Bacteria biotope annotation guidelines

Robert Bossy, Claire Nédellec, Julien Jourde, Mouhamadou Ba, Estelle Chaix, Louise Deleger

Bacteria Biotope Annotation Guidelines

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

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

Contents

0 Note

This document is an updated version of the annotation guidelines of BioNLP-ST 2016. The main update concerns the addition of the phenotype entity and exhibits relation.

1 Introduction

This document specifies the guidelines for the annotation of the BioNLP-OST 2019 Bacteria Biotope corpus. The task consists of the extraction of places where microorganisms live and of microbial phenotypes in a set of scientific texts (Pubmed abstracts, full-text extracts, web pages). In concrete terms, this is specified by two types of relations : (1) relations between microbial taxon names on one hand, and habitats mentions and geographical names on the other hand; and (2) relations between microbial taxon names on one hand, and phenotype mentions on the other hand.
Microorganism taxon names are organized by relevant subtrees from the NCBI Taxonomy (version downloaded on Feb. 7, 2019)
Habitat mentions are organized by the *microbial habitat* subtree of the OntoBiotope ontology (OntoBiotope_BioNLP-OST-2019 version).
Phenotype mentions are organized by the *microbial phenotype* subtree of the OntoBiotope ontology.

1.1 Copyright and License

1.2 Conventions

Annotation schema vocabulary is denoted in `fixed-width font`.
Excerpts from text and surface forms are denoted “between double quotes”.
Habitat concepts and microbial taxa are denoted in *emphasis* (*italic font*).
Examples are given in two parts: an annotated piece of text, then a comment about this annotation. In the examples, `Microorganism` annotations are highlighted in orange, `Habitat` annotations in blue, `Geographical` annotations in red, and `Phenotype` annotations in pink.

2 Microbial taxon names

2.1 Entity domain

2.1.0 Microorganism definition

In Bacteria Biotope, a microorganism is annotated if and only if the taxon has an ancestor specified in the following table:

Microorganism taxon	NCBI ID
Alveolata	33630
Amoebozoa	554915
Nematoda	6231
Choanoflagellida	28009
Cryptophyta	3027
Diplomonadida	5738
Euglenozoa	33682
Fungi	4751
Haptophyceae	2830
Ichthyosporea	127916
Oxymonadida	66288
Parabasalia	5719
Glaucocystophyceae	38254
Chlorella	3071
Prototheca	3110
Chlamydomonadales	3042
Volvox	3066
Desmidiales	131210
Retortamonadidae	193075
Rhizaria	543769
Stramenopiles	33634
Crenarchaeota	28889
Euryarchaeota	28890
Korarchaeota	51967
Nanoarchaeota	192989
Bacteria	2
Viruses	10239

2.1.1 Gram staining

Gram-positive and *gram-negative* bacteria are not microorganism taxa since they have both been proved to be polyphyletic. However the following mentions are microorganism taxa and must be annotated:

- “low G+C gram-positive bacteria”, synonym of *Firmicutes*
- “high G+C gram-positive bacteria”, synonym of *Actinobacteria*

2.1.2 Abbreviations

Annotated abbreviations include:

- genus name abbreviated by its first letter in capital;
- loose strain names;
- widely accepted abbreviations.

Non standard abbreviated taxon names are annotated, if and only if there is an occurrence of the complete non-abbreviated name of the same entity in the same document.

> **Example.** _____
> *…Here, we determined that during Mycobacterium tuberculosis (Mtb) infection…*
>
> → “Mtb” is clearly introduced as an abbreviation as an apposition. Furthermore it is a globally understood abbreviation. Every occurrence in the document is annotated.
>
> _____

2.1.3 Lactic acid bacteria

The term “lactic acid bacteria” (or “LAB”) does not correspond to a single microbial taxon in the taxonomy, therefore **it must not be annotated** as Microorganism. It is considered as a phenotype (cf section 5.1.8).

> **Example.** _____
> *The production of exopolysaccharides by LAB has been correlated to specific gene clusters tagged as eps or cps, located, as in Streptococcus thermophilus*
>
> → “LAB” (meaning lactic acid bacteria) is not annotated as `Microorganism`. It is annotated as `Phenotype`.
>
> _____

2.1.4 Too general

Microorganism mentions that are too general should not be annotated. This includes general words like:

- “microorganism”
- “microbe”
- “microbial”
- “bacterium”
- “bacteria”
- “bacterial”
- “fungi”
- “fungus”
- “virus”

- “yeast” (when it does not specifically refer to *Saccharomyces cerevisiae*)
- “mold”
- “phage”
- “protozoa”

If the authors emphasize on the phylum using capitalization (e.g., “Bacteria”, “Fungus”), then the mention is annotated.
If the mention “yeast” clearly refers to the species *Saccharomyces cerevisiae* then it should be annotated.

> **Example.** _____
> *Streptococcus salivarius is the principal commensal bacterium of the oral cavity in healthy humans.*
>
> → “bacterium” is not annotated.
>
> _____

2.2 Boundaries

The boundaries of the microorganism annotations must be as wide as possible to delimitate the most precise taxon. Thus the boundaries must include strain names, isolate identifiers, etc.

