplomonadida*, *Glaucocestophyceae*, *Haptophyta*, *Ichthyosporea*, *Oxymonadida*, *Parabasalia*, *Prototheca*, *Retortamonadidae*, *Rhizaria*.

<https://doi.org/10.1371/journal.pone.0272473.g005>

### Fermentation of soy milk use case

Here, we present an example of Omnicrobe application for applied research in food innovation. It illustrates both the effective retrieval of relevant information in a short time and its use for further biological development.



**Fig 6. Distribution of habitats in Lives\_In relations extracted from PubMed.** This chart represents the habitats at the four highest levels in the OntoBiotopie ontology. The arc is proportional to the number of Lives\_In relations extracted from PubMed that involve the habitat or any descendant in OntoBiotopie. Only habitats with more than 20,000 occurrences are shown for readability.

<https://doi.org/10.1371/journal.pone.0272473.g006>

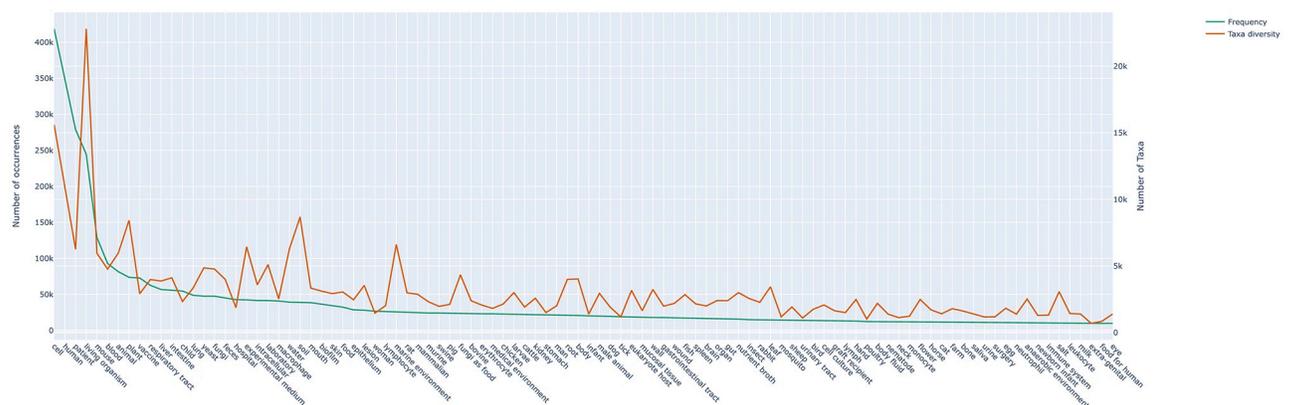
**Use case aim.** This work aims to create a novel fermented food product based on the fermentation of soy milk [33]. Soymilk represents an interesting alternative to animal milk as a sustainable food. It could also be a valuable protein source for lactose-intolerant and vegan populations. Lactic fermentation of soy juice by lactic acid bacteria (LAB) to produce a yogurt-type fermented soy product can contribute to improving the organoleptic properties of soy juice by reducing “off-flavors” and to lowering the content of non-digestible oligosaccharides. For these reasons, soymilk fermentation attracts recent interest [34].

For this purpose, we would like to identify bacterial species that exhibit the desired properties. We first search the literature for relevant candidate strains. Once a subset of relevant strains is selected, their *in vitro* cultivation and screening for acidification in soy juice requires the strains to be available in BRC catalogs for ordering. CIRM-BIA dedicated to bacteria of



**Fig 7. Proportion of taxa-habitat relations in each source, which are also extracted from PubMed.** The height of the bar represents the proportion of relations per source that were also extracted from PubMed. For instance, only 10% of relations in GenBank were also extracted from PubMed (the same taxon-habitat pair), leaving 90% of relations exclusive to GenBank.

<https://doi.org/10.1371/journal.pone.0272473.g007>



**Fig 8. Frequency of habitats and number of different taxa to which they are linked.** The green line (left scale) represents the number of *Lives\_In* relations extracted from PubMed that involves each of the 100 most frequent habitats. The brown line (right scale) represents the number of distinct taxa to which each habitat is linked with *Lives\_In* relations extracted from PubMed.

