Guidelines for Annotation of Bacteria Biotopes

Robert Bossy, Claire Nédellec, Julien Jourde

December 2, 2010

Contents

1 Introduction **2**

2 Bacterial taxon names **2**

3 Mentions of localization relations **2**
3.1 Coreferences . . . 3
3.2 Partial localization . . . 3
3.3 Effect of bacteria on the environment . . . 4
3.3.1 Diseases . . . 4
3.3.2 Host-part symptoms . . . 4
3.3.3 Localized diseases . . . 4
3.3.4 Symbioses . . . 4
3.4 Localization transitivity . . . 5
3.5 Experiments and artificial location . . . 5

4 Localizations **5**
4.1 Adjectival forms . . . 6
4.2 Localization types . . . 6
4.2.1 Hosts . . . 6
4.2.2 Host part . . . 6
4.2.3 Geographical . . . 7
4.2.4 Food . . . 7
4.2.5 Medical . . . 7
4.2.6 Soil . . . 7
4.2.7 Water . . . 8
4.2.8 Environment . . . 8
4.3 Boundaries . . . 8
4.3.1 Location properties . . . 8
4.3.2 Enumerations . . . 8
4.3.3 Disambiguation of geographical names . . . 9
4.4 Excluded locations . . . 9
4.4.1 Environment properties . . . 9
4.4.2 Too general . . . 9
4.4.3 Experimental growth substrates . . . 9
4.4.4 Names containing location names . . . 9

1 Introduction

This document specifies the annotation guidelines for Bacteria Biotopes. Annotators abided by the rules of these guidelines in order to annotate the training and test corpora for the Bacteria Biotopes BioNLP Shared Task 2011. The task consists of the extraction of places where bacteria live in a set of defining textbook documents. In concrete terms, this is specified by relations between bacterial taxon names on one hand and localization mentions on the other hand.

Localizations are organized into different localization types:

- host: living beings
- host-part: part of living beings
- geographical: named administrative zone
- environment: place specified by environment properties, which has four sub-types:
  - food: processed human or animal food
  - water: environmental water
  - soil: soil
  - medical: medical places

Additionally, if a relation between a bacteria and a localization of type host-part is annotated, the host to which the host-part belongs must be annotated as well.

2 Bacterial taxon names

A bacterial taxon is a sub-taxon of the *Bacteria* domain, formerly called *Eubacteria*. *Archaea*, monocellular eukaryotes and all viruses are excluded.

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.

> "*The type strain of the species is* ***Bradyrhizobium japonicum USDA 6*** *which was isolated from [...]*"

Example 1: the annotation includes the genre, species and strain.

3 Mentions of localization relations

This section specifies which relations must be annotated.

The localization relation must be explicit within the scope of the document. The bacteria must be alive in the mentioned localization. This section specifies the minimal shared knowledge which is necessary to annotate a localization relation. If a knowledge not mentioned here is necessary, then the relation must not be annotated.

When the localization is implicit but is certain from the point of view of a reader, then the relation must be annotated:

> "***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.*"

Example 2: the localization in "North America" is implicit but certain, it must be annotated

3.1 Coreferences

The scope of the mention of relations is the document, thus it may span over several sentences, even several paragraphs. This means that documents may contain many coreferences. The relation argument must be the most precise coreference:

- pronoun coreferences must be resolved, a relation argument may never be "it", "these", "they", etc.
- higher-level taxon name references must be resolved, for instance:
  - if "these bacteria" refer to a named taxon elsewhere in the document, then the relation argument must be the precisely named taxon;
  - if a bacterial genre name is used, but refers to a species (ellipsis of the specific name), then the relation argument must be an occurrence of the name *with* the specific name.

> "***Streptococcus salivarius*** *is the principal commensal* ***bacterium*** *of the oral cavity in healthy humans*"

Example 3: the relation is between "oral cavity" and "Streptococcus salivarius", not "bacteria"

.

In some cases several occurrences of the same object may equivalently be an argument of a relation; the equivalence must be annotated. Indeed the task evaluation will take into account a relation between equivalent occurrences as equivalent relations. Equivalent occurrences must have the same level of specificity, although this does not mean that they necesserily have the same form. In the case of the abbreviation of a taxon generic name ("B. anthracis"), it is accepted as an equivalent form of the name with the full generic part ("Bacillus anthracis").

3.2 Partial localization

The relation is annotated even if only part of the taxon is mentioned as living in the localization.

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

Example 4: the relation between "Yersinia" and "human" is annotated even is only sub-taxa are said to be human pathogens.

3.3 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:

3.3.1 Diseases

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.

3.3.2 Host-part symptoms

Pathogen bacteria are often described by the disease they cause, in particular the symptoms on the host parts. 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 plant parts 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

3.3.3 Localized diseases

Location of diseases are not common knowledge. Localized diseases must not be annotated as host parts.

"*streptococci are the most common causes of endocarditis*"

Example 5: "endocarditis" is not annotated as an host part (heart)

3.3.4 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").