> **Example.** _____
> *The type strain of the species is Bradyrhizobium japonicum USDA 6 which was isolated from […]*
>
> → The annotation includes the genre, species and strain.
>
> _____

However the span of the *Microorganism* annotation should not include the trailing or leading words “strain”, “genus”, “species”, etc.

> **Example.** _____
> *The strain 96-OK-85-24 significantly differed from the existing mosquitocidal B. thuringiensis strains.*
>
> → The leading “strain” is excluded from the annotation.
>
> _____
>
> **Example.** _____
> *inhibition of PMN ROS production with diphenyleneiodonium chloride resulted in a reduction of PMN cell death similar to that induced by the virulence plasmid-containing strain Y. pestis KIM5.*
>
> → The leading “strain” is excluded from the annotation.
>
> _____

The annotation excludes modifiers that qualify a taxon or a strain but that are not part of the

taxon name.

Example. _____
…the fur gene was cloned from a pathogenic Pseudomonas fluorescens strain isolated from diseased Japanese flounder.

→ “pathogenic” is excluded.

Example. _____
[…] reservoirs of viable methicillin-resistant Staphylococcus aureus.

→ “methicillin-resistant” is excluded from the `Microorganism` annotation.

2.2.1 Phenotype acronyms designating microorganisms

Phenotypes and qualifiers are not included in `Microorganism` annotations (see above). Acronyms that abbreviate both the phenotype and the species name must be annotated as `Microorganism` (for widely accepted abbreviations) and as Phenotype. In particular, the following abbreviations should be annotated:

- “MRSA”: methicillin-resistant Staphylococcus aureus
- “EPEC”: enteropathogenic Escherichia coli
- “EHEC”: enterohemorrhagic Escherichia coli
- “NTHi”: nontypeable Haemophilus influenzae
- “MDRTB”: Multi-drug-resistant tuberculosis
- “VRE”: Vancomycin-Resistant Enterococci
- “MDRP”: multidrug resistant Pseudomonas aeruginosa

This list might grow with acronyms widely used in papers.

Example. _____
We examined jail environmental surfaces to explore whether they might serve as reservoirs of viable methicillin-resistant Staphylococcus aureus (MRSA).

→ “methicillin-resistant” is excluded from the first Microorganism annotation. “MRSA” is annotated as `Microorganism` since it is a widely used abbreviation, and as `Phenotype`.

2.2.2 Strain specification

When the species name is followed by a strain name, then a single annotation must both contain the species and the strain names, including words like “strain”, “isolate”, or “serovar” in between.

Example. _____
Heat-shock response and its contribution to thermotolerance of the nitrogen-fixing cyanobacterium Anabaena sp. strain L-31.

→ A single annotation includes the species and the strain.

Example. _____
gram-negative plant pathogen Xanthomonas campestris pv. vesicatoria.

→ "pv." means "pathovar".

The following are considered as strain specifications:

- serovars
- serotypes
- mutants

Example. _____
[…] nonvirulent Ara+ Burkholderia pseudomallei isolates […]

→ The mutation specification "Ara+" is included in the Microorganism annotation.

2.2.3 Nomenclatural suffixes: sp., spp., gen. nov., sp.nov.

After a genus name, "sp." and "spp." mean unspecified single or multiple species of the genus. These abbreviations must be included in the taxon name.
After a genus name, "gen. nov." means the document introduces a new genus name. This abbreviation is not included in the taxon name.
After a species name, "sp. nov." means the document introduces a new species name in the genus. This abbreviation is not included in the taxon name.
After a species binomen, "gen. nov., sp. nov." means the document introduces a new genus name and a new species name. These abbreviations are not included in the taxon name.

Example. _____
Of the 104 isolations of Salmonella sp. from egg pulp, 97 were obtained from strontium chloride M broth.

→ "sp." is included in the taxon name.

Example. _____
A novel species of a new genus in the family Chitinophagaceae, for which the name Taibaiella smilacinae gen. nov., sp. nov. is proposed.

→ "gen. nov., sp. nov." has a meaning that is circumstantial to the document and is not included in the taxon name.

2.3 Taxon ID

The attribute `NCBI_Taxonomy` must be filled in the AlvisAE annotation interface for each occurrence of *Microorganism* entities. It informs the taxon identifier from the NCBI Taxonomy.

Example.

NCBI_taxonomy=1280 NCBI_taxonomy=1280

[…] reservoirs of viable methicillin-resistant Staphylococcus aureus (MRSA) […] Finding

NCBI_taxonomy=1280

MRSA-contaminated surfaces on a variety of environmental surfaces […]

→ "Staphylococcus aureus" and "MRSA" are assigned the taxid 1280.

2.3.1 Unknown taxon identifier

If the taxon identifier is unknown because it is missing in the NCBI Taxonomy:

- If the taxon is of a rank *below* the species (*e.g.* strain), then the entity is assigned the taxon identifier of the species.
- If the taxon is a species, or a higher rank, then the situation is exceptional and must be notified in the annotation discussion forum.

Note that some mutants have their own entry in the NCBI taxonomy.

