Documentation for S1000 corpus annotation

Original Curator Guidelines for S800 corpus

The annotation guidelines as described in the original publication of the S800 corpus

The guidelines to curators were to annotate all substrings, which can meaningfully be identified as *referring to a taxon*. While the main focus was on annotating **species** mentions, strings referring to **any taxonomic level**, (e.g.kingdoms, orders, genera, strains) were also considered. The data for upper taxonomic level annotations were never officially released.

The main guidelines were:

- All document substrings must be evaluated and all mentions including repetitions should be listed in the order of appearance in the text.
- The annotated name types among others should include: Linnaean binomials, common names, strain names, author defined acronyms.
- For each annotated string, curators must record the name as it appeared in text and report the corresponding NCBITaxonomy database identifier.
- Special cases of adjectives being used to indicate a taxon, misspellings, typographic or other errors and enumerations were indicated as such.
- Taxonomic mentions that did not correspond to an existing NCBI Taxonomy database entry were also indicated.

Additional guidelines for the annotation of the S1000 corpus

- The first resource that is trusted to resolve issues is NCBI Taxonomy. If there is still not enough information there we have resolved inconsistencies using the Catalogue of Life, ICTV, ITIS, Avibase and WoRMs.
- The NCBI Taxonomy Ranking has been adopted from Schoch, *et. al*, 2020

General annotation guidelines

- Taxonomic mentions that correspond to **Species**, **Genus** and **Strain** will receive annotations and normalization to taxonomy reference databases (priority will be the normalizations to NCBI taxonomy, and when that's not possible the resources mentioned above will be used)
- Adjectival forms like **murine** (**taxid:10090**), **bovine** (**taxid:9913**), **pneumococcal** (**taxid:1313**) that map to a specific species should be annotated as such
- The role in which common species names are mentioned should **not** be taken into account and all species names mentions should be annotated (e.g. *rice* mentioned as food or *tobacco* as cigarettes should still be annotated).
- Genus or higher level mentions (e.g. *Arabidopsis*, *yeast*) should only be annotated as the real taxinomic level (or not annotated at all), and not as synonyms of species names. (e.g. *Arabidopsis* should be annotated as **Genus** and assigned the genus **taxid:3701**)

```
1 The second face of a known player: Arabidopsis (Genus) silencing suppressor AtXRN4 acts organ-specifically
```

- Former **Species** annotations in the original S800 coprpus that belong to the **Genus** taxinomic rank have been annotated as the latter in the corpus. Ranks higher than *Genus* (e.g. *Phylum*, *Kingdom*, *Class*, *Order*, *Family*) should receive an **OOS** annotation or not be annotated at all
- For annotations above *Species* only the "coarse" ranks should be considered, thus mapping mentions at fine-grained levels to their coarse equivalents, e.g. *Subgenus* should be mapped back to *Genus*.
- For **Subspecies** mentions: when a subspecies name immediately follows a species name the entire mention is simply annotated as one slightly longer **Species** mention, e.g. *Phocoenoides dalli dalli* annotated as **Species** + **taxid: 9745** (Rank: Subspecies).
- **Biotypes** should be treated the same way as **Subspecies**, i.e. they are annotated as **Species**
- **common name** (*scientific name*)" mentions should be annotated as **two mentions** e.g from PMID: 21054435:

```
2 We studied seasonal dynamics in delta^1^3C of CO2 efflux (delta^1^3C(E)) from non-leafy branches, upper and lower trunks
  and coarse roots of adult trees, comparing deciduous Fagus sylvatica (Species) (European beech (Species)) with evergreen Picea abies (Species)
  (Norway spruce (Species)).
```

- Species names in noun phrase premodifier positions (e.g. **Arabidopsis** EDR1, **Aspergillus nidulans** cells) also in cases where they appear as part of the name of an entity of a non-organism type (e.g. **human** epidermal growth factor receptor 2 (HER2)) are annotated.
- Species names are annotated when they are part of **hyphenated compound words** (e.g. *human-infecting*)
- **Clade** mentions will receive **Clade** normalizations and will be assigned type according to nearest **non-**