<https://doi.org/10.1371/journal.pone.0272473.g008>



**Fig 9. Correlation between temperature tropism phenotypes in Omnicrobe.** Each box represents the intersection between the sets of taxa to which the two phenotypes are linked with Exhibits relations in Omnicrobe. The color intensity indicates the Jaccard index between the sets of taxa.

<https://doi.org/10.1371/journal.pone.0272473.g009>

food interest contains more than 4,000 different strains of lactic and propionic acid bacteria and is thus our primary source of strains for this study.

**Searching Omnicrobe.** Classical bibliographic search is complex due to the high number of previous studies and the distribution of the information per bacteria in a high number of publications, books, and websites. We used the Omnicrobe interface to express the combination of criteria that the candidate bacteria have to meet. The targeted properties and their translation into Omnicrobe criteria were defined as follows:

- *targeted bacterial species have been previously reported as detected in soy milk* is translated as “soy milk” habitat value
- *they can perform acidification* is translated as “acidification” use value
- *at medium or warm temperature* is translated as “mesophile” or “thermophile” phenotype values
- and *they are safe for human food consumption* is translated as QPS criteria set to yes.
- The Advanced search of the Search menu was used for the expression of these complex embedded queries by combining “and” and “or” operators as shown in Fig 10.

The resulting list of 103 taxa (20 species, 15 subspecies, and 68 strain names) with all requested properties was retrieved from the Omnicrobe interface in spreadsheet format. The file includes links and IDs of PubMed and GenBank source documents that allow the biologist

The screenshot shows a query builder interface with a main query and an embedded query. The main query is set to 'AND' and contains four rules: Taxon = Bacteria, Habitat = soy milk, Use = acidification, and QPS = only QPS (Qualified presumption of safety). The embedded query is set to 'OR' and contains two rules: Phenotype = thermophile and Phenotype = mesophile. Each rule has a 'Delete' button. There are also '+ Add rule' and '+ Add group' buttons for both the main and embedded queries. An 'Apply' button is located at the bottom left of the interface.

**Fig 10. An example of complex embedded queries.** These queries are used to retrieve mesophilic or thermophilic bacteria present in soy milk and capable of acidification, and with a qualified presumption of safety.

<https://doi.org/10.1371/journal.pone.0272473.g010>

to check by herself the accuracy of the text mining prediction. The information is spread across 501 documents. For example, *L. acidophilus* has acidification properties according to PubMed ID: 27384493, 9839223, 22264421, 30776138, 9633657, 15591363, and is known to be present in soy milk and soy milk yogurt according to PubMed ID: 18541163, 20477889, 16943081, 28985138, 21775184, 29656125, and is thermophilic according to PubMed ID: 9057296, 1115513.

The analysis of the source texts reveals that two of the retrieved species are reported as both thermophile and mesophile: *Lactiplantibacillus plantarum* and *Streptococcus thermophilus*. In the corresponding articles the enumeration of species and phenotypes, and the mixture of species mentions make difficult the extraction of the relationships by automatic text mining. An example is “A mesophilic-thermophilic mixed culture of *Lactococcus lactis ssp. lactis*, *Lactococcus lactis ssp. cremoris*, and *Streptococcus thermophilus* was also used.” where *Streptococcus thermophilus* was incorrectly classified as mesophilic by the text mining process. We discarded the *Bifidobacterium longum* species because it was considered as an ‘obligate anaerobe’ and incompatible with food fermentation conditions.

We aggregated the relationship results at the species level. The limited number of species in our study makes manual checking feasible. One additional constraint was the availability of the strains of the selected species in the CIRM-BIA collection (Rennes, France). The eight selected species that meet all requirements were *Lactobacillus acidophilus*, *Lactobacillus delbrueckii*, *Lactobacillus helveticus*, *Lactococcus lactis*, *Lactiplantibacillus plantarum*, *Streptococcus thermophilus*, *Lacticaseibacillus casei*, *Lacticaseibacillus paracasei*.