"*This nitrogen-fixing bacterium develops a symbiotic relationship with the soybean plant Glycine max*"

Example 6: no relation between a bacterium and "Glycine max" shall be annotated because this symbiosis relationship may be on nutrient exchange basis only

"***S. glossinidius*** *is an endosymbiont of the tsetse fly*"

Example 7: this relation is annotated since the "endo-" prefix makes the bacterium localization explicit

3.4 Localization transitivity

Some bacteria localizations are said to be located in another location. Depending on the type of the two localizations, the relation between the bacterium and the second localization must be annotated or not:

Type of first localization	**Transitivity of the relation**
host	only if the second localization is geographical
host-part	only if the second localization is host or another host-part
geographical	only if the second localization is geographical
environment including all subtypes	always transitive

"*sheep and goats in Europe*"

Example 8: if a bacterium is located in "sheep" and "host", then it is in "Europ"

3.5 Experiments and artificial location

Experiments where bacteria are placed in unnatural environments are not annotated. The following relations are excluded:

- growth on rich substrates (peptone, galactose, glucose, lactate, acetate, etc.);
- vaccines: most vaccines contain dead bacteria, living bacteria vaccines are not annotated either;
- the geographical place where the laboratory where experiments are conducted is located.

4 Localizations

This section specifies which localization mentions are candidates to be the argument of a relation and which are not. It also specifies the contour of each localization type and the boundaries of the annotations.

"*Maronti Beach*, *Ischia*, *Italy*"

Example 9: bacteria lives in all places mentioned in hierarchical geographical names

"*an isolate of this species was studied by researchers at University of California*"

Example 10: "California" is not annotated, it just the place where an experiment was conducted

A localization is generally a complete noun phrase including its head and some of its modifiers. The exceptions are adjectival localization forms listed in 4.1 and some enumerations specified in 4.3.2

4.1 Adjectival forms

The only adjectives to be annotated are listed here:

- aquatic
- enteroinvasive
- epiphytic
- extraintestinal
- foodborne
- intracellular
- marine
- nosocomial
- phytopathogen
- saprophyte

4.2 Localization types

4.2.1 Hosts

A host must be alive to be annotated with the type host. If the mentioned living being is dead, it is either an environment or a food.

4.2.2 Host part

A localization of type host-part is something bound to a living being. It includes the host living parts: organs, tissues, cells and organelles. It also includes non-living parts:

- abcesses
- excretions

- fluids
- phyllome
- rhizome
- secretions
- tumors
- wounds

If the localization is detached from its host, or if the host is not alive, then the annotation is either an environment or a food.

"*in abusive-chilled retail cooked meat*"

Example 11: "meat" is not a host part since the host is manifestly dead

"*It is also isolated from the carcasses of dead animals*"

Example 12: "carcasses" is not a host part since the animals are explicitely dead

"*This bacterium can remain in the excrement of cattle and the aborted fetuses of the cattle for quite some time*"

Example 13: "excrement" and "aborted fetuses" are not host parts since they were explicitely detached from the animal's main body

4.2.3 Geographical

Only geographical names are annotated as geographical. Annotators are required to check if the mentioned names are present in gazeteers and administrative name lists.

4.2.4 Food

Localizations of type food are processed food for humans or animals. Food also includes drinking water. Food excludes ingested bodies which were not meant for nutrition. Food excludes non-animal (i. e. plant) nutrients.

4.2.5 Medical

Medical localizations are places where or instruments with which medicine is practiced. This excludes the medical personnel which are considered as hosts.

4.2.6 Soil

Soil localizations are layered solid mineral environments. Soil excludes sediments and dust in the air, which are environments.

4.2.7 Water

Localizations of type water are environmental waters (sea, river, spring, etc.). Water excludes drinking water, which is a food.

4.2.8 Environment

Environment is the default type for localizations. If a localization does not fit any other type.

Environment notably includes:

- climatic zones
- sediments

4.3 Boundaries

4.3.1 Location properties

The boundaries of the localization annotation must include all specific modifiers relevant to the bacteria living conditions. Conversely the boundaries shall exclude modifiers that are irrelevant to the bacteria.

"*chromate and xylene enriched soil microcosms*"

Example 14: the "microcosm" modifiers are all relevant for bacteria who live there, thus they are included in the annotation span

4.3.2 Enumerations

When several localizations are enumerated, there are two cases:

The factorized part is important : a single annotation covering all the enumeration including the factorized part.

"*wild and domesticated animals*"

Example 15: the factorized part "animals" is crucial for the understanding of the bacteria loacalization, thus the whole enumeration is included in a single annotation

The factorized part is purely grammatical : one annotation for each item.

"*oral*, *vaginal*, *and* *intestinal* *regions*"

Example 16: the factorized part "region" is not informative with regard to the bacteria loacalization, thus each item is covered by a different annotation even if they consist in adjectives

4.3.3 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.

"*island of Malta*"

Example 17: the annotation includes "island" in order to distinguish between the main island of Malta and the state of Malta

4.4 Excluded locations

4.4.1 Environment properties

The ecology of bacteria is often described by the optimal and observed conditions for their growth. These mentions are **not** localizations. The following nouns are generally not considered as localization anchors:

- condition
- environment
- medium
- place
- property

"*as a means to survive for much longer periods of time in the aerobic environment*"

Example 18: "anaerobic environment" because it is not a characterized place

4.4.2 Too general

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

- host
- nature
- underdevelopped countries

Note that "body" is not considered too general, it must be annotated as a host part.

4.4.3 Experimental growth substrates

See 3.5.

4.4.4 Names containing location names

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

"*African river blindness*"

Example 19: "African" is part of a disease name, thus not annotated as a localization

"*Tick-borne encephalitis viruses*"

Example 20: "Tick" is part of a virus name, thus not annotated as a localization