Clade ancestor (if that falls within the scope of the current annotation effort)

- Similarly, **no rank** mentions will receive **no rank** normalizations and will be assigned type according to nearest **ranked** ancestor (if that falls within the scope of the current annotation effort)

Rules for common names

- In general, when a **Species** and a higher-level entry in the taxonomy (e.g. **Genus**) share a common name or synonym, the **Species interpretation should be preferred** when it is not clear from context which is intended.
- Common names like **human**, **goat**, **horse**, and **rats** should be **always** annotated.
- Common names that should be annotated in the **Genus** level:
  - **fire ant**: **Genus** and **taxid:13685** (**Solenopsis**); Note: red fire ant, little fire ant, black fire ant etc should be tagged as the corresponding species)
  - **sunflower**: **Genus**+**taxid:4231** (**Helianthus**)
  - **galaxias** : **Genus**+**taxid:51242** (**Galaxias**)
- Common names that should be annotated in the **Species** level (but could be annotated in a higher taxonomic level):
  - **rat**: synonym for *Rattus norvegicus* and **Rattus**. Should be annotated as *Rattus norvegicus* (**taxid:10116**), unless explicitly referring to a different taxonomic unit (e.g. **cotton rat**: **Genus** + **taxid:42414** (**Sigmodon**))
  - **fruit fly**: synonym for *Drosophila melanogaster* and **Drosophila** genus and **Tephritidae** family. Should be annotated as *Drosophila melanogaster* (**taxid:7227**), unless explicitly referring to a different taxonomic unit
  - **bee**: synonym for *Apis mellifera*, and **Apoidea** superfamily. Should be annotated as *Apis mellifera* (**taxid:7460**), unless explicitly referring to a different taxonomic unit (e.g. **bumble bee**)
  - **duck**: synonym for *Anas platyrhynchos*, but can be a synonym for other **Anatidae**. Should be annotated as *Anas platyrhynchos* (**taxid:8839**), unless explicitly referring to a different taxonomic unit
  - **midge**: synonym for *Chironomus thummi*, but can refer to several species of flies. Should be annotated as *Chironomus thummi* (**taxid:7154**), unless explicitly referring to a different taxonomic unit

Mentions that should **NOT** receive annotations

- Adjectival forms of Phyla (e.g. **cyanobacterial**: **taxid:1117**) can only be annotated as **OOS** or not be annotated at all
- Adjectival forms of **Kingdoms** (e.g. **viral**, **bacterial**) can only be annotated as **OOS** or not be annotated at all
- Non-name mentions (e.g. **woman**) and species clues (e.g. **patients, children, men, women**) should not be annotated. This includes the non-name mention **man** which should not be annotated as a synonym for *Homo sapiens* (**taxid: 9606**)
- Mentions that are **not monophyletic** (e.g. *fish*) should be annotated as **Out-of-scope (OOS)** with *Note: not monophyletic* or not be annotated at all
- Forms identified by place names, like **ecotype**, are not annotated.

```
3 For investigating cadmium uptake, we incubated protoplasts obtained from leaves of Thlaspi caerulescens (Ganges ecotype) with a Cd-specific fluorescent dye.
```
(Species: Thlaspi caerulescens)

- Species names are **NOT** annotated when they appear as a **substring in a word not separated by a boundary** such as a hyphen (e.g. *nonhuman*)
- Abbreviations are marked if the abbreviation stands for an organism mention in scope of the annotation, but not if the full form merely includes an organism mention e.g. in modifier position. For example, the **H** in **HER2** is not annotated despite it standing for **human**.
- *Cultivars* should be annotated as **OOS** or not be annotated at all.

```
4 The physiological traits underlying the apparent drought resistance of 'Tomatiga de Ramellet' (TR) cultivars.
```
(Out-of-scope: 'Tomatiga de Ramellet')