**Fermentation results.** We selected 206 strains of the eight species from the CIRM-BIA collection to be tested (S5 Table). The fermentation of the strains was performed in glass vessels with inoculation of 1% in the soy juice and a fermentation time of 48 h. Two different temperatures of fermentation were chosen: 30 °C for mesophilic species and 43 °C for thermophilic species. *Streptococcus thermophilus* was cultivated at 43 °C and *L. plantarum* was cultivated at 30 °C. The details of the experiment can be found in [33]. Fermentation success was checked by pH monitoring after 48 hours of fermentation. The pH of soy milk was 7.2 at the time of inoculation. A pH below six means that an acidification process occurred.

Among selected strains, 148 strains succeeded to acidify soy juice in 48 h (S5 Table). All *S. thermophilus* strains acidified soy juice while none of the *L. helveticus* strains acidified soy juice. Acidification is strain-dependent for *L. acidophilus*, *L. delbrueckii*, *L. plantarum*, *L. casei*, *L. paracasei* and *L. lactis*. Except for *L. helveticus*, the pre-screening of relevant strains by using the Omnicrobe database was efficient, cost and time effective. The microbiologist saved significant time in bibliographical search and wet lab experiments because she could quickly focus on a relevant subset of species among potential candidates.

## Discussion and conclusion

We presented the Omnicrobe online application, with its unique database of information on microbe habitats and phenotypes. The whole framework was developed to automate as much as possible the update of the database content according to the evolution of the original sources and reference vocabulary and the improvement of the text mining process. Specifically, the annotation process of the textual source data and the indexing by standard metadata are fully automatic.

Other projects aim at aggregating microbe biodiversity data. The Prego database [35] is the most similar to Omnicrobe in its scope and goal, although it differs in several aspects. As Omnicrobe, Prego gathers information about microbes and their environments computed from textual data. Prego indexes taxa and environments using the NCBI Taxonomy and EnvO respectively. Additionally, Prego includes molecular processes using the Gene Ontology (GO) whereas Omnicrobe contains information about microbial phenotypes and uses. With regards to microbe-habitat information, which is common to both databases, Omnicrobe is significantly more comprehensive than Prego. Prego contains fewer habitat categories than Omnicrobe (1,000 versus 4,000), and Omnicrobe contains twice as many relationships between microbes and their environments (700,000 compared to 300,000).

As with any resource automatically computed from textual data, the quality of Omnicrobe information depends on the quality and richness of the primary sources and on the performance of the text mining process. We noticed that frequent errors of normalization were due to ambiguities as illustrated by S3 Table examples. They are currently handled by applying hand-coded rules that use the ambiguous word context. Long distance relationship (see S3 Table for examples) is also a source of errors in relation detection. Future work includes ambiguity resolution and relation detection by using machine learning methods trained on the Bacteria Biotope dataset [9].

The extension of Omnicrobe to newly available data and sources is straightforward as long as the semantic interoperability of the data is ensured. The increasing standardization of database metadata and the use of controlled vocabulary are heading in the right direction. The next step is to extend Omnicrobe to include other information available in the current processed sources. For example, work is underway on BacDive DSMZ data to extract other types of data than habitat data. Geographical information is one type of data that will require dedicated text mining analysis to ensure semantic interoperability since the lack of standardization allows mixing in the same record different types of places that should be distinguished, e.g. address, landscape, country. Consideration is also being given to integrating data from GOLD. The scope of the Use part of the database focuses on food application but will be extended to other domains in future versions, among which are biotechnology and sewage treatment. We are considering extending scientific literature sources to full-texts. We have compared the information present in the abstracts to the information of full papers of a sample of the *International Journal of Systematic and Evolutionary Microbiology*, which is the reference journal for the description of new prokaryote species and thus particularly rich in microbe habitats

and phenotypes. The habitat and phenotype mentioned in the full-text are more precise but not significantly more diverse than in the abstract. Accessing and preprocessing full-text documents remain challenging because of the diversity of access protocols and formats.

