Guidelines for Annotation of Bacteria Biotopes

Robert Bossy, Claire Nédellec, Julien Jourde, Mouhammadou Ba

March 1, 2016

Contents

1 Introduction

This document specifies the guidelines for the annotation of the BioNLP-ST 2016 Bacteria Biotopes corpus. The task consists of the extraction of places where bacteria live in a set of PubMed abstracts. In concrete terms, this is specified by relations between bacterial taxon names on one hand, and habitats mentions and geographical names on the other hand.
Bacteria taxon names are organized by the *Bacteria* subtree of the NCBI Taxonomy.
Habitat mentions are organized by the *bacteria habitat* subtree of the OntoBiotope ontology.

1.1 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 bacterial taxa are denoted in *emphasis (italic font)*.
Examples are given in two parts: an annotated piece of text, then a comment about this annotation. The annotation in comments are always correct.

2 Bacterial taxon names

2.1 Entity domain

A bacterial taxon is a sub-taxon of the *Bacteria* phylum, formerly called *Eubacteria*. All taxon ranks are annotated, including genera, species, strains, serovars, families, classes, and all intermediary ranks.
Archaea, monocellular eukaryotes and all viruses are excluded.

2.1.1 Gram staining

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

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. Annotated abbreviations include:

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

Example 1.

. . . reservoirs of viable methicillin-resistant Staphylococcus aureus (MRSA). . . . Finding MRSA-contaminated surfaces on a variety of environmental surfaces in the absence of an overt outbreak . . .

→ “MRSA” is clearly introduced as an abbreviation as an apposition. Furthermore it is a globally understood abbreviation. Every occurrence in the document is annotated.

2.2 Boundaries

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

Example 2.

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 *Bacteria* annotation should not include the trailing or leading words "strain", "genus", "species", etc.

Example 3.

The strain 96-OK-85-24 significantly differed from the existing mosquitocidal B. thuringiensis strains.

→ The leading "strain" is excluded from the annotation.

Example 4.

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 5.

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

→ "pathogenic" is excluded.

Example 6.

[...] reservoirs of viable methicillin-resistant Staphylococcus aureus.

→ "methicillin-resistant" is excluded.

The words “bacterium”, “bacteria”, and “bacterial” in lower case are never annotated. If the authors emphazise on the phylum by capitalizing “Bacteria” then it is annotated.

Example 7.

Streptococcus salivarius is the principal commensal bacterium of the oral cavity in healthy humans.

→ “bacterium” is not annotated.

2.2.1 Phenotype acronyms

Phenotypes and qualifiers are not included in *Bacteria* annotations (see above). Acronyms that abbreviate both the phenotype and the species name must not be annotated with the following notable exceptions:

- “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 8.

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 *Bacteria* annotation. “MRSA” is annotated as a *Bacteria* since it is in the exception list.

2.2.2 Strain specification

When the species is followed by a strain name, then there must be a single annotation that contains the species and the strain names, including words like “strain”, “isolate”, or “serovar”.

Example 9.

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 10.

gram-negative plant pathogen Xanthomonas campestris pv. vesicatoria.

→ "pv." means "pathovar".

The following are considered as strain specifications:

- serovars
- serotypes
- mutants

Example 11.

[. . .] nonvirulent Ara+ Burkholderia pseudomallei isolates [. . .]

→ The mutation specification "Ara+" is included in the *Bacteria* 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 12.

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 13.

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 each occurrence of *Bacteria* entities. It informs the taxon identifier in the NCBI Taxonomy.

Example 14.

[. . .] reservoirs of viable methicillin-resistant NCBI_Taxonomy=1280 *Staphylococcus aureus* (NCBI_Taxonomy=1280 *MRSA*) *. [. . .] Finding* NCBI_Taxonomy=1280 *MRSA -contaminated surfaces on a variety of environmental surfaces in the absence of an overt outbreak [. . .]*

→ "MRSA" is clearly introduced as an abbreviation as an apposition. Furthermore it is a globally understood abbreviation. Every occurrence in the document is annotated.

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.

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 15.

XopC and XopJ, two novel type III effector proteins from NCBI_Taxonomy=456327 *Xanthomonas campestris pv. vesicatoria*. *. . . Both genes encode* NCBI_Taxonomy=456327 *Xanthomonas outer proteins (Xops) that were shown to be secreted by the TTS system.*

→ 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 16.

NCBI_Taxonomy=2742 *Marinobacter belong to the class of* NCBI_Taxonomy=1236 *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 bacteria habitats are expressions and phrases that denote a place where bacteria can live. 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.

3.1.1 Too generic

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

- "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”
- “underdevelopped countries”
- “vector”
- “world”

These words, if they are not attached to more precise modifiers must not be annotated. Note that “body” is not considered too generic.

3.1.2 Diseases, symptoms

Disease and symptom names do not denote bacteria habitats and must never be annotated.

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 non-healthy 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. Occurrences of the word "cell" must be annotated only if they denote a potential host cell. They must not be annotated if they denote bacteria cells.

Example 17.

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

→ "ureter cell" denotes an eukaryote cell and is thus annotated.

Example 18.

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 bacteria cells, and thus is not annotated.

Example 19.

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

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

3.1.4 Experimental materials and methods

Experimental material including devices, equipment, and media must be annotated except for the most commonly used molecular biology devices:

- “pulsed field gel” and variations of thereof
- “tube”

Experimental method names must never be annotated.

Example 20.

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

→ “pulsed-field gel” is not annotated even though it is an equipment, “pulsed-field gel electrophoresis” is not annotated because it is a method.

Example 21.

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 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 22.

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 23.

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 24.

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 25.

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

→ Here, qwater is a H_2O molecule.

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 bacteria living conditions.

