Problem Statement

1. Goal

The goal is to develop a systematic approach to identify and correct mismatches in trait ontology mappings. The process ensures that trait names are accurately linked to their respective IDs, which is crucial for consistency in biological databases and downstream applications like genomics and phenomics research.

2. Dataset overview

The dataset (trait_ontology_details.txt) contains comprehensive trait ontology details in a structured text format. It serves as the foundation for building a trait mapping and correction system. Each trait is represented as a block of text containing:

- **Trait ID:** Unique identifier for each trait
- **Name:** Official term describing the trait
- **Definition:** A textual description of the trait
- **Synonym:** The synonym of the trait
- **Hierarchy (is_a relationships):** Parent-child relationships that define the hierarchical structure of the ontology

```
[Term]
id: TO:0000004
name: reversible male sterility
def: "Trait of a plant in which the action of the cytotoxic gene used to
 introduce male sterility is suppressed by the application of a chemical
to the plant. Reversion of the sterility allows the RMS parent to be
self-fertilized, a step that overcomes the need to remove fertile sib plants
 prior to making the hybrid cross." [PMID:10645437]
synonym: "RMS (related)" RELATED []
is_a: TO:0000111 ! genetically engineered male sterility
```

Figure 1: one block of term ontology details dataset

The dataset contains hierarchical relationships and detailed trait information, including synonyms and definitions. This information provides critical insights into the quality and structure of the ontology data, guiding the development of our methods for error correction and trait mapping. The dataset statistics are summarized below, providing insights into the structure and completeness of the trait data. Understanding the dataset's structure through these statistics allows for informed decision-making in downstream tasks, such as semantic matching and ontology refinement.

Number of Trait IDs: 1667
Number of Traits without Synonyms: 759
Number of Traits without Definitions: 99
Number of Traits without Both Synonyms and Definitions: 43
Traits without 'is_a' relationships: 34
Average Synonyms per Trait: 0.98
Maximum Hierarchy Depth: 12
Average Hierarchy Depth: 5.59

3. Key Challenges

Ambiguity in Trait Names:
Different researchers or databases may use varying terms to refer to the same trait. For instance:

"Flowering time trait" vs. "Days to flowering trait"

" Plant height uniformity " vs. " Plant height "

High Volume of Data:
The dataset contains thousands of trait IDs, each associated with names, definitions, synonyms, and hierarchical relationships. Manually resolving mismatches in such a dataset is time-consuming and error-prone.

Complex Ontology Structure:
The hierarchical relationships (is_a relationships) introduce another layer of complexity. A trait might inherit characteristics from one or more parent traits, creating dependencies that need to be considered when suggesting corrections.

Methodology

1. Task 1

The goal is Developing a quantitative metric to evaluate the correctness of each row in the CSV files. To achieve this, we introduced two complementary metrics: *string similarity* and *semantic similarity*. These metrics strike a balance between strictness and flexibility, ensuring greater accuracy in matching ontology terms.

a. String similarity

In string similarity, we focus on the textual structure rather than the contextual meaning of the terms. This method evaluates both exact and partial matches of the trait name:

1. **Exact Match:**
   Exact matching calculates the similarity between two strings by comparing their entire content. It is particularly useful when the strings are expected to be highly similar in their entirety.

   - **Formula:** Exact Match Score = Levenshtein Similarity / 100
   - *The score is normalized to the range [0, 1].*

2. **Partial Match:**
   Partial matching compares substrings and identifies the highest similarity between parts of the strings. This approach is ideal when the input might be a partial or abbreviated version of the ontology term.

   - **Formula:** Partial Match Score = FuzzyWuzzy Partial Ratio / 100
   - *The score is also normalized to the range [0, 1].*

The overall String Similarity is determined as:

String similarity= max(Exact Match, Partial Match)

b. Semantic similarity

Semantic matching is a sophisticated technique that evaluates the similarity between terms by focusing on their contextual meanings rather than their surface-level textual representations. This method leverages pre-trained language models, SentenceTransformer, to generate high-dimensional vector embeddings for textual content. These embeddings capture contextual relationships and semantic nuances, enabling a deeper understanding of terms, definitions, and synonyms.

Unlike string-based methods like exact or partial matching, which struggle when terms differ significantly in syntax but share similar meanings, semantic matching excels in such scenarios. It compares the meaning of terms by converting text into embeddings and calculating their similarity, providing a more nuanced and robust evaluation.

This method computes two key contextual similarities:

- **Synonym Similarity:** Measures the semantic alignment between the trait name and its synonyms.
- **Definition Similarity:** Assesses the contextual similarity between the trait name and its definition.

The overall Semantic Similarity is determined as:
Semantic Similarity = max(Synonym Similarity, Definition Similarity).

The table 1 show the comparison between the string and semantic similarity.

Table 1: The comparison between the string and semantic similarity

Aspect	**String Similarity**		**Semantic Similarity**
	Exact Matching	**Partial Matching**	
Definition	Compares the trait ontology name to the trait name for an exact match	Matches substrings or portions of the trait name and trait ontology name	Uses embeddings to compare the contextual meaning of the trait name and trait ontology synonym and definition
Technique	Levenshtein Distance	Fuzzy Matching	Embedding-based comparison using cosine similarity.
Data Used	Only the trait name.	Only the trait name.	Trait name, definition, and synonyms.
Strengths	- Quick and straightforward. - Suitable for identical matches.	- Tolerates partial matches. - Works for abbreviations or partial overlaps.	- Captures nuanced meanings. - Handles synonyms and paraphrases effectively. - Robust to phrasing variations.
Weaknesses	- Fails for synonyms, rephrasings, or	- Sensitive to substring patterns.	- Computationally intensive. - Depends on quality and completeness of ontology data.

	variations. - Sensitive to typos.		
Example (Trait