- *Rootstocks* should be annotated as **OOS** e.g. in PMID: 20837155
- Non-taxonomic groupings such as **Gram-positive/negative bacteria**, **marine bacteria** or **enteric bacteria** should not be annotated. e.g.

```
5 The redox-sensitive transcription factor SoxR in enteric bacteria senses and regulates the cellular response to superoxide and nitric oxide.
```

```
6 Oscillochloris trichoides is a mesophilic, filamentous, photoautotrophic, nonsulfur, diazotrophic bacterium which is capable of carbon dioxide fixation via the reductive pentose phosphate cycle and possesses no assimilative sulfate reduction.
```
(Species: Oscillochloris trichoides)

- **tree** and **bush** are non-taxonomic mentions and thus not annotated or annotated as **OOS** + *Note: non-taxonomic*
- Standalone **alga** (**algae**, **microalgae**, **macroalgae**): can only be annotated as **OOS** + *Note: non-taxonomic* or not be annotated at all e.g. Algae is an informal term for a large and diverse group of photosynthetic eukaryotic organisms.
- **protist** (any eukaryotic organism that is not an animal, plant, or fungus) is a non-taxonomical expression and can be annotated as **OOS** + *Note: non-taxonomic* or not be annotated at all e.g.
- **protozoa** can also be annotated as **OOS** + *Note: non-taxonomic* or can not be annotated at all
- **methanotroph** is a non-taxonomical expression and can be annotated as **OOS** + **no taxid** or not be annotated at all e.g.

- **methanogen(s)** over 50 *Archaea* species can be annotated as **OOS** with *Note: not monophyletic* or not be annotated at all
- **prokaryotes** includes *Bacteria* and *Archaea* in the current three-domain system, so this can be annotated as **OOS** + **no taxid** or not be annotated at all e.g.
- **heterokonts** and **alveolates** are clades of microorganisms and can be annotated as **OOS** + **taxid:33634** and **taxid:33630** respectively or not be annotated at all e.g. and e.g.
- **cyanobacteria**, **eubacteria** and the like can be annotated as **OOS** or not be annotated at all unless it's clear from context that the reference is definitely to the genus **Cyanobacterium** or **Eubacterium** respectively.
- **Non-taxonomic groupings** of organisms by their behaviour (e.g. *herbivores*, *predators*, *parasites*) should not receive an annotation or should receive an **OOS** annotation
- **actinorhiza(l)**, **mycorrhiza(l)**, **ectomycorrhiza** can be **OOS** + *Note: non-taxonomic* or not be annotated at all
- **species complex** and **clonal complex** rank can be **OOS** or not be annotated at all

Rules for common names

- Common names that should not be annotated in the **Species** level:
  - **ant(s)**: **OOS**+**taxid:36668** (**Formicidae**) or no annotation
  - **insect(s)**: when standalone assign **OOS**+**taxid:50557** (**Insecta**) or no annotation
  - **mite**: **OOS**+**taxid:6933** (**Acari** subclass) or no annotation
  - **trout**: several species of fish, annotate as **OOS** + **no taxid** or no annotation
  - **leafminer** and **leaf miner**: insects that eat the tissue of plants, annotate as **OOS** + *Note: non-taxonomic* or no annotation
  - **fishes**: **OOS** (Clade-like concept, non-tetrapoda vertebrata) or no annotation
  - **bug**: **OOS** + *Note: non-taxonomic* or no annotation
  - **field cricket**: **OOS** + *Note: non-taxonomic* or no annotation
  - **mirid bug**: **OOS**+**taxid:30084** (**Miridae**) or no annotation
  - **clownfish**: **OOS**+**taxid:30863** (**Pomacentridae**) or no annotation
  - **elephant**: 3 species, not monophyletic (both **Elephas** and **Loxodonta** genera), annotate as **OOS** + **no taxid** or no annotation
  - **crab**: infraorder containing 850 species, so it should be annotated as **OOS** + **taxid:6752** (**Brachyura**) or no annotation
  - **grass**: **OOS**+**taxid:4479** (**Poaceae**) or no annotation
  - **seabird(s)**: **OOS** with *Note: non-taxonomic* or no annotation
  - **marsupial** (animals carry the young in a pouch) is a mammalian clade, e.g. and will be annotated as **OOS** + **taxid:9263** or not annotated at all
  - **coral(s)**: Hexacorallia + Octocorallia, but paraphyletic because sea anemones are also part of Hexacorallia: annotated as **OOS** with *Note: not monophyletic* or not annotated
  - **DNA viruses**, **RNA viruses** map to no rank entries: annotated as **OOS** or not annotated at all
  - **dsRNA mycoviruses**: **OOS** with *Note: non-taxonomic*
  - **cereal**: **OOS** with *Note: non-taxonomic*
  - **kittiwake**: **OOS** and *Note: non-taxonomic*
- **Young animals** (e.g. chicks, calfs etc) should not receive an annotation or should receive an **OOS** annotation