2.3.2 Partial coreference

A common practice is to mention a precise taxon at the beginning of the document, then refer to the same taxon using a higher and shorter taxon name. In this case the coreference, even partial, is assigned to the identifier of the antecedent taxon.

Example.

NCBI_taxonomy=456327

XopC and XopJ, two novel type III effector proteins from Xanthomonas campestris pv.

NCBI_taxonomy=456327

vesicatoria … Both genes encode Xanthomonas outer proteins (Xops) that were [...]

→ The "Xanthomonas" occurrence in the second sentence clearly denotes the taxon mentioned in full in the first sentence, and not the genus Xanthomonas. Thus it inherits the same taxon identifier.

Example. _____

NCBI_taxonomy=2742 NCBI_taxonomy=1236

Marinobacter belong to the class of Gammaproteobacteria

→ In this case “Gammaproteobacteria” is not a coreference for “Marinobacter”, it is a statement of the relationship between both taxa. “Gammaproteobacteria” is thus assigned to the taxon Gammaproteobacteria.

_____

3 Habitat mentions

3.1 Entity domain

Mentions of microorganism habitats are expressions and phrases that denote a physical place where microorganisms could be observed. This includes:

- biomes (natural habitats, soil, sea, etc.);
- hosts (living beings of any phylum) and their parts (organs, secretions, excretions);
- human artefacts (food, buildings, equipment, farms);
- environments qualified by their physical or chemical properties.

This excludes:

- General places (3.1.1)
- Diseases, symptoms, diagnostic methods (3.1.2)
- Molecule and drugs (3.1.5)

3.1.1 Too general

When a localization is too general or too imprecise, it must not be annotated. The following list is a vocabulary of terms which are too general:

- “antibiotic”
- “antimicrobial”
- “biopsy specimens”
- “biotope”
- “carrier”
- “cohort”
- “culture” (exception: meaning crop)
- “drug”
- “ecosystem”
- “environment”
- “extract”
- “extracellular”
- “field” (exception: meaning crop)
- “growth medium”

- “host”
- “in vitro”
- “in vivo”
- “media”
- “medium”
- “microbe” / “microbial” / “microorganism”
- “nature”
- “niche”
- “population”
- “product” (exception: meaning food)
- “site”
- “solution”
- “subject”
- “substrate”
- “substrat”
- “suspension”
- “underdeveloped countries”
- “vector”
- “eukaryote”
- “eukarya”
- “facility”
- “world”

These words, if they are not attached to more precise modifiers must not be annotated. Note that “body” is not considered too general when it refers to a host.

3.1.2 Diseases, symptoms, diagnostic methods

Disease, symptom and diagnostic method names do not denote microorganism habitats and must never be annotated.

> **Example.** _____
> *Groups were stratified on the basis of age, Injury Severity Score (ISS), Glasgow Coma Scale (GCS) Score, base excess (BE), ICP days, transfusions in 24 h, ICU days, ventilator days, head Abbreviated Injury Score (AIS), and chest AIS.*
>
> → “Injury” and “chest” are not annotated because they are all part of a name of a diagnostic method.
>
> _____

3.1.3 Part of living organisms

Parts of living organisms are habitats, their names must be annotated. Parts of living organisms include organs, tissues, fluids, also non-living parts, and unhealthy parts:

- “abscesses”

- “excretions”
- “fluids”
- “phyllome”
- “rhizome”
- “secretions”
- “tumors”
- “wounds”

However part of living organisms are annotated from the macroscopic scale down to the cell included. Subcellular scale parts of living organisms are not annotated. Thus organelles, cytoplasm, membranes, cell walls are excluded, as well as surface or border of cells.
Occurrences of the word “cell” must be annotated only if they denote a potential host cell. They must not be annotated if they denote microorganism cells.

Example. _____

[...] one ureter cell line (SV-HUC-1) was incubated in artificial urine with five Proteus mirabilis strains.

→ “ureter cell” denotes a eukaryote cell and is thus annotated.

_____

Example. _____
When we cloned FlgF, a flagellar rod protein, from Salmonella typhimurium and overproduced it in Escherichia coli, FlgF was highly susceptible to cleavage by endogenous proteases after cell disruption even in the presence of various protease inhibitors.

→ “cell” denote microorganism cells, and thus is not annotated.

_____

Example. _____
This greater permeability of the H. influenzae cell to penicillins appeared to reduce the protective effect of its beta-lactamase.

→ “H. influenzae cells” denote microorganism cells, and thus is not annotated.

_____

Example. _____
Escherichia coli [...] was found to adhere only to the brush border of epithelial cells

→ “brush border of epithelial cells” is subcellular and is not annotated, but “epithelial cells” is annotated.

_____

3.1.4 Experimental materials and methods

Experimental method names must never be annotated.
Experimental material including devices, equipment, and media must be annotated except when the mention is part of a method name.

Example. _____
We ran pulsed-field gel electrophoresis on six resistant isolates and observed three patterns.

→ “pulsed-field gel electrophoresis” is not annotated because it is a method. “pulsed-field gel” is not annotated even if it is an equipment because it is part of the method name.

Example. _____
In this report, we introduce a liquid chromatography single-mass spectrometry method for metabolome quantification, using the LTQ Orbitrap high-resolution mass spectrometer.

