Clear Earth Annotation Guidelines: Biology

(February 2018)

Anne Thessen & Jenette Preciado

Annotating Entities & Eventualities

Key for Annotation Examples:

- [brackets] represent annotation spans.
- [ ]$_{\text{SUBSCRIPT}}$ represents the annotation categories applied to annotation spans.
- The underscore inside an entity or property name may read as 'and' or 'or.'
- Underscores underneath multiple words represent the proper length of a span for discontinuous spans.

Annotation Categories:

Eventuality

Refers to any mention, generic or specific, of a process, event, or state related to biological processes. This include any event or state that can occur within an abiotic or biotic entity, to an abiotic or biotic entity, or between abiotic and abiotic entities, biotic and biotic entities, or abiotic and biotic entities. They typically occur over time and often, but not always, have inputs and outputs.

Eventualities can be annotated wherever they occur, whether or not they are the syntactic head of a span, and can be any part-of-speech. Typically, they are single word spans. Their pre-modifiers can be annotated too if they are relevant to one of the annotation categories listed below or if a trait or eventuality is modifying the eventuality, though in this later case the entire expression is annotated as a MWE.

Any eventuality that is performed, directed, or experienced by a person, computer, or observational instrument should be left unannotated as we are only interested in capturing those eventualities that are related to biological processes.

Examples:

- The [sunflower]$_{\text{BIOTIC\_ENTITY}}$ [grew]$_{\text{EVENTUALITY}}$.
- The [fox]$_{\text{BIOTIC\_ENTITY}}$ [ate]$_{\text{EVENTUALITY}}$ the [bird]$_{\text{BIOTIC\_ENTITY.}}$

- [Kingbirds]$_{BIOTIC\_ENTITY}$ [attack]$_{EVENTUALITY}$ other [birds]$_{BIOTIC\_ENTITY}$ that come too close.

- The [top]$_{QUALITY}$ [trophic level]$_{BIOTIC\_ENTITY}$ of the [food web]$_{ABA\_ENTITY}$ has [top or [apex predators]$_{BIOTIC\_ENTITY}$]$_{BIOTIC\_ENTITY}$] which no other [species]$_{BIOTIC\_ENTITY}$ [kills]$_{EVENTUALITY}$ directly for its [food resource]$_{BIOTIC\_ENTITY}$ [needs]$_{EVENTUALITY}$.

- Therefore, [viruses]$_{BIOTIC\_ENTITY}$ in the [microbial food web]$_{ABA\_ENTITY}$ [act]$_{EVENTUALITY}$ to [reduce]$_{EVENTUALITY}$ the [population]$_{BIOTIC\_ENTITY}$ of [bacteria]$_{BIOTIC\_ENTITY}$ and, by [lysing]$_{EVENTUALITY}$ [bacterial]$_{BIOTIC\_ENTITY}$ [cells]$_{BIOTIC\_ENTITY}$.

- Soil ecologists discovered that [omnivory]$_{QUALITY}$ in [food webs]$_{ABA\_ENTITY}$ was common, and that [food chains]$_{ABA\_ENTITY}$ could be [long]$_{QUALITY}$ and [complex]$_{QUALITY}$ and still remain [resistant]$_{EVENTUALITY}$ to [disturbance]$_{EVENTUALITY}$ by drying[1], freezing, and fumigation.

- [Fire]$_{EVENTUALITY}$ [frequency]$_{QUALITY}$ in the [communities]$_{BIOTIC\_ENTITY}$ where [lavender thistle]$_{BIOTIC\_ENTITY}$ [occurs]$_{EVENTUALITY}$ is generally [low]$_{QUALITY}$. In [creosote bush scrub communities]$_{BIOTIC\_ENTITY}$, [fires]$_{EVENTUALITY}$ generally [occur]$_{EVENTUALITY}$ only [in years]$_{TIME}$ when exceptionally [heavy]$_{QUALITY}$ [winter]$_{TIME}$ [rains]$_{EVENTUALITY}$ have [produced]$_{EVENTUALITY}$ abnormally [high]$_{QUALITY}$ [numbers]$_{QUALITY}$ of [annuals]$_{BIOTIC\_ENTITY}$.

- [Caribou]$_{BIOTIC\_ENTITY}$ [migrate]$_{EVENTUALITY}$ [between summer and winter]$_{TIME}$ [habitats]$_{ABA\_ENTITY}$. [Spring]$_{TIME}$ [migration]$_{EVENTUALITY}$ [begins]$_{EVENTUALITY}$ [as early as mid-February]$_{TIME}$ and is typically [completed]$_{EVENTUALITY}$ [by June]$_{TIME}$.

