**Annotation Protocol**

This protocol outlines the process of establishing concepts or entities to annotate a core set of papers in the field of restoration ecology and train natural language processing (NLP) tools for information extraction. This document serves as both an essential component in this thesis as well as a guideline for further annotations as part of a larger project aiming at developing an AI-based tool to synthesize academic data into a more readable and easier format to handle a set of hypotheses, relational variables and ultimately an interactive tool for practitioner.

Annotation Task: When provided with a scientific article on restoration ecology, the goal is to mark certain words and phrases in the text corresponding to pre-established entities, with the goal of creating a knowledge graph.

Document Organization: This protocol is divided into four parts. The first establishes the codes and provides them with detailed definitions. The second part provides examples of entities in the texts that are normally excluded from the annotations and other terms to watch out for. The third part details the process behind obtaining scientific articles, formatting them into plaintext files, and organizing them to be processed by the software used, INCEpTION.

These instructions should provide users with an understanding of the nature of the annotated sections of the text, the rationale behind it, and what to avoid annotating.

1. **Established Entities**

There are eight classifications we are working with.

*Table 1*. Information based on Extraction Scheme v.4 created by **Team?**

|  |  |  |
| --- | --- | --- |
| **Entity** | **Description** | **Examples** |
| Time period of study | Refers to information about the timing of the study, beginning and end date, total duration, timing, and duration of fieldwork.  Usually found in the Abstract/Introduction/Methods sections. | December 2024 |
| Location of study | Refers to information about the physical setting, coordinates, geographical location, and name of country/city.  Usually found in the Abstract/Introduction/Methods sections. | New York, plains of America |
| Ecosystem | Refers to information that describes the ecosystem of the study or where the species is present, which includes the type of ecosystem, land use history (any previous degradation events or usage patterns), abiotic conditions, and occasionally the location of the study if it contains any descriptor.  Usually found in the Methods sections. | Temperature, terrain, weather, precipitation, nutrients, plains of America |
| Focal object | Refers to the species discussed | Tulipa gesneriana |
| Method | Refers to the type of the study and includes type of study (experiment, fieldwork, surveys), its conditions, number and types of samples, measured variables. | Collecting 50 samples from a designated area, performing regression analysis, using GIS |
| Research questions | Refers to the problems or questions and sometimes encompasses a clearly stated research gap.  Usually found in the last part of the Introduction/Method sections. | “We performed this study to determine whether X impacts Y” |
| Main hypothesis and corresponding results | Refers to the hypotheses and theories posed by the authors in the beginning of the paper, and the main outcomes and results of the study.  Usually found in the Results/Discussion sections. | “Habitat X had the highest amount of species mortality”.  “Our study showed significant variance”. |
| Causal statements | Refers to statements that clearly depict a cause-effect relationship.  Does not have a designated location and not always found in every text. | “The absence of species X resulted in the growth in the number of species Y”. |
| Recommendations and suggestions | Refers to statements not backed by hard evidence or results of the study, rather the author’s own knowledge or conclusions.  Can also include suggestions to alter the methods of the study to achieve desired results.  Found in the Discussion sections. |  |

1. **Annotating Specific Entities**
   1. Annotating Entities as *Time period of study*

The time period of the study is mostly explicitly stated at the beginning of the methods section, usually in the form of a month and a year, and is annotated as such, e.g., December 2024. Other instances include time periods (i.e., month to month), duration of certain steps in the study (i.e., analysis taking months or years), length of study, and start and end dates. We also annotate descriptors of dates (e.g., *early* December, *mid-*October). There usually is no repetition of the time period in the rest of the text.

Excluded are experiment timestamps (i.e,. samples being observed on day X of study), and how often steps are performed (i.e., daily, monthly, yearly).

Examples of sentences containing annotated entities, marked in italics:

* “*From 14 to 15 August 2019* we transplanted 21 individuals of each species into each of our herbicide treated areas and untreated reference areas”.
* “In *early September 2019* we recorded survival and measured growth of all transplants”.
* “In *June and August 2020*, *June 2021*, and again in *August 2022*, we repeated our assessments of survival, growth, reproduction, and browse on our sentinel plants, and percent cover of associated vegetation”.
  1. Annotating Entities as *Location of study*

The location of the study is where it is carried out and is expressed in the names of cities, countries, parks, or nature reserves. It is explicitly stated in the abstract, introduction, and methods sections. It also includes coordinates, its proximity or location with respect to other significant markers, and which part of the country or island it is located on, as well as crucial ecosystem characters (e.g. wetlands, grasslands, rainforests).