→ “LTQ Orbitrap high-resolution mass spectrometer” is an equipment and is not in the excluded vocabulary, it is thus annotated.

3.1.5 Experimental media, molecules, drugs, and substances

Molecule names, including drugs, are not habitats and not annotated as `Habitat`. However some substances, experimental media, and habitats are designated by the most relevant molecules. In this case, the mention is annotated as `Habitat`.

Example. _____
These β-CAs could serve as novel antimicrobial drug targets for this pathogen.

→ “antimicrobial drugs” is not an habitat because it denotes a set of molecules.

Example. _____
MRSA were isolated by oxacillin screening agar.

→ “oxicillin” is not an habitat, it is a molecule. However “agar” is an habitat and thus annotated with its modifiers, including “oxicillin”.

Example. _____
When biphenyl-grown cells were transferred back to a fructose medium, they required 25 generations to […]

→ “biphenyl” and “fructose” are molecules, thus not annotated as habitats. “fructose medium” is annotated as an habitat characterized by its molecule contents.

Example. _____
This same residue would serve to deprotonate the incoming water and reprotonate the enolate in the second half of the catalytic cycle.

→ Here, "water" is a H_2O molecule.

_____

3.1.6 Experimental conditions and environment properties

Experimental conditions and environment properties on their own must not be annotated as `Habitat`.

Example. _____
Potent bactericidal properties were maintained at high salt concentrations, under acidic or basic conditions, and at extreme temperatures.

→ "high salt concentrations", "acidic conditions", "basic conditions", "extreme temperatures" are not annotated.

_____

3.1.7 Microscopic habitats

Microbial taxon names as well as common terms denoting microorganisms (such as "yeasts", "fungi", "molds", "viruses" and "protozoans", see the list: 2.1.4 Too general) are not annotated as `Habitat`.
However, microbial communities (i.e., microflora) as a whole are annotated as `Habitat`. This includes terms such as "microflora", "bacterial population", "microbial community", "microbiota", "starter culture", and so on.

3.2 Boundaries

Habitat mentions are noun phrases or isolated adjectives.

3.2.1 Noun phrases

The annotation of a noun phrase habitat must contain the head of the noun phrase as well as all significant modifiers. A significant modifier is a modifier relevant to the microorganism living conditions.
Conversely, the boundaries shall exclude modifiers that are irrelevant to the microorganism. Excluded modifiers are:

- general adjectives ("diverse", "common");
- relative adjectives or adverbs ("different", "other");
- cardinals and ordinals.

Example. _____

In a group of 17 patients with duodenal ulcers the authors investigated the effect of omeprazole.

→ "17" is a cardinal and thus excluded. "with duodenal ulcers" specify the host and is included.

_____

3.2.2 Host characterization

Host characterizations are included in the annotation forming a single *Habitat*.

Example. _____
[…] green algae Acrosiphonia […]

→ A single annotation.

_____

Example. _____
[…] blood-sucking tsetse fly […]

→ A single annotation.

_____

Note that if the characterization of the host is denoted by an apposition, then two separate *Habitat* are annotated (see next).

3.2.3 Appositions

Appositions are annotated separately.

Example. _____
[…] mosquito (Aedes albopictus) […]

→ "mosquito" and "Aedes albopictus" are annotated separately.

_____

However a *Habitat* that includes appositions must be annotated in a single fragment (not discontinuous).

Example. _____
… assessed by determining the degree of attachment to hydrophilic tissue culture plates and human corneal epithelial (HCE) cells.

→ The parenthesis is included in the Habitat annotation.

_____

3.2.4 Geographical position modifier

Geographical position modifiers introduced by the prepositions “in” or “from” are not included in the *Habitat* annotation.

Example. _____
The prevalence of H. pylori infection in dyspeptic patients in Yemen is very high.

→ “Yemen” is not included in the annotation of “dyspeptic patients”.

_____

Example. _____
[…] the principal mycoplasmosis of sheep and goats in Europe.

→ “Europe” is excluded from “sheep” and “goat” annotations.

_____

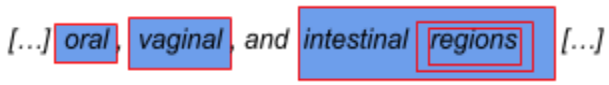
3.2.5 Enumerations

When several habitats are enumerated, there are two cases:

The enumeration denotes a conjunction : the habitat is specified by the intersection of the enumerated items. In this case a single `Habitat` annotation covers the whole enumeration.

The enumeration denotes a disjunction : several related habitats are enumerated. In this case one `Habitat` mention for each enumeration item must be annotated. If the factored part is leftward, then all annotations but the first are discontinuous. If the factored part is rightward, then all annotations but the last are discontinuous.

Example. _____



→ Three habitats are annotated, two of them (“oral region” and “vaginal region”) are discontinuous, the last one (“intestinal regions”) is continuous.