We presented the Omnicrobe user interface and an example of its use for food fermentation studies. The complementary display by the semantic search engine AlvisIR of the source text and its semantic annotation has proved very useful in checking the quality of the predicted result and putting it into context. The food fermentation use case also confirms the relevance of expressing a complex combination of criteria by a query that combines variables and operators. This advanced search allows users to express complex queries while requiring less technical skills than the API.

Beyond the inventory of the observations of habitats and phenotypes of microbes offered by Omnicrobe, we believe that it can contribute to hypothesizing microbe spread scenarios, to anticipating new disease spread and enabling appropriate control procedures, while conversely supporting positive flora studies. The 'One Health' concept with the interdependence of human, animal, plant, and environmental health has stressed the role of new contamination paths that involve intermediate hosts like wildlife, insect vectors, water, and air [36]. The microbial ecology of natural environments rapidly evolves as a consequence of anthropogenic activities, e.g., climate change, deforestation, water pollution [37], increasing human-wildlife interactions, and globalization of plant, and animal trade. Predicting the ability of given microorganisms to grow in habitats where they have not been observed strongly depends on the knowledge of their phenotypes and of the connection of their known habitats including vectors and dissemination pathways.

For instance, plant health management aims at anticipating epidemiological risk of outbreak in crops. An important challenge is that many plant pathogens do not have obligate dependence on plant hosts, furthermore biological interactions between pathogen agents and their vectors, and hosts varies widely in specificity, ranging from a few to thousands of different species (see, e.g. *Xylella fastidiosa* or *Pseudomonas syringae*). Thus health management professionals need to discover unsuspected reservoirs of pathogens of wild or cultivated plants, or plant disease vectors. We measured the richness of plant pathogen habitat inventory in Omnicrobe to evaluate how it could contribute to the identification of potential reservoirs. Omnicrobe currently contains more than 110,000 habitat relationships between plants or plant parts, among which 82,000 microbes are pathogens. It also contains 94,000 microbes which have both insect and plant hosts. We believe that this large-scale survey is a valuable first step in the development of new tools for epidemiological risk estimation and for surveillance and management of plant diseases. Currently our research in plant health on Omnicrobe focuses on how Omnicrobe information could feed into prediction models such as connectivity models that combine biophysical data, spatial data, or air circulation information to predict pathogen dispersal based on knowledge of connected habitats [38].

## Supporting information

**S1 Table. Root taxa of microorganisms included in Omnicrobe.**

(DOCX)

**S2 Table. Alignment of CIRM database fields to Omnicrobe database fields.**

(DOCX)

**S3 Table. Examples of challenges faced by text-mining methods.**

(DOCX)

**S4 Table. Ten most frequent taxa, habitats, phenotypes and uses in relationships extracted from PubMed in Omnicrobe.**

(DOCX)

**S5 Table. Phenotypes registered in Omnicrobe DB V1.0.**

(DOCX)

**S1 Text. Queries used to extract data from PubMed and GenBank.**

(DOCX)

## Acknowledgments

The authors thank the Migale platform for providing the resources to run Omnicrobe services (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: [10.15454/1.5572390655343293E12](https://doi.org/10.15454/1.5572390655343293E12)).

The current affiliation of Estelle Chaix is the French Agency for Food, Environmental and Occupational Health & Safety (ANSES), 14 rue Pierre et Marie Curie, 94701 Maisons Alfort Cedex, France.

## Author Contributions

**Conceptualization:** Robert Bossy, Claire Nédellec.

**Data curation:** Hélène Falentin.

**Formal analysis:** Robert Bossy.

**Funding acquisition:** Hélène Falentin, Claire Nédellec.

**Methodology:** Robert Bossy, Louise Deléger.

**Project administration:** Sandra Dérozier.

**Resources:** Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Valentin Loux, Claire Nédellec.

**Software:** Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba.

**Supervision:** Claire Nédellec.

**Validation:** Robert Bossy, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Hélène Falentin, Claire Nédellec.

**Visualization:** Sandra Dérozier.

**Writing – original draft:** Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Hélène Falentin, Claire Nédellec.

**Writing – review & editing:** Sandra Dérozier, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Valentin Loux, Hélène Falentin, Claire Nédellec.