Special rules for Strains

- Strain aliases such as **CC-12301**(T) (=**DSM 45298**(T) =**CCM 7727**(T)) should be annotated in all instances as type **Strain**.
- *name strain* mentions should be annotated as **two mentions** of **Species**+**Strain**, e.g. from PMID: 20154326

```
7 Strain GSW-R14(T) exhibited 97.6 % 16S rRNA gene sequence similarity to F. gelidilacus LMG 21477(T) and similarities of
  91.2-95.2 % to other members of the genus Flavobacterium
```


- mentions of the form *[Genus] sp. [Strain]*, should have a separate **Genus** and **Strain** annotation e.g.
- descriptive references to *Strains* using gene names are not annotated as organisms e.g.

Special rules for Viruses

- **Viruses** (or other taxonomic units) that have species level of entry as "unidentified" (e.g. "retrovirus" **taxid:31931** ("unidentified retrovirus" equivalent: "retrovirus") or "adenovirus" **taxid:10535** ("unidentified adenovirus" equivalent: "adenovirus")) should **NOT** be annotated in the **Species** level.
- The following mentions should be annotated as **OOS** or not be annotated at all:
  - "virus"/"viral" **OOS**+**taxid:10239** "Viruses" *superkingdom*
  - "retrovirus" **OOS**+**taxid:11632** "Retroviridae" *family*
  - "influenza virus" **OOS**+**taxid:11308** "Orthomyxoviridae" *family*
  - "herpesvirus" **OOS**+**taxid:10292** "Herpesviridae" *family*
  - "adenovirus" **OOS**+**taxid:10508** "Adenoviridae" *family*
  - "baculovirus" **OOS**+**taxid:10442** "Baculoviridae" *family*
  - "reovirus" **OOS**+**taxid:10880** "Reoviridae" *family*
- The following mentions should be annotated as **Genus**:

    - “norovirus” **Genus+taxid:142786** “Norovirus” *genus*
    - “ebola virus” **Genus+taxid:186536** “Ebolavirus” *genus*
    - “cytomegalovirus” **Genus+taxid:10358** “Cytomegalovirus” *genus*
- **dengue**: dengue is synonym for dengue fever (disease), annotate as **OOS** + **no taxid** unless *dengue virus* is mentioned when it should be annotated as **taxid:12637** (**Species**)
- **smallpox**: smallpox is synonym for smallpox disease, annotate as **OOS** + **no taxid** unless *smallpox virus* is mentioned when it should be annotated as **taxid:10255** (**Species**)
- **influenza**: influenza is synonym for the flu (disease), annotate as **OOS** + **no taxid** unless *influenza X virus* is mentioned when it should be annotated as **Species**. EXCEPTION: **standalone influenza** may be marked when organism sense is clear from context (e.g. **influenza strains**)
- **human adenovirus** (or similar cases): when a mention cannot be normalized in an “identified” virus species it should be annotated e.g. as **Species+taxid:9606** (*Homo sapiens*) for **human** and **OOS+taxid:10508** (*Adenoviridae*) for *adenovirus* (or no annotation for the latter)

Special rules for Yeasts

- All text spans including “yeast” should have an **OOS** annotation if the taxonomy level is higher than **Genus** or should not be annotated at all:
    - standalone *yeast*: **OOS+taxid:147537** (“true yeast” subphylum) (Note: an even higher level may be included)
    - *black yeast*: **OOS+taxid:34395** (“black yeast” order)
    - *budding yeast*: **OOS+taxid:4892** (“budding yeasts” order)
    - *fission yeast*: **OOS+taxid:4894** (“fission yeasts” family)
    - *truffle*: **Genus** + **taxid:36048** (*Tuber* genus)