_____

3.2.6 Adjectives

Adjectives relating to an habitat, or a tropism must be annotated.
List of adjectives to be annotated:

- “aquatic”
- “enteroinvasive”
- “foodborne”
- “marine”
- “nosocomial”

Moreover, all organs and parts of living beings mentioned as adjectives must be annotated.
The “clinical” adjective must be annotated if it qualifies a microbial strain, most often the heads of clinical strains are “isolate”, “strain”, or “sample”. However “clinical” must not be annotated if it qualifies studies, or surveys. “clinical samples” must not be annotated if “sample” designate a human population sample in clinical studies. “clinical” must be associated with the concept patient, or one of its sub-concepts.
Trophisms **are not annotated** as habitats:

- “phototroph”
- “methanotroph”

3.2.7 Overlapping habitats

Habitat mentions whose boundaries are contained in another one are annotated, except if and only if:

- the containing and the contained mentions share the same head, and
- the contained mention denotes an habitat that is a super-concept of the containing mention.

Example. _____

[…] 4 isolates from intestinal tracts of healthy fish and 98 isolates from sediments […]

→ Both “intestinal tracts of healthy fish” and “healthy fish” are annotated because (1) they have distinct heads, and (2) *fish* is not a super-concept of *intestinal tract*.

_____

Example. _____

The most likely location for the proliferation of resistant lineages is in farmed chickens.

→ “chickens” is not annotated since it has the same head as and is a super-concept of “farmed chickens”.

_____

Example. _____

[…] the intestinal environment of healthy individuals with soft stools were evaluated.

→ Each embedded fragment is a different habitat and annotated separately.

_____

3.3 OntoBiotope concepts

Each `Habitat` annotation must be associated to one or several OntoBiotope concepts through the attribute `OntoBiotope` in the AlvisAE interface. The annotation of an `Habitat` by several concepts means that the conjunction of this concepts is true. It is not possible to represent a disjunction. In case of doubt, the most probable concept should be associated.

Example. _____

Fermented milk, probiotic food

Effects of a probiotic fermented milk beverage containing Lactobacillus casei strain Shirota on defecation frequency.

→ "probiotic fermented milk beverage" is both a *fermented milk* and a *probiotic food* at the same time.

_____

3.3.1 Partial coreference

A common practice is to mention a precise habitat at the beginning of the document, then refer to the same habitat using a shorter more general habitat. In this case the most precise concept is assigned to the coreference, even partial.

Example. _____

duodenal ulcer

[...] the authors found that 5 of 17 chronic duodenal ulcers were healed in the remainder

duodenal ulcer patient

substantial regression was found. [...] the ulcer healed in 11 instances, in 7 patients residual bulbitis persisted.

→ The "ulcer" entity is a coreference of "chronic duodenal ulcers" hence it is assigned the same OntoBiotope concept *duodenal ulcer*.

_____

3.3.2 Creation of new synonyms and concepts

New synonyms and concepts of habitats can be created during the annotation by extending OntoBiotope.
New synonyms must be widely recognized forms of the concept. In OntoBiotope the synonymy is strict.
New concepts must fill a gap in the ontology. The ontology will be curated regularly.
Use your best judgment.

4 Geographical names

Only geographical names are annotated as `Geographical`. Annotators are required to check if the mentioned names belong to gazetteers and administrative name lists.

4.1 Disambiguation of geographical names

Some geographical names may be ambiguous and document authors take care to add clues for its disambiguation. The annotation must span over these clues.

Example. _____
[…] island of Malta […]

→ The annotation includes “island” in order to distinguish between the main island of Malta and the state of Malta.

_____

4.2 Succession of geographical names

When geographical names follow one each other, in an enumeration, usually each name designate a place included in the following one. Each name must be annotated as a separate entity.

Example. _____
Surveillance for upper respiratory tract disease and Mycoplasma in free-ranging gopher tortoises (Gopherus polyphemus) in Georgia, USA.

→ Each geographical location, “Georgia” and “USA”, is annotated as a distinct entity.

_____

4.3 Zone specification

The specification of a particular zone of a geographical location must be included in the annotation.

Example. _____
To estimate the prevalence of thermotolerant Campylobacter spp. in commercially reared partridges (Perdix perdix) in southern Italy.

→ “southern” is included in the annotation because it specifies a specific zone of the location (“Italy”).

_____

4.4 Names containing geographical names

Some disease names or common taxon names contain geographical names or names of species. The included geographical names shall not be annotated.

Example. _____

[…] African river blindness […]

→ “African” is part of a disease name, thus not annotated.

Example. _____
Groups were stratified on the basis of age, Injury Severity Score (ISS), Glasgow Coma Scale (GCS) Score, base excess (BE), ICP days, transfusions in 24 h, ICU days, ventilator days, head Abbreviated Injury Score (AIS), and chest AIS.

→ “Glasgow” is annotated because it is part of a name of a diagnostic method.

4.5 Geographical habitats

Some geographical names also denote `Habitats` like lake and river names. These mentions must be annotated twice with two entities: one `Geographical` entity, and one `Habitat` entity.

Example. _____
In meromictic Lake Cadagno, Switzerland, [...]

→ “Lake Cadagno” is annotated as a `Geographical` and is also part of the “meromictic Lake Cadagno” `Habitat` entity.

4.6 Nationalities and adjectives

Adjectives that denote geographical places are annotated. Nationalities are not annotated because they are not microorganism habitat.