## References

1. Ducklow H. Microbial services: challenges for microbial ecologists in a changing world. *Aquat Microb Ecol.* 2008; 53: 13–19. <https://doi.org/10.3354/ame01220>
2. Imhoff J. New Dimensions in Microbial Ecology—Functional Genes in Studies to Unravel the Biodiversity and Role of Functional Microbial Groups in the Environment. *Microorganisms.* 2016; 4: 19. <https://doi.org/10.3390/microorganisms4020019> PMID: 27681913

3. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science*. 2015; 348: 1261359. <https://doi.org/10.1126/science.1261359> PMID: 25999513
4. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012; 486: 215–221. <https://doi.org/10.1038/nature11209> PMID: 22699610
5. Krause S, Le Roux X, Niklaus PA, Van Bodegom PM, Lennon JT, Bertilsson S, et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front Microbiol*. 2014; 5. <https://doi.org/10.3389/fmicb.2014.00251> PMID: 24904563
6. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*. 2017; 551: 457–463. <https://doi.org/10.1038/nature24621> PMID: 29088705
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016; 3: 160018. <https://doi.org/10.1038/sdata.2016.18> PMID: 26978244
8. Ivanova N, Tringe SG, Liolios K, Liu W-T, Morrison N, Hugenholtz P, et al. A call for standardized classification of metagenome projects: Genomics update. *Environ Microbiol*. 2010; 12: 1803–1805. <https://doi.org/10.1111/j.1462-2920.2010.02270.x>
9. Bossy R, Deléger L, Chaix E, Ba M, Nédellec C. Bacteria Biotope at BioNLP Open Shared Tasks 2019. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. pp. 121–131. <https://doi.org/10.18653/v1/D19-5719>
10. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2010; 38: D5–D16. <https://doi.org/10.1093/nar/gkp967> PMID: 19910364
11. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res*. 2012; 40: D136–D143. <https://doi.org/10.1093/nar/gkr1178> PMID: 22139910
12. Nédellec C, Bossy R, Chaix E, Deléger L. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. *arXiv*; 2018. <http://arxiv.org/abs/1805.04107>
13. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, et al. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res*. 2021; 49: D723–D733. <https://doi.org/10.1093/nar/gkaa983> PMID: 33152092
14. Pignatelli M, Moya A, Tamames J. EnvDB, a database for describing the environmental distribution of prokaryotic taxa. *Environ Microbiol Rep*. 2009; 1: 191–197. <https://doi.org/10.1111/j.1758-2229.2009.00030.x> PMID: 23765793
15. Wu L, Sun Q, Desmeth P, Sugawara H, Xu Z, McCluskey K, et al. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Res*. 2017; 45: D611–D618. <https://doi.org/10.1093/nar/gkw903> PMID: 28053166
16. Reimer LC, Vetcinova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, et al. Bac Dive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res*. 2019; 47: D631–D636. <https://doi.org/10.1093/nar/gky879> PMID: 30256983
17. Parente E, De Filippis F, Ercolini D, Ricciardi A, Zotta T. Advancing integration of data on food microbiome studies: FoodMicrobionet 3.1, a major upgrade of the FoodMicrobionet database. *Int J Food Microbiol*. 2019; 305: 108249. <https://doi.org/10.1016/j.ijfoodmicro.2019.108249> PMID: 31207407
18. Parte AC. LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res*. 2014; 42: D613–D616. <https://doi.org/10.1093/nar/gkt1111> PMID: 24243842
19. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics*. 2017; 18: 114. <https://doi.org/10.1186/s12864-017-3501-4> PMID: 28361695
20. Tara Oceans Consortium Coordinators, Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data*. 2015; 2: 150023. <https://doi.org/10.1038/sdata.2015.23> PMID: 26029378
21. Buttigieg P, Morrison N, Smith B, Mungall CJ, Lewis SE, the ENVO Consortium. The environment ontology: contextualising biological and biomedical entities. *J Biomed Semant*. 2013; 4: 43. <https://doi.org/10.1186/2041-1480-4-43> PMID: 24330602
22. Buttigieg PL, Pafilis E, Lewis SE, Schildhauer MP, Walls RL, Mungall CJ. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperability. *J Biomed Semant*. 2016; 7: 57. <https://doi.org/10.1186/s13326-016-0097-6> PMID: 27664130
23. Chibucos MC, Zweifel AE, Herrera JC, Meza W, Eslamfam S, Uetz P, et al. An ontology for microbial phenotypes. *BMC Microbiol*. 2014; 14: 294. <https://doi.org/10.1186/s12866-014-0294-3> PMID: 25433798

24. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. Sarkar IN, editor. PLoS ONE. 2012; 7: e29715. <https://doi.org/10.1371/journal.pone.0029715> PMID: 22238640
25. European Food Safety Authority (EFSA). The food classification and description system FoodEx 2 (revision 2). EFSA Support Publ. 2015; 12. <https://doi.org/10.2903/sp.efsa.2015.EN-804>
26. Tamang JP, Cotter PD, Endo A, Han NS, Kort R, Liu SQ, et al. Fermented foods in a global age: East meets West. *Compr Rev Food Sci Food Saf.* 2020; 19: 184–217. <https://doi.org/10.1111/1541-4337.12520> PMID: 33319517
27. Janda JM, Abbott SL. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *J Clin Microbiol.* 2007; 45: 2761–2764. <https://doi.org/10.1128/JCM.01228-07> PMID: 17626177
28. Chaix E, Deléger L, Bossy R, Nédellec C. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiol.* 2019; 81: 63–75. <https://doi.org/10.1016/j.fm.2018.04.011> PMID: 30910089
29. Cohen KB, Hunter L. Getting Started in Text Mining. *PLoS Comput Biol.* 2008; 4: e20. <https://doi.org/10.1371/journal.pcbi.0040020> PMID: 18225946
30. Aubin S, Hamon T. Improving Term Extraction with Terminological Resources. In: Salakoski T, Ginter F, Pyysalo S, Pahikkala T, editors. *Advances in Natural Language Processing.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. pp. 380–387. [https://doi.org/10.1007/11816508\\_39](https://doi.org/10.1007/11816508_39)
31. Golik W, Warnier P, Nédellec C. Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. Paris; 2021. pp. 37–39.
32. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. *F1000Research.* 2021; 10: 33. <https://doi.org/10.12688/f1000research.29032.2> PMID: 34035898
33. Harlé O, Falentin H, Niay J, Valence F, Courselaud C, Chuat V, et al. Diversity of the metabolic profiles of a broad range of lactic acid bacteria in soy juice fermentation. *Food Microbiol.* 2020; 89: 103410. <https://doi.org/10.1016/j.fm.2019.103410> PMID: 32138982
34. Tangyu M, Muller J, Bolten CJ, Wittmann C. Fermentation of plant-based milk alternatives for improved flavour and nutritional value. *Appl Microbiol Biotechnol.* 2019; 103: 9263–9275. <https://doi.org/10.1007/s00253-019-10175-9> PMID: 31686143
35. Zafeiropoulos H, Paragkamian S, Ninidakis S, Pavlopoulos GA, Jensen LJ, Pafilis E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms.* 2022; 10: 293. <https://doi.org/10.3390/microorganisms10020293> PMID: 35208748
36. van Bruggen AHC, Goss EM, Havelaar A, van Diepeningen AD, Finckh MR, Morris JG. One Health—Cycling of diverse microbial communities as a connecting force for soil, plant, animal, human and ecosystem health. *Sci Total Environ.* 2019; 664: 927–937. <https://doi.org/10.1016/j.scitotenv.2019.02.091> PMID: 30769316
37. Labbate M, Seymour JR, Lauro F, Brown MV. Editorial: Anthropogenic Impacts on the Microbial Ecology and Function of Aquatic Environments. *Front Microbiol.* 2016; 7. <https://doi.org/10.3389/fmicb.2016.01044> PMID: 27458441
38. Morris CE, Géniaux G, Nédellec C, Sauvion N, Soubeyrand S. One Health concepts and challenges for surveillance, forecasting, and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathol.* 2022; 71: 86–97. <https://doi.org/10.1111/ppa.13446>