Conversely, the boundaries shall exclude modifiers that are irrelevant to the bacteria. Excluded modifiers are:

- generic adjectives (“diverse”);
- relative adjectives or adverbs (“different”, “other”);
- cardinals and ordinals.

Example 26.

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 27.

[. . .] green algae Acrosiphonia [. . .]

→ A single annotation.

Example 28.

[. . .] blood-sucking tsetse fly [. . .]

→ A single annotation.

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

3.2.3 Appositions

Appositions are annotated separately.

Example 29.

[. . .] 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 30.

. . . 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 with the preposition "in" are not included in the *Habitat* annotation.

Example 31.

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

→ "Yemen" is not included in the annotation with "dyspeptic patients".

Example 32.

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

→ "Europe" is excluded from "sheep" and "goat" annotations since it refers to "mycoplasmosis".

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 33.

[. . .] oral, vaginal, and intestinal regions [. . .]

→ 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"
- "saprophyte"

Moreover, all organs and parts of living beings mentioned as adjectives must be annotated.
The "clinical" adjective must be annotated if it qualifies a bacterial 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:

- "phototroph"
- "methanotroph"

3.2.7 Overlapping habitats

Habitat mentions whose boundaries are contained in another one are omitted, 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 34.

[. . .] 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 35.

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 36.

[. . .] 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`. If a `Habitat` annotation is associated to several concepts, then it is assumed to be a conjunction.

Example 37.

`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 coreference, even partial, is assigned to the most precise concept.

Example 38.

After two weeks of treatment the authors found that 5 of 17

`duodenal ulcer`

chronic duodenal ulcers were healed in the remainder substantial

`patient with gastric ulcer`

`duodenal ulcer`

regression was found. In subjects with ulcers and bulbitis (12

`patient with gastric ulcer` `duodenal ulcer`

patients) the ulcer healed in 11 instances, in 7

`patient with gastric ulcer`

patients residual bulbitis persisted.

→ The "patient" entities are coreferences of "subjects with ulcers and bulbitis" hence they're assigned the same OntoBiotope concept *patient with gastric ulcer*. The same goes for "ulcer" entities that inherit the concept *duodenal ulcer*.

3.3.2 Creation of new synonyms and concepts

New synonyms and concepts can be created during the annotation.
New synonyms must be widely recognised 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 judgement.

4 Geographical names

Only geographical names are annotated as `Geographical`. Annotators are required to check if the mentioned names are present in gazeteers 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 39.

[. . .] 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 with a separate entity.

Example 40.

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 41.

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 42.

[. . .] African river blindness [. . .]

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

Example 43.

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.

→ None of “Injury”, “Glasgow”, and “chest” are annotated because they are all part of a name of a diagnistic method.

4.5 Geographical habitats

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

4.6 Nationalities and adjectives

Adjectives that relate to geographical places, like nationalities, are not annotated.

Example 44.

Eight patients shared a strain identical to a previously reported Australian transmissible strain (pulsotype 1).

→ “Australian” is not annotated.

5 Lives_In relation

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

- `Bacteria` must be a bacteria taxon name;
- `Location` must be either a habitat mention or a geographical name.

The argument entities must be as close as possible graphically.
The `Lives_In` relation must be explicit within the scope of the document. The bacteria 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 45.

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.

5.1 Topological constraints

The `Lives_In` relation does not carry strong topological semantics. The relation may be annotated wether the bacteria live on the surface, inside, or on top of the habitat.
However if the habitat is a cell, the relation is annotated only if the bacteria is inside the cell. The relation is not annotated if the bacteria adhere to the cell membrane or bind to a surface protein.

Example 46.

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

→ "myelinated Schwann cells" is a cell and the bacteria do not live inside, so the relation is not annotated.

Example 47.

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

→ Here the bacteria are explicitely said to be living inside the qhuman cells, thus the relation is annotated.

5.2 Partial localization

The relation is not annotated if only part of the taxon is mentioned as living in the localization. However observations and experiments can be generalized for the whole taxon.

Example 48.

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 49.

Marinobacter belong to the class of Gammaproteobacteria and these motile, halophilic or halotolerent 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.

5.3 Effect of bacteria on the environment

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

5.3.1 Diseases and symptoms

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

- "abscess"
- "colonization"
- "commensality"
- "infection"
- "invasion"

Example 50.

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 bacteria causes a symptom ("gastric atrophy") but the bacteria is not necessarily located in the stomach ("gastric").

Example 51.

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 bacteria and the skin is not established, even though it causes a symptom on the skin.

5.3.2 Symbioses

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

Example 52.

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 53.

S. glossinidius is an endosymbiont of the tsetse fly.

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

5.4 Experimental settings

Descriptions of bacteria 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 54.

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

→ "California" is not annotated, it just the place where an experiment was conducted.

5.5 Vaccines

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

5.6 Hypothesis sentence

Utterances for a working hypothesis must not be annotated as a Lives_In relation. Positive evidence of the hypothesis further in the documet must be anotated 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 55.

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).

5.7 Relation transitivity

This section covers the case where a bacteria 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

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

Example 56.

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

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

5.8 Selection media

Selection media are experimental media in which only bacteria of one species survive. The relation between the *Bacteria* entity of this taxon entity and the selection media must be annotated. The relation between any other *Bacteria* entity and the selection media must not be annotated.
Here is a list of knwon selection media and the taxon that survives:

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

6 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 must never cross paragraph boundaries.
Coreference groups must contain entities of the same type. Coreference groups of Habitat entities must contain entities all associated with the exact same OntoBiotope concept. Coreference groups of Bacteria entities must contain entities all associated with the exact same NCBI taxon identifier. Entities in a coreference group are not required to have the exact same surface form.