Example. _____
[...] isolated from the culturable microbiota associated with the Mediterranean seagrass Posidonia oceanica.

→ “Mediterranean” is annotated.

Example. _____
[...] the development of clinical AIDS among Italian patients [...]

→ “Italian” is not annotated as a `Geographical`. “Italian patients” is annotated as a `Habitat`.

Example. _____
Eight patients shared a strain identical to a previously reported Australian transmissible

strain (pulsotype 1).

→ “Australian” is not annotated.

---

5 Phenotypes

5.1 Entity definition

`Phenotype` mentions are spans of text that describe microbial characteristics, such as morphology, development and physiological properties or behavior, source of energy, adaptation to physicochemical properties (acidity, oxygen, temperature).

Example. _____
[…] reservoirs of viable methicillin-resistant Staphylococcus aureus.

→ “methicillin-resistant” is annotated as `Phenotype`.

Other examples of phenotype mentions include: methanotroph, acidophilus, aerobe, halophile, motile, antibiotic resistant, yellow, not pigmented, pathogenic, commensal.

5.1.1 Phenotype acronyms

Acronyms that abbreviate both the phenotype and the species name must be annotated as `Phenotype` and (for widely used acronyms corresponding to microbial taxa) as `Microorganism`.

Example. _____
We examined jail environmental surfaces to explore whether they might serve as reservoirs of viable methicillin-resistant Staphylococcus aureus (MRSA).

→ “MRSA” is annotated both as `Microorganism` and as `Phenotype`.

Example. _____
The production of exopolysaccharides by LAB has been correlated to specific gene clusters tagged as eps or cps, located, as in Streptococcus thermophilus

→ “LAB” which means *lactic acid bacteria* is annotated as `Phenotype` but not as `Microorganism`

5.1.2 Taxon names denoting the form of the bacteria

Taxon names denoting the form of the bacteria (such as *Streptobacillus*, *Streptococcus*, *Staphylococcus*) should **not** be annotated as `Phenotype`. They are annotated as `Microorganism`.
However, words from classical roots (Latin, Greek) that denote only the form the bacteria and are not part of a taxon name should be annotated (e.g., *coccus*).

5.1.3 Microscopic eukaryotes defined by their phenotypes

Microscopic eukaryotes defined by their phenotypes, such as protozoans and molds, should not be annotated as `Phenotype`.

5.1.4 Metabolic activity

Metabolic activities of microorganisms, such as production and degradation of molecules, are annotated as `Phenotype` entities, if and only if they are expressed as noun phrases or adjectives (cf. section 5.2). They are usually normalized using the general OntoBiotope concept *phenotype wrt metabolic activity* concept (cf section 5.3.2).

5.1.5 Virulence and diseases

Virulence is considered a synonym of pathogen and is annotated as `Phenotype`, including in the context of virulence genes.

Diseases and symptoms are not considered as phenotypes and are not annotated. However noun phrases that indicate that a microorganism causes the disease or the symptom are annotated as `Phenotype` and linked to the concept *pathogen*.

> **Example.**
> Chlamydia trachomatis is globally the predominant infectious cause of blindness
>
> → “cause of blindness” is annotated as `Phenotype`.

5.1.5 Adherence

Mentions of “adhesion” and “attachment” of a microorganism should be annotated. They correspond to the *adherent* concept in OntoBiope.

5.1.6 Too general

Phenotypes that are too general should not be annotated. For instance:

- “development”
- “survival”
- “growth”
- “viable”
- “group-beneficial traits”

- “selfish”

5.1.7 Mutant and wild-type

Mentions describing a microorganism as being a mutant or a wild-type strain are annotated as `Phenotype` entities, and normalized using either the *mutant* or the *wild-type* OntoBiotope concepts.

> **Example.** _____
> In this study, we utilized wild-type and mutant strains of Yersinia
>
> → “wild-type” and “mutant” are annotated as `Phenotype`.
>
> _____

5.1.8 Lactic acid bacteria

The term “lactic acid bacteria” (or “LAB”) refers to bacteria that are able to produce lactic acid. Since the production of lactic acid is a phenotype, the whole phrase “lactic acid bacteria” is annotated as `Phenotype` (with the OntoBiope concept *phenotype wrt metabolic activity*).

5.2 Phenotype boundaries

Phenotype mentions are either noun phrases or adjectives. **They should only include significant modifiers**, i.e. modifiers that are relevant and necessary to the phenotypes. In particular, general modifiers like “various”, “several”, “main”, “major”, and so on, should not be included in the annotation. Thus phenotype mentions are generally short phrases, rather than complex phrases.

5.2.1 Intensity qualifiers

Adjectival and adverbial modifiers that qualify the intensity of the phenotypes (“high”, “low”, “poor”, “moderately”, and so on) should be included in the annotation. However, comparative and superlative adjectives (higher, highest, lower...) are excluded.

> **Example.** _____
> *E. faecalis, E. faecium, E. durans and L. lactis ssp. lactis were placed in quadrant 3, while L. paracasei ssp. paracasei and L. plantarum were in quadrant 2. In fact, both groups are associated with high reduction activity.*
>
> → The qualifier “high” is included in the phenotype span.
>
> _____
>
> **Example.** _____
> *The L. lactis ssp. lactis strains had the highest reduction activity*
>
> → The superlative adjective “highest” is excluded from the phenotype span. Only “reduction activity” is annotated.