Special rules for Amoebae

- All **amoebae** instances have been revised to resolve confusion of non-taxonomical expression **amoebae** (type of cell or unicellular organism which has the ability to alter its shape), of **taxid:554915** (**OOS**: *Clade: Amoebozoa*), and **taxid:55774** (**Genus**: *Amoebae*). Most of the cases were non-taxonomical expressions (**OOS** + **no taxid**)
    - **testate amoebae**: very common combination of mentions, which means *shelled amoebae*, which explains the form of microorganism(s): **OOS** + **no taxid** or no annotation
    - Interesting article PMID: 21112814, where both *non-taxonomical* and *Genus* amoebae are mentioned (only one real “amoebae” Genus in the corpus)

Special cases

- Four mentions of “astomes” in this document PMID: 21398102 are **OOS**
- Astome ciliates in this document PMID: 21398102 are also **OOS**
- FGSC should not be annotated as it refers to a complex which is **OOS**, namely Fusarium graminearum complex PMID: 22004876
- Mentions of carnivores in PMID: 21323921 are **OOS** (interpreting these to refer generally to meat-eating animals)
- **human** and **primates** in a context of **non-human primates** are considered **two mentions** 21295520
- PMID: 2435057 is discussing retroviruses, but terminology there is quite old (published in 1987). ICTV (International Committee on Taxonomy of Viruses) was used to figure out how those viruses are called/classified in that period tracing its history.
- **GII.4** in PMID: 20980508 has been annotated as **Species**, following the general rule about **Clade** mentions
- **arbuscular mycorrhizal fungi** (**AMF**) e.g. in PMID: 20880038 is **OOS**
- **tropical japonica rice** (e.g. PMID: 20946420): following rule about **no rank** entries: normalization to NCBI txid: 1736656 and type **Species**

Span consistency guidelines

- The expressions *sp. nov.* and *gen. nov., sp. nov.* are not included in the **Species** name, since these are supposedly used only the first time a *genus* and/or a *species/subspecies* is described to denote that it’s new, so they are not part of the scientific name and shouldn’t be found anywhere else other than the first paper describing them.
- An annotated span should **not** end with *sp.* or *spp.*
- Superscript T - to denote type strain - should **not** be included in species’ names
- The person’s name should **not** be included in the species name, especially when it is in parentheses. The non-parenthesized form is a bit more complex (at least in the example above *Pseudacteon tricuspis* Borgmeier is a valid name shown as a synonym for *Pseudacteon tricuspis* in NCBI taxonomy). For annotation consistency the suggestion is to **drop these names in all appearances**. (The confusion with subspecies can be avoided because of the capital letter at the start of the second word, e.g. *Ursus arctos arctos* would be easy to distinguish from *Ursus arctos* Linneaus and then drop the name for the latter.)
- Do **not** include common head nouns such as “plants” in annotation spans
- Do **not** include adjectival premodifiers such as “native” in annotation spans
- Model words like **SCID** mouse should be excluded from annotations
- *“species complex”* should **not** be part of a species name, e.g. from PMID: 20682355

8 The splicing activity of the PRP8 intein from the B. dermatitidis, E. parva and P. brasiliensis species complex was demonstrated in a non-native protein context in Escherichia coli.

- *f. sp.* (forma specialis) should be included in the annotated mention (e.g. *Blumeria graminis f. sp. tritici*)
- Do **not** include nouns identifying levels of taxonomy in annotation spans. For example, the words **strain**, **serotype**, **serovar**, and **serogroup** should be excluded from the spans of annotated *Strain* mentions. e.g from 20154326

9 Strain GSW-R14(T) exhibited 97.6 % 16S rRNA gene sequence similarity ...

- Annotate antibodies e.g **anti-HCV** with species annotation for the organism (**HCV**) and *Note: "anti-" prefix*

For information on Annodoc, see http://spyysalo.github.io/annodoc/.