Biotic_entity

Refers to a living organism, group of living organisms, or substance derived from a living organism. Most contain carbon and are capable of decay.

It is common to find biotic entities modifying other biotic entities. Frequently they appear as common names modifying a generic biotic entity, like *three-tip sagebrush communities*, *dothideomycete leaf blotch fungus, Asian horse-chestnut scale insect, giant sequoia seedlings.* These kinds of instances should always be annotated as one, MWE span.

Examples:

- The [sunflower]$_{BIOTIC\_ENTITY}$ [grew]$_{EVENTUALITY}$.

[1]We do not annotate the actions and processes of scientists in these texts, which is why “drying, freezing, and fumigation” are not annotated as eventualities. We do, however, annotate eventualities related to humans when they occur in the text in a biological context (.i.e. People living in the U.S. have an average life expectancy of 75 years.)

- [White-tailed deer]$_{BIOTIC\_ENTITY}$ [eat]$_{EVENTUALITY}$ [shoots]$_{BIOTIC\_ENTITY}$, [leaves]$_{BIOTIC\_ENTITY}$, [cacti]$_{BIOTIC\_ENTITY}$, and [grasses]$_{BIOTIC\_ENTITY}$.

- [Crabs]$_{BIOTIC\_ENTITY}$ are generally covered with a [thick]$_{QUALITY}$ [exoskeleton]$_{BIOTIC\_ENTITY}$.

- Elton organized [species]$_{BIOTIC\_ENTITY}$ into [functional groups]$_{BIOTIC\_ENTITY}$, which was the basis for Raymond Lindeman's classic and landmark paper in 1942 on [trophic dynamics]$_{EVENTUALITY}$.

- These [microbes]$_{BIOTIC\_ENTITY}$ include [viruses]$_{BIOTIC\_ENTITY}$, [bacteria]$_{BIOTIC\_ENTITY}$, [algae]$_{BIOTIC\_ENTITY}$, [heterotrophic]$_{QUALITY}$ [protists]$_{BIOTIC\_ENTITY}$ (such as [ciliates]$_{BIOTIC\_ENTITY}$ and [flagellates]$_{BIOTIC\_ENTITY}$).

- Certain [predators]$_{BIOTIC\_ENTITY}$ or [parasites]$_{BIOTIC\_ENTITY}$, when added to the [soil]$_{ABA\_ENTITY}$, can have a large effect on [root herbivores]$_{BIOTIC\_ENTITY}$ and thereby indirectly affect [plant]$_{BIOTIC\_ENTITY}$ [fitness]$_{QUALITY}$.

- There is a considerable amount of [variability]$_{EVENTUALITY}$ in [coyote]$_{BIOTIC\_ENTITY}$ [social organizations]$_{EVENTUALITY}$. In many [areas]$_{ABA\_ENTITY}$, most [coyotes]$_{BIOTIC\_ENTITY}$ are [solitary]$_{QUALITY}$ [outside of the breeding season]$_{TIME}$.

Abiotic_Entity

Refers to non-living chemical and physical parts of the environment.

Examples:

- There was plenty of [nitrogen]$_{ABIOTIC\_ENTITY}$ in the [soil]$_{ABA\_ENTITY}$.

- [Plants]$_{BIOTIC\_ENTITY}$ use their [roots]$_{BIOTIC\_ENTITY}$ to [take]$_{EVENTUALITY}$ in [water]$_{ABIOTIC\_ENTITY}$ from the [soil]$_{ABA\_ENTITY}$.

- [Carbon]$_{ABIOTIC\_ENTITY}$ [produced]$_{EVENTUALITY}$ during [photosynthesis]$_{EVENTUALITY}$ is not [used]$_{EVENTUALITY}$ for the [synthesis]$_{EVENTUALITY}$ of [proteins]$_{BIOTIC\_ENTITY}$ (and subsequent [cell growth]$_{EVENTUALITY}$), but is limited due of a lack of the [nutrients]$_{ABA\_ENTITY}$ necessary for [macromolecules]$_{BIOTIC\_ENTITY}$.

- It is now widely recognized that [bacteria]$_{BIOTIC\_ENTITY}$ and [fungi]$_{BIOTIC\_ENTITY}$ are critical to the [decomposition]$_{EVENTUALITY}$ of [carbon]$_{ABIOTIC\_ENTITY}$ and [nitrogen]$_{ABIOTIC\_ENTITY}$.

Aggregate_Biotic_Abiotic_Entity

Refers to entities that are made up of both biotic and abiotic entities. Typically, these entities will appear as biology terms that encapsulate concepts of living and nonliving things like *food web* as well as mentions of types of environments or ecosystems. For example, the *open ocean* will include biotic_entities like fish and plankton as well as abiotic_entities such as water so it should be considered an aggregate_biotic_abiotic_entity (ABA_ENTITY).