5.3 OntoBiotope concepts

Each `Phenotype` annotation must be associated to one or several OntoBiotope concepts through the attribute `OntoBiotope`.

Example.

[...] the origin of pigments seems essentially related to the presence of yellow bacteria such as Arthrobacter species

→ "yellow" is assigned the *yellow pigmented* concept.

Example.

The nonpathogen enterococci and L. lactis ssp. lactis reach Eh values of −120 and −220 mV in a very short time.

→ "nonpathogen" is assigned the *commensal* concept.

5.3.1 Adjectival vs. nominal phenotypes

When choosing an OntoBiotope concept for a given text span, we do not make a difference between phenotypes expressed as adjectival phrases and phenotypes expressed as noun phrases. For instance, the phrase "multi-drug resistance" is normalized with the same concept as the phrase "multi-drug resistant", i.e. the *drug resistant* concept. This is the case even if the phenotype is only tested and is not actually present in the microorganism. In this case, there will be no relation between the phenotype and the microorganism, but the normalization of the phenotype will not change.

Example.

Spirometry, anthropometrics, hospitalisations and antibiotic sensitivity data were recorded.

→ "antibiotic sensitivity" is assigned the *antibiotic sensitive* concept.

5.3.2 Metabolic activity concept

Phenotype mentions denoting metabolic activity, such as production and degradation of molecules, are assigned the *phenotype wrt metabolic activity* concept, if there is not a more precise ontology concept corresponding to the mention.

> **Example.** _____
> *Remus et al. (2012) identified four cps genes clusters in the chromosome of L. plantarum WCFS1, which are associated with surface polysaccharide production.*
>
> → surface polysaccharide production should be annotated as a phenotype and assigned to the *phenotype wrt metabolic activity* concept.
> _____

6 Lives_In relation

This section specifies which relations must be annotated. The `Lives_In` relation is oriented. It has two arguments:

- Argument 1: `Microorganism` must be a microorganism taxon name, it is the source of the relation;
- Argument 2: `Location` must be either a habitat mention or a geographical name, it is the target of the relation.
  NB: in some rare cases, a microorganism may be found inside another microorganism. In these cases, the target of the relation is a `Microorganism` entity.

The argument entities must be as close as possible graphically.
The `Lives_In` relation must be explicit within the scope of the document. The microorganism must be alive in the mentioned localization.
When the localization is implicit but is certain from the point of view of a reader, then the relation must be annotated:

> **Example.** _____
> *Vibrios[…] are ubiquitous to oceans, coastal waters, and estuaries. […] The bacterial pathogen is a growing concern in North America, particularly in places where seafood is popular.*
>
> → The localization in “North America” is implicit but certain, it must be annotated.
> _____

6.1 Topological constraints

The `Lives_In` relation does not carry strong topological semantics. The relation may be annotated whether the microorganism lives on the surface, inside, or on top of the habitat.
However if the habitat is a cell, the relation is annotated only if the microorganism is inside the cell. The relation is not annotated if the microorganism adheres to the cell membrane or binds to a surface protein.

Example. _____

M. leprae directly bind to ErbB2 on myelinated Schwann cells.

→ “myelinated Schwann cells” is a cell and the microorganisms do not live inside, so the relation is **not** annotated.

_____

Example. _____

Francisella tularensis, a causative agent of human tularemia, displaying the ability to proliferate inside the human cells.

→ Here the microorganisms are explicitly said to be living inside the human cells, thus the relation is annotated.

_____

6.2 Partial localization

The relation is not annotated if only some members of the taxon are mentioned as living in the localization, but not all. However observations and experiments can be generalized for the whole taxon.

Example. _____

Yersinia. This genus consists of 11 species, 3 of which are human pathogens.

→ The relation between “Yersinia” and “human” is not annotated since only sub-taxa are said to be human pathogens.

_____

Example. _____

Marinobacter belong to the class of Gammaproteobacteria and these motile, halophilic or halotolerant bacteria are widely distributed throughout the world’s oceans.

→ The relation between “Marinobacter” and “world’s oceans” is annotated. However there is no relation annotated between “Gammaproteobacteria” and “world’s oceans” because this relation would not be universal.

_____

6.3 Effect of microorganisms on the environment

The localization of a microorganism may be mentioned through its effect on the immediate environment. These must be annotated with special care, in particular:

6.3.1 Diseases and symptoms

Pathogen microorganisms are often described by the disease they cause. A microorganism that causes a disease on a host is always considered to be located in this host.

Also a microorganism that causes an epidemic on a geographical place is always considered to be located in this geographical place.
Pathogen microorganisms are often described by the symptoms of the disease they cause. The annotator must be careful to distinguish whether this effect means the microorganisms are located in the host part or not. For instance inflammatory reactions and necroses do not imply the microorganisms are located there.
On the other hand, some terms mention a symptom as well as a localization:

- “abscess”
- “colonization”
- “commensality”
- “infection”
- “invasion”

Example. _____
Long-term data show that H. pylori infection can lead to gastric atrophy and may play an important role in the development of gastric cancer.