As the location is usually mentioned throughout the text, we only annotate it in the introduction and methods sections, and in the results section only when it is included in a sentence that contains another annotated entity (i.e. in the sentence “our study showed a higher rate of species X in location A”, we annotate species X as a focal object, location A as location of study, and the entire sentence as main hypothesis and corresponding results). We exclude phrases that contain background information on the location such as construction dates in the case of parks or educational institutions where some of the studies were carried out.

Examples:

* “We used the ongoing long-term P. australis management program in Adirondack Park in northern New York State, United States…”
* “The Adirondack Park is one of the largest protected areas in the Eastern United States…”
* “The first research site was located on the Crane Creek Forest, hereafter “Crane Creek,” a private research forest, in Latah County, Idaho (lat 46.9875 N, long 116.88 W).”
* “We conducted our experiment at three study sites in northern Idaho, United States.”
* “Site 2, hereafter Ukwle, was located on Ukwle Mountain in Benewah County, Idaho (lat 47.1919 N, long 116.8892 W)”.
  1. Annotating Entities as *Ecosystem*

The ecosystem comprises every attribute of the physical environment from abiotic conditions to land use history. It is normally found in the methods section when describing the location of the study. Similar to the time period entity, it is not commonly discussed in the rest of the text.

Examples:

* “*Low levels of anthropogenic disturbance*, *limited road development*, and *conservation* status have led to *minimal encroachment by invasive species*, compared to the surrounding regions.”
* “*Chemical treatments of select populations using glyphosate began in 2010*, and the Adirondack Park Invasive Plant Program (APIPP) conducted strategic treatments aiming for *suppression and long-term eradication of these mostly small population*.”
* “We selected sites with *unsuitable conditions for natural huckleberry establishment*—*no canopy cover, hot and dry soils, and below 900 m.a.s.l*.”
* “Our chosen field sites were consistent in their *low elevations* and *completely open overstories*—habitat qualities that made them *unsuitable for natural black huckleberry establishment*.”
  1. Annotating Entities as *Focal object*

Focal object represents the species at the center of the study and is usually its scientific name. Similar to the location of study, it is significantly repeated throughout the text. Therefore, we annotate every instance the species is stated from the abstract to the results sections but exclude instances where the species is stated in sentences that do not have any other annotations, and are thus irrelevant to our study, or contain background information on the species itself. There are also instances when the general name of the species is mentioned, such as “birds”, that do not get annotated.

It is important to note that when transferring the plain text files to the INCEpTION software, there is often a symbol of question mark in a box on both sides of the word. In cases where we have to annotate a phrase including a species, these symbols are included.

Examples:

* “Introduced *Phragmites australis* represents a widespread threat to North American wetlands.”
* “Treatment duration (short- or long-term) did not significantly influence survival of A*. incarnata* or *S. alba*.”
* “Finally, we chose to study the microsite effect on *black huckleberry shrubs* due to the cultural, economic, and ecological importance of this species…”

Examples of instances where we do not annotate the species:

* “Prescribed burn treatment was not important in determining survival and fluorescence of black huckleberry seedlings, nor in determining SST and VWC trends.”
* “Black huckleberries are flowering and fruiting earlier in the year, leading to phenological mismatch between huckleberries and important pollinators…”
  1. Annotating Entities as *Methods*

The methods of the study are the most detailed and are found beginning in the end of the abstract and introduction, make up the entire methods section, and are rarely found in the title of the study or scattered in the discussion section when the authors state their adherence to or avoidance of certain steps in the study. They are usually annotated as sentence phrases.

Examples:

* *Legacy effects of Phragmites australis and herbicide treatments* can reduce survival but do not prevent native plant establishment.
* “We used a *coordinated management program* targeting mostly small P. australis populations in the Adirondack Park in New York State, United States, to assess the potential for such *legacy effects*.”
* “We *used a drip torch as an ignition device*.”
* “When an *estimate of the rate of selfing for this population was carried out after excluding these three individuals*, the rate was estimated to be zero (s = 0)” (as found in the Results section).
  1. Annotating Entities as *Research questions*

Research questions are usually found at the end of the introduction. They can either be straightforward questions that are answered in the study, can take the form of a clearly written research gap also at the end of the introduction, or stated as the aim of this study. They aren’t always found in every article we annotate.

Examples:

* “The poor outcomes of allowing unassisted succession to proceed (Pearson et al. 2016) raise the question whether *reseeding or replanting could prevent reinvasion or result in more diverse plant communities*.”
* “This research provides a *comparison of genetic diversity and inbreeding between restored and wild populations for three subtropical rainforest tree species in Northeast New South Wales (NE NSW), Australia*.”
* A secondary question was, “*Are trees in restored populations more inbred than those in the wild reference populations?*”
* “However, *evaluation of the success of reef restoration in reestablishing structural complexity and at what scales this is achieved is lacking in the literature*.”
  1. Annotating Entities as *Main hypothesis and corresponding results*

This entity is two-fold. If even found in the study, the hypothesis is usually in the introduction section along with a brief overview of the results. Mostly, the results section and some of the discussion form the bulk of this entity. It includes statements that involve the discussed study and results they have reached, and excludes background information, information on related studies, reasoning behind some of the results, and suggestions or recommendations.

Examples:

* “We found only *limited support for our hypothesis that herbicide treatment and years of treatment would negatively impact seedling survival*.”
* “We observed a *limited incidence of herbivore browsing during the study*, with *20 individuals being browsed a total of 22 times across all sites and species*.”

Examples of instances where we do not annotate:

* “The spatial and directional pattern of survival can be explained by the timing of amelioration of extreme heat throughout the growing season.”
* “Many shrub and tree seedlings, including black huckleberry, experience high mortality during initial establishment and during the seedling life stage.”
  1. Annotating Entities as *Causal statements*

Causal statements are not commonly found in the examined studies but are easily located due to their easy sentence structure represented as ‘[cause] results in/causes/leads to [effect]’. They are normally found in the results and discussion sections.

Examples:

* “*As the size of treatment areas increases, costs for treatment and replanting will also increase*,…”
* “We consider it likely that this *enhancement in plant growth was a function of reduced plant competition in treatment areas*…”
* “Our results indicate the *microsite benefit weakens as distance from DWD increases*…”
  1. Annotating Entities as *Recommendations and suggestions*

Recommendations and suggestions are most commonly found in the discussion section and represent the authors’ opinions on study conditions that may be altered to achieve more desirable results or simply their opinion on the obtained evidence. It excludes any information that was referenced from other authors (such as background information).

Examples:

* “Although the removal of invasive plant species is an essential step in riparian forest restoration, we *may improve outcomes by considering effects on mycorrhizal fungi*.”
* “In particular, we *advocate for the use of a local reference inoculum instead of a commercial inoculum*, since our results failed to detect the presence of any EMF associated with the commercial inoculum.”\*

\* The second part of this sentence is actually annotated as *Results*.

Examples of instances where we do not annotate:

* “However, many riparian communities in the western United States are extensively invaded, increasing the need for high-disturbance removal (Goodwin et al. 1997; USGS 2002; Boland 2016).”

1. **Use of INCEpTION**

After accessing the website of the paper from the PDF located in the ‘Restoration Ecology’ shared folder, we copy the entire text and paste into a Word file. The file is then saved as a plain text file titled ‘(N) AuthorNameYear’ e.g. ‘(30) Smith2024’.

Images and figures are lost in translation and thus erased from the txt file except for headings, numerical values, and titles, all of which we disregard in our annotations.

We then ‘clean’ and organize the txt files so that it is readable on INCEpTION and to avoid many unknown symbols that interfere with the annotations. Steps in the procedure include:

* Deleting the ‘Open in PowerPoint’ phrase under many of the figures
* Adding full stops to the end of each heading to obtain sentences on different lines
* Deleting points after authors’ middle initials to avoid excessive line breaks

Other steps in this procedure involve using the ‘Find and replace’ option to get rid of symbols such as:

* Replacing “?” with a space
* Replacing “°” with “ degrees ”
* Replacing “±” with “+/-“ and “ ≤” with “</=”
* Replacing “-“ or “—” with “\_”
* Replacing quotation marks “ ” with spaces

The txt file is then uploaded to INCEpTION after manually creating the layers that correspond to our entities. The paper is displayed on the tool such that each sentence is on one line. This facilitates the annotation process and allows the model to extract precise and concise statements.

1. **Concerns between Annotators**

* ~~Results: we have been annotating results directly correlated to the study and the hypothesis (including sampling/experiment results, conclusions) but excluding the authors’ personal opinions and recommendations~~  **🡪 addressed by adding new category**
  + Some annotated phrases include citations from different papers (check Deschenes) 🡪 **not included**
  + ~~On the other hand, the sentences which include these aforementioned recommendations are divided into two: the first part states a result of the study (which we annotate) and the second contains these suggestions 🡪 do we annotate the entire sentence?~~
  + ~~Should there be a separate layer?~~
* ~~The issue of repetition: focal points/location of study are mentioned throughout the text but we’ve annotated them in the abstract/introduction/methods where they are the subject of interest. We did not annotate in the following sections unless they are included in sentences that already include other annotations.~~ **🡪 addressed above in respective focal object/location sections**
* ~~Methods: sometimes included in discussion/conclusions~~ **🡪 we annotate**
* ~~Focal point:~~
  + ~~should we annotate both the common name and the species name (e.g. bats, birds, fruit-bearing trees)~~ **🡪 addressed by opting to not annotate the general name**