Note: These entities are frequently mentions of habitats, environments, or ecosystems. Habitats can be very broad (see the use of “rabbit warrens” and “cultivated fire breaks” below).

Examples:

- [African elephants]$_{\text{BIOTIC\_ENTITY}}$ [live]$_{\text{EVENTUALITY}}$ on the [savannas]$_{\text{ABA\_ENTITY}}$ of [Africa]$_{\text{LOCATION}}$.-
- The [microbial food web]$_{\text{ABA\_ENTITY}}$ refers the combined [trophic interactions]$_{\text{EVENTUALITY}}$ among [microbes]$_{\text{BIOTIC\_ENTITY}}$ in [aquatic environments]$_{\text{ABA\_ENTITY}}$.
- A [photic zone]$_{\text{ABA\_ENTITY}}$ is a [region]$_{\text{ABA\_ENTITY}}$ of the [ocean]$_{\text{ABA\_ENTITY}}$ through which [light]$_{\text{ABIOTIC\_ENTITY}}$ [penetrates]$_{\text{EVENTUALITY}}$; and the [place]$_{\text{ABA\_ENTITY}}$ where [photosynthetic]$_{\text{QUALITY}}$ [marine]$_{\text{QUALITY}}$ [organisms]$_{\text{BIOTIC\_ENTITY}}$ [live]$_{\text{EVENTUALITY}}$.
- An [ecosystem]$_{\text{ABA\_ENTITY}}$ includes all the [organisms]$_{\text{BIOTIC\_ENTITY}}$ in a particular [region]$_{\text{ABA\_ENTITY}}$ and the [environment]$_{\text{ABA\_ENTITY}}$ in which they [live]$_{\text{EVENTUALITY}}$.
- The [seedlings]$_{\text{BIOTIC\_ENTITY}}$ [prefer]$_{\text{EVENTUALITY}}$ [disturbed]$_{\text{EVENTUALITY}}$ [soils]$_{\text{ABA\_ENTITY}}$ which [provide]$_{\text{EVENTUALITY}}$ suitable [bare areas]$_{\text{ABA\_ENTITY}}$ for [litter-free]$_{\text{QUALITY}}$ [germination]$_{\text{EVENTUALITY}}$. Therefore, [sheep camps]$_{\text{ABA\_ENTITY}}$, [rabbit warrens]$_{\text{ABA\_ENTITY}}$, [cultivated fire breaks]$_{\text{ABA\_ENTITY}}$, [roadsides]$_{\text{ABA\_ENTITY}}$, [overgrazed pastures]$_{\text{ABA\_ENTITY}}$ and the like are ideal [propagation sites]$_{\text{ABA\_ENTITY}}$.
- In [aquatic environments]$_{\text{ABA\_ENTITY}}$, [microbes]$_{\text{BIOTIC\_ENTITY}}$ constitute the [base]$_{\text{ABA\_ENTITY}}$ of the [food web]$_{\text{ABA\_ENTITY}}$.

  Note: The “base” of the food web is referring to all the minerals, plants and animals that form this part of the food web qualifying it for the ABA_ENTITY tag.

Time

Refers to any expression of a point or period when an action, process, or state that has occurred, is occurring or is anticipated to occur in the future. Times can be captured anywhere they occur in a document, and the whole time phrase should be annotated in one span.

Time phrases can vary from being the exact date and time of an occurrence to durations such as “breeding season,” “growing season,” ”during this period,” “2 million years,” and “between summer and winter.”

In order to capture the full time phrase, we will always annotate any article preceding the time phrase and include full durations such as [over a period of weeks].

Examples:

- [Grasslands]$_{ABA\_ENTITY}$ [occur]$_{EVENTUALITY}$ in the [dry]$_{QUALITY}$ [temperate]$_{QUALITY}$ [interiors]$_{ABA\_ENTITY}$ of [continents]$_{ABA\_ENTITY}$, and first appeared [in the Miocene]$_{TIME}$.

- [The Paleogene]$_{TIME}$ is notable for being [the time]$_{TIME}$ during which [mammals]$_{BIOTIC\_ENTITY}$ [diversified]$_{EVENTUALITY}$ from relatively [small]$_{QUALITY}$, [simple]$_{QUALITY}$ [forms]$_{BIOTIC\_ENTITY}$ into a [large]$_{QUALITY}$ [group]$_{BIOTIC\_ENTITY}$ of [animals]$_{BIOTIC\_ENTITY}$ in the wake of the [Cretaceous]$_{TIME}$–[Paleogene]$_{TIME}$ [extinction]$_{EVENTUALITY}$ [event]$_{EVENTUALITY}$ that [ended]$_{EVENTUALITY}$ [the preceding Cretaceous Period]$_{TIME}$.

- [Pupation]$_{EVENTUALITY}$ for [monarch butterflies]$_{BIOTIC\_ENTITY}$ is [two weeks]$_{TIME}$ although this [process]$_{EVENTUALITY}$ may last [months]$_{TIME}$ or even [years]$_{TIME}$ for other [insects]$_{BIOTIC\_ENTITY}$.

- Some [species]$_{BIOTIC\_ENTITY}$, such as [seed predators]$_{BIOTIC\_ENTITY}$, [seedling-eaters]$_{BIOTIC\_ENTITY}$, and [tree-killing]$_{EVENTUALITY}$ [bark beetles]$_{BIOTIC\_ENTITY}$, are [true plant predators]$_{BIOTIC\_ENTITY}$, but most [herbivores]$_{BIOTIC\_ENTITY}$ [function]$_{EVENTUALITY}$ as [plant parasites]$_{BIOTIC\_ENTITY}$ because they normally do not [kill]$_{EVENTUALITY}$ their [hosts]$_{BIOTIC\_ENTITY}$, but instead [feed]$_{EVENTUALITY}$ on the [living]$_{EVENTUALITY}$ [plant]$_{BIOTIC\_ENTITY}$ without [causing]$_{EVENTUALITY}$ [death]$_{EVENTUALITY}$ (Price 1980)[2].

- [Caribou]$_{BIOTIC\_ENTITY}$ [reach]$_{EVENTUALITY}$ [sexual maturity]$_{EVENTUALITY}$ at [16 to 17 months]$_{TIME}$ of [age]$_{QUALITY}$, but [yearlings]$_{BIOTIC\_ENTITY}$ rarely [mate]$_{EVENTUALITY}$. [Females]$_{BIOTIC\_ENTITY}$ [begin]$_{EVENTUALITY}$ [mating]$_{EVENTUALITY}$ at [28 to 41 months]$_{TIME}$ of [age]$_{QUALITY}$. Most [males]$_{BIOTIC\_ENTITY}$ do not [breed]$_{EVENTUALITY}$ successfully until they are [4 to 5 years old]$_{QUALITY}$.

[2]We do not annotate time expressions when they occur in citations or footnotes such as this one.

- [Diurnal]$_{QUALITY}$ [Activity]$_{EVENTUALITY}$: [Red squirrels]$_{BIOTIC\_ENTITY}$ are [diurnal]$_{QUALITY}$ and [active]$_{EVENTUALITY}$ [year-round]$_{TIME}$. [Peak activity]$_{EVENTUALITY}$ [occurs]$_{EVENTUALITY}$ [2 hours after sunrise]$_{TIME}$ and [just before sunset]$_{TIME}$, but [red squirrels]$_{BIOTIC\_ENTITY}$ are [active]$_{EVENTUALITY}$ [most of the daylight hours]$_{TIME}$.

Location

Refers to places as defined by political boundaries (proper names of cities, states, provinces, countries, islands, continents, etc.) or geographic features that have proper names.

Test: If you can pinpoint the place mentioned on a map or globe then it should be annotated as a location (i.e. Rocky Mountains). If the place mentioned is generic like grasslands or mountains then it should be annotated as aggregate_biotic_abiotic_entity.

***Note**:* Do not annotate locations that are part of an organization's name (ex: University of Nevada).

Examples:

- [Herds]$_{BIOTIC\_ENTITY}$ of [elk]$_{BIOTIC\_ENTITY}$ [live]$_{EVENTUALITY}$ in the [Rocky Mountains]$_{LOCATION}$.

- A rain shadow is the [dry]$_{QUALITY}$ [region]$_{ABA\_ENTITY}$ on the leeward [side]$_{ABA\_ENTITY}$ of a [mountain range]$_{ABA\_ENTITY}$, where rainfall is noticeably less than on the windward [side]$_{ABA\_ENTITY}$. For example, the [White Mountains]$_{LOCATION}$ in east central [California]$_{LOCATION}$ are in the [rain shadow]$_{ABA\_ENTITY}$ of the [Sierra Nevada]$_{LOCATION}$.

- [Tundra]$_{ABA\_ENTITY}$ is a [vast]$_{QUALITY}$, mostly [flat]$_{QUALITY}$, [treeless]$_{QUALITY}$ [Arctic]$_{LOCATION}$ [region]$_{ABA\_ENTITY}$ of [Europe]$_{LOCATION}$, [Asia]$_{LOCATION}$, and [North America]$_{LOCATION}$ in which the [subsoil]$_{ABA\_ENTITY}$ is permanently [frozen]$_{EVENTUALITY}$.

Quality

Refers to a feature of an entity that can be measured or observed. See the list of qualities and their values here for more information.

Examples:

- [Carnations]$_{BIOTIC\_ENTITY}$ come in many [colors]$_{QUALITY}$.

- Charles Elton pioneered the concept of [food cycles]$_{ABA\_ENTITY}$, [food chains]$_{ABA\_ENTITY}$, and [food size]$_{QUALITY}$ in his classical 1927 book "Animal Ecology".

- The [distribution]$_{QUALITY}$ of [plant]$_{BIOTIC\_ENTITY}$ [biomass]$_{BIOTIC\_ENTITY}$.

- [Environment]$_{ABA\_ENTITY}$ includes measures like [moisture]$_{QUALITY}$ and [temperature]$_{QUALITY}$, as much as it refers to the actual [physical place]$_{ABA\_ENTITY}$[3] where an [organism]$_{BIOTIC\_ENTITY}$ is found.

- [Herds]$_{BIOTIC\_ENTITY}$ in [mountainous areas]$_{ABA\_ENTITY}$ may [move]$_{EVENTUALITY}$ from [alpine tundra]$_{ABA\_ENTITY}$ [in summer]$_{TIME}$ to [forests]$_{ABA\_ENTITY}$ at [lower elevations]$_{ABA\_ENTITY}$ [in winter]$_{TIME}$ instead of [undertaking]$_{EVENTUALITY}$ [long-distance]$_{QUALITY}$ [migrations]$_{EVENTUALITY}$.

- [Mature]$_{QUALITY}$ [flowering]$_{EVENTUALITY}$ [plants]$_{BIOTIC\_ENTITY}$ over [1]$_{VALUE}$ [meter]$_{UNIT}$ [tall]$_{QUALITY}$ had up to [24]$_{VALUE}$ [stems]$_{BIOTIC\_ENTITY}$ per [square meter]$_{UNIT}$. Based on the extremely [large]$_{QUALITY}$ [number]$_{QUALITY}$ of [immature]$_{QUALITY}$ [flowering]$_{EVENTUALITY}$ [plants]$_{BIOTIC\_ENTITY}$ (ranging from [15]$_{VALUE}$-[61]$_{VALUE}$ [cm]$_{UNIT}$ [tall]$_{QUALITY}$), it would seem that the [germination]$_{EVENTUALITY}$ [rate]$_{QUALITY}$ for [seeds]$_{BIOTIC\_ENTITY}$ must be very [high]$_{QUALITY}$.

Value

Refers any numerical measurements that coincide with descriptions of an entity's qualities.

- The [African grey parrot]$_{BIOTIC\_ENTITY}$ is a predominantly [grey]$_{QUALITY}$, [black]$_{QUALITY}$-[billed]$_{BIOTIC\_ENTITY}$ [parrot]$_{BIOTIC\_ENTITY}$ which [weighs]$_{QUALITY}$ [400]$_{VALUE}$ [g]$_{UNIT}$, with a [length]$_{QUALITY}$ of [33]$_{VALUE}$ [cm]$_{UNIT}$ and an average [wingspan]$_{QUALITY}$ of [46]$_{VALUE}$–[52]$_{VALUE}$ [cm]$_{UNIT}$.

- The [polar bear]$_{BIOTIC\_ENTITY}$ is the largest [living]$_{EVENTUALITY}$ [land]$_{AREA}$ [carnivore]$_{BIOTIC\_ENTITY}$ in the world today, with [adult]$_{QUALITY}$ [males]$_{BIOTIC\_ENTITY}$ [growing]$_{EVENTUALITY}$ up to [2.6]$_{VALUE}$ [meters]$_{UNIT}$ in [length]$_{QUALITY}$.

Unit

Refers to the unit used to measure a feature of an entity. For example, an animal will have a length measured in centimeters or a weight measured in kilograms. Centimeters and kilograms are the units used to measure the length and weight of an animal respectively.

Examples:

[3] A generic mention like "place" should be annotated as an aggregate_biotic_abiotic_entity to be consistent with these generic location/environment mentions.

- The average [length]$_{QUALITY}$ of a [millipede]$_{BIOTIC\_ENTITY}$ is [20]$_{VALUE}$ [cm]$_{UNIT}$.

- [Blue whales]$_{BIOTIC\_ENTITY}$ can be up to [30]$_{VALUE}$ [m]$_{UNIT}$ [long]$_{QUALITY}$ and [weigh]$_{QUALITY}$ upwards of [200]$_{VALUE}$ [tons]$_{UNIT}$.

- [Polar bear cubs]$_{BIOTIC\_ENTITY}$ [weigh]$_{QUALITY}$ up to [0.7]$_{VALUE}$ [kilograms]$_{UNIT}$ at [birth]$_{EVENTUALITY}$.

Proper Span Selection

Relevancy

If you are not sure whether or not a mention should be annotated, ask yourself the following questions:

- Does the mention fall into one or more of the annotation categories described above?

  If the answer is no, do not annotate the mention. If the answer is yes, then ask yourself the following question:

- Does its location within its span allow you to annotate it? If it's a headword and falls into one of our annotation categories, then it's eligible for annotation.

There are some exceptions:

- Not all entities need to be headwords in order to be annotated. All time expressions should be annotated wherever they occur regardless of their syntactic location unless they are part of a citation. Similarly, eventualities and multiword expressions can be annotated wherever they occur.

Span Information

A span refers to the section of the text that is annotated. For Clear Earth annotation, we are primarily capturing **minimum spans** with a few exceptions. A minimum span is one in which only the syntactic heads are annotated. (This generally results in single-word annotations.)

The head of a phrase is the word that determines what syntactic type of phrase it is.

For example, in the sentence:

"Our neighbor [John] bought a new, green [car]."

We have "our neighbor John" in which the syntactic head of the noun phrase (NP) is the proper noun "John" and we have the NP "a new, green car" in which the head is the noun "car."

We can also think of the head of a phrase as the word that provides the essential meaning of the phrase. Take the NP “a new, green car” for example. What is essential to our understanding of what exactly John bought is the head (and object) “car.” Although the dependents or pre-modifiers “new” and “green” give us additional information, they are less essential to our interpretation of the sentence.

This is important to know because **we only annotate words that are the syntactic head of the phrase themselves with the exception of markable multiword expressions (MWE) (again this is better defined below):**

Example of a markable, syntactic head of an NP:

- [Predators]$_{\text{BIOTIC\_ENTITY}}$ may or may not [kill]$_{\text{EVENTUALITY}}$ their [prey]$_{\text{BIOTIC\_ENTITY}}$ prior to [feeding]$_{\text{EVENTUALITY}}$ on them.

  Take a look at the noun phrase “their prey” in the sentence above and notice that we have only annotated the span [prey] and not [their prey]. We have left out the possessive pronoun [their] because it is not the syntactic head of the phrase and is not part of a MWE.

Although this minimum span approach generally results in single-word annotations, this annotation task calls for some exceptions described in the text below.

Exceptions to the rule of grabbing the syntactic head only:

There are certain multiword expressions (MWE) that we treat as though they are minimum spans and annotate the complete MWE. These can be annotated regardless of where they occur syntactically.

Note: When in doubt, check the glossary and EBV+ of biology terms. If the phrase is in the glossary then annotate the whole phrase as a single unit. If there is a phrase that is not in the glossary but you think it might be a MWE then ask the project team.

Examples:

- The [microbial food web]$_{\text{ABA\_ENTITY}}$ refers the combined [trophic interactions]$_{\text{EVENTUALITY}}$ among [microbes]$_{\text{BIOTIC\_ENTITY}}$ in [aquatic environments]$_{\text{ABA\_ENTITY}}$.

We always want to capture the full quantitative measurements of distance, weight, size or amounts.

Examples:

- [Blue whales]$_{\text{BIOTIC\_ENTITY}}$ can be up to [30]$_{\text{VALUE}}$ [m]$_{\text{UNIT}}$ [long]$_{\text{QUALITY}}$ and [weigh]$_{\text{QUALITY}}$ upwards of [200]$_{\text{VALUE}}$ [tons]$_{\text{UNIT}}$.

We want to capture the full measurement span which is both the value and the unit. In the sentence above we have two measurements [30 m] (value = 30 and unit = m) and [200 tons] (value = 200 and unit = tons).

Only annotate modifiers if the headword warrants an annotation category listed above.

Examples:

- A gradient exists between [trophic levels]$_{BIOTIC\_ENTITY}$ running from complete [autotrophs]$_{BIOTIC\_ENTITY}$ that [obtains]$_{EVENTUALITY}$ their sole source of [carbon]$_{ABIOTIC\_ENTITY}$ from the [atmosphere]$_{ABIOTIC\_ENTITY}$ to [mixotrophs]$_{BIOTIC\_ENTITY}$ such as [carnivorous]$_{QUALITY}$ [plants]$_{BIOTIC\_ENTITY}$ that are [autotrophic]$_{QUALITY}$ [organisms]$_{BIOTIC\_ENTITY}$ that partially [obtain]$_{EVENTUALITY}$ [organic matter]$_{BIOTIC\_ENTITY}$ from sources other than the [atmosphere]$_{ABIOTIC\_ENTITY}$, and complete [heterotrophs]$_{BIOTIC\_ENTITY}$ that must [feed]$_{EVENTUALITY}$ to [obtain]$_{EVENTUALITY}$ [organic matter]$_{BIOTIC\_ENTITY}$

- The [ant]$_{BIOTIC\_ENTITY}$ [Novomessor cockerelli]$_{BIOTIC\_ENTITY}$ [interferes]$_{EVENTUALITY}$ with the ability of the [red harvester ants]$_{BIOTIC\_ENTITY}$ to [forage]$_{EVENTUALITY}$ by [plugging]$_{EVENTUALITY}$ the entrances to their [colonies]$_{ABA\_ENTITY}$ with [small]$_{QUALITY}$ [rocks]$_{ABIOTIC\_ENTITY}$.

The highlighted sections above illustrate instances in which we can annotate modifiers (*carnivorous, autotrophic,* and *small*) when their headwords (*plants, organism*, and *rocks*, respectively) correspond to the annotation categories biotic_entity or abiotic_entity and are annotated appropriately. If those terms did not fit into one of the annotation categories in our scheme they would not be annotated, which would make their modifiers ineligible for annotation as well. The only exceptions to this rule are multiword-expressions, eventualities, and time expressions, all of which can be annotated wherever they occur regardless of their syntactic location.

Tricky Cases

1. How to annotate an entity that is part of a discontinuous span:

Discontinuous spans add an extra step to entity annotation. A discontinuous span is a span that is interrupted in some way (see example immediately below). Often, the entity we'd like to tag is discontinuous because of a conjunction. An annotator can capture the complete span by linking (or adding) the disparate parts of the span together.

- [Morphology]$_{QUALITY}$: The [form]$_{QUALITY}$ and [structure]$_{QUALITY}$ of anything, usually applied to the [shapes]$_{QUALITY}$, [parts]$_{BIOTIC\_ENTITY}$, and [arrangement]$_{QUALITY}$ of [features]$_{QUALITY}$ in [living]$_{BIOTIC\_ENTITY}$ and [fossil organisms]$_{BIOTIC\_ENTITY}$.
  The discontinuous span in the example above is [living...organisms].

- The [top]$_{QUALITY}$ [trophic level]$_{BIOTIC\_ENTITY}$ of the [food web]$_{ABA\_ENTITY}$ has [top] or [apex predators]$_{BIOTIC\_ENTITY}$ which no other [species]$_{BIOTIC\_ENTITY}$ [kills]$_{EVENTUALITY}$ directly for its [food resource]$_{BIOTIC\_ENTITY}$ [needs]$_{EVENTUALITY}$.
  This example has one discontinuous span: [top...predators]. [apex predators] is an multiword expression in our glossary so we annotate it as a single span.

2. How to annotate possessive adjectives:

Occasionally, a possessive adjective will modify a relevant headword. If the possessive adjective corresponds to one of the annotation categories below then it is fine to annotate it, but the " **'s** " should not be included in the span.

Example:

The [human]$_{BIOTIC\_ENTITY}$ [retina]$_{BIOTIC\_ENTITY}$ has about [10]$_{VALUE}$ times more [cones]$_{BIOTIC\_ENTITY}$, the [light receptors]$_{BIOTIC\_ENTITY}$ that [function]$_{EVENTUALITY}$ best in bright [light]$_{ABIOTIC\_ENTITY}$, than [**cats]**$_{BIOTIC\_ENTITY}$' [eyes]$_{BIOTIC\_ENTITY}$ have.

In the example above, we capture the biotic_entity "cats" but omit the apostrophe from the span.

3. How to annotate plural nouns:

The pluralization of an entity should be annotated the same way as it would if it were singular.

Example:

The [bats]$_{\text{BIOTIC\_ENTITY}}$ [flew]$_{\text{EVENTUALITY.}}$
The [bat]$_{\text{BIOTIC\_ENTITY}}$ [flew]$_{\text{EVENTUALITY}}$.

4. How to annotate a part of a biotic entity?:

The definition of a biotic_entity can be reasonably extended to include parts of a living thing. Mentions can be specific like leaves to plants or organs to mammals.

An exception to this rule are any mentions of elements on the periodic table. For example, nitrogen will always be abiotic whether or not it is found in the atmosphere, soil, or an organism.

Examples:

- The [leaves]$_{\text{BIOTIC\_ENTITY}}$ of [sunflowers]$_{\text{BIOTIC\_ENTITY}}$ are [bright green]$_{\text{QUALITY}}$ in [color]$_{\text{QUALITY}}$, [heart-shaped]$_{\text{QUALITY}}$ and covered with [fine]$_{\text{QUALITY}}$ [hairs]$_{\text{BIOTIC\_ENTITY}}$.

5. It is important to consider the context when deciding on an annotation category.:

Take the term *biomass*. In certain contexts, *biomass* should be considered a quality while, in other contexts, *biomass* should be considered a biotic_entity.

Examples:

- [Microbial]$_{\text{QUALITY}}$ [biomass]$_{\text{QUALITY}}$ is a measure of the [mass]$_{\text{QUALITY}}$ of the living components (i.e. [bacteria]$_{\text{BIOTIC\_ENTITY}}$ and [fungi]$_{\text{BIOTIC\_ENTITY}}$) of [organic matter]$_{\text{BIOTIC\_ENTITY}}$ in [soil]$_{\text{ABA\_ENTITY}}$.

- The [microbial loop]$_{\text{ABA\_ENTITY}}$ describes a pathway in the [microbial food web]$_{\text{ABA\_ENTITY}}$ where [DOC]$_{\text{BIOTIC\_ENTITY}}$ is returned to [trophic levels]$_{\text{BIOTIC\_ENTITY}}$ via the [incorporation]$_{\text{EVENTUALITY}}$ into [bacterial]$_{\text{QUALITY}}$ [biomass]$_{\text{BIOTIC\_ENTITY}}$.

  Biomass in the first example should be annotated as quality because it is referring to a particular feature or quality of soil that can be measured. Whereas, *biomass* in the second example is referring to part of the biotic_entity *bacteria*, which qualifies it for the annotation category biotic_entity as well.

6. It is important to consider the part-of-speech of an instance when deciding on an annotation category.:

Why are adjectives like “planktonic” and “autotrophic” annotated as qualities when they are modifiers, but annotated as biotic_entities when they occur as the nouns “plankton” and “autotroph” in the text?

It is clear that “plankton” and “autotroph” are referring to living organisms when they occur as nouns. These words represent the actual, physical entities in the world. However, when these words occur as adjectives, it is not so much the entities themselves that are being referred to, but rather the qualities they represent. This is why we choose different annotation categories depending on the terms part-of-speech.

7. **When should mentions about individuals be annotated and when should they be ignored?:**

We do not annotate the proper names of people (authors, scientists, scholars) in this project. However, we do annotate references to humans and humanity when they occur in a biological context.

An example of when not to annotate:

- Elton organized [species]$_{\text{BIOTIC\_ENTITY}}$ into [functional groups]$_{\text{BIOTIC\_ENTITY}}$, which was the basis for Raymond Lindeman's classic and landmark paper in 1942 on [trophic dynamics]$_{\text{EVENTUALITY}}$.

An example of when to annotate:

- The [ingenuity]$_{\text{QUALITY}}$ and [adaptability]$_{\text{QUALITY}}$ of [*Homo sapiens]*$_{\text{BIOTIC\_ENTITY}}$ has led to its becoming the most influential [species]$_{\text{BIOTIC\_ENTITY}}$ on the [Earth]$_{\text{LOCATION}}$.

8. **Food webs and food chains are often followed by sublevels and sub-categorizations like “levels” or “top two levels” in the example below:**

Examples:

- Many of them also share the same fate—which is to become food for the carnivorous animals of the food chain's **top two levels**.

Mentions of sublevels and sub-categorizations should not be annotated. These are non-standardized references in biology and therefore are too generic to annotate in this schema.

Note: The same rule applies for modifiers like “primary, secondary, tertiary, etc” in noun phrases like “the primary marine food web.”

9. How to annotate the modifiers -- generic and specific -- of environments:

Authors of biology texts use modifiers to offer more precise descriptions of environments. These can be standard multiword expressions within the field of biology such as the earth’s biomes like tropical forest, temperate grasslands, and coral reefs. If you are not sure if it is a recognized MWE then check the biology glossary. If you find it there then it is okay to annotate.

However, more often than not, modifiers are uniquely combined with headwords and are not standard biology terms. In these cases, you will likely not find them listed in the glossary, but you should still treat environments and ecosystems (aggregate_biotic_abiotic_entities) as one span and then add a note about the phrase in the “potential MWE” tab of the glossary.

Note: If a “quality” is found modifying an environment or ecosystem mention then default to the rule above and mark it as one, aggregate_biotic_abiotic_entity rather than dividing the span into two annotations: one quality and the other aggregate_biotic_abiotic_entity.

Example:

- There are other [deep-ocean ecosystems]$_{\text{ABA\_ENTITY}}$ that are entirely independent of the [sunlight energy]$_{\text{ABIOTIC\_ENTITY}}$ that kick-starts the main [marine ecosystem]$_{\text{ABA\_ENTITY}}$.