→ The microorganism causes a symptom (“gastric atrophy”) but the microorganism is not necessarily located in the stomach (“gastric”). It may be an indirect cause.

_____

Example. _____

Non-O1 Vibrio cholerae bacteremia in patients with cirrhosis: 5-yr experience from a single medical center. …The overall case-fatality rate was 23.8%, but 75% of the deaths were observed in patients with skin manifestation.

→ The relation between “Non-O1 Vibrio cholerae” and “patients with cirrhosis” is annotated. However the relation between the microorganism and the skin is not established, even though it causes a symptom on the skin.

_____

6.3.2 Symbioses

The annotator must be careful that a symbiosis between a microorganism and another living being (the host) does not necessarily mean that the microorganism lives in the host. If the microorganism lives in the host, it is generally mentioned explicitly either independently from the mention of symbiosis, or the symbiosis is specified with explicit localization terms (“endosymbiosis”).

Example. _____
This nitrogen-fixing bacterium develops a symbiotic relationship with the soybean plant Glycine max.

→ No relation between a bacterium and “Glycine max” shall be annotated because this symbiosis relationship may be on nutrient exchange basis only.

_____

Example. _____
S. glossinidiusis an endosymbiont of the tsetse fly.

→ This relation is annotated since the "endo-" prefix makes the microorganism localization explicit.

_____

6.4 Experimental settings

Descriptions of microorganism cultures grown on enriched substrates (peptone, galactose, glucose, lactate, acetate, etc.) must be annotated as `Lives_In` relations.
However the laboratory, university, research center, country where the experiment was conducted are excluded.

Example. _____
[…] an isolate of this species was studied by researchers at University of California.

→ No relation is annotated with "California", it is just the place where an experiment was conducted.

_____

6.5 Vaccines

If a vaccine is a living vaccine, then the `Lives_In` relation must be annotated between the antigenic microorganism and the vaccine habitat. Note however that most vaccines are made of dead microorganisms.

6.6 Hypothesis sentence

Utterances for a working hypothesis must not be annotated as a `Lives_In` relation. Positive evidence of the hypothesis further in the document must be annotated as a `Lives_In` relation. This relation might have as an argument one of the entities in the hypothesis sentence when there is no closer alternative.

Example. _____
We examined jail environmental surfaces to explore whether they might serve as reservoirs of viable methicillin-resistant Staphylococcus aureus (MRSA).

→ This sentence is an hypothesis, the relation between "Staphylococcus aureus" and "jail environmental surfaces" is not proved (yet).

_____

6.7 Relation transitivity

This section covers the case where a microorganism taxon lives in a location, and that this

location is included in, inside of, or part of another location. Depending on the nature of both locations, the `Lives_In` is transitive or not.

First location	Second location	Transitive
Experimental setting	`Geographical`	No
Any other `Habitat` or `Geographical`	`Geographical`	Yes
Part of living organism	Living organism	Yes
Living organism	Living organism	No
Living organism	Environment of the living organism	No

When the Lives_In relation is transitive, then all relations must be annotated.

Example.
[…] sheep and goats in Europe […]

→ If a microorganism is located in “sheep” and “host”, then it is in “Europe”.

6.8 Selection media

Selection media are experimental media in which only microorganisms of one species grow. The relation between the *Microorganism* entity of this taxon entity and the selection media must be annotated. The relation between any other *Microorganism* entity and the selection media must not be annotated.
Here is a list of known selection media and the taxon that grows or survive there:

Selection medium	taxon
PALCAM	*Listeria monocytogenes*
LPM	*Listeria monocytogenes*
HCLA	*Listeria monocytogenes*

7 Exhibits relation

The `Exhibits` relation has two arguments:

- `Microorganism` must be a microorganism taxon name;
- `Phenotype` must be a phenotype mention.

The argument entities must be as close as possible graphically.
The `Exhibits` relation must be explicit within the scope of the document.

Example.
It has also shown that the origin of pigments seems essentially related to the presence of

yellow bacteria such as Arthrobacter species

→ There is an Exhibits relation between the “Arthrobacter” microorganism and the “yellow” phenotype.

Example.
[…] mesophilic heterofermentative Lactobacilli […]

→ Two Exhibits relations are annotated: the first one between “Lactobacilli” and “mesophilic” and the second one between “Lactobacilli” and “heterofermentative”.

7.1 Modality for mutants

`Exhibits` relations involving a mutant strain of a microorganism (as opposed to a wild-type strain) should be tagged using the specific modality “mutant”, which can be selected in the annotation editor.

8 Coreferences

The argument entities of `Lives_In` relations must be as close as possible graphically. In the case two or more entities are acceptable as a relation argument, then the equivalent entities must be part of a coreference group. The relation references either one of the entities in the group.
Coreference groups should only be made between equivalent entities occurring next to each other, in particular in the context of appositions.
Coreference groups must contain entities of the same type. Coreference groups of `Habitat` entities must contain entities that are all associated with the exact same OntoBiotope concept. Coreference groups of `Microorganism` entities must contain entities that are all associated with the exact same NCBI taxon identifier. Entities in a coreference group are not required to have the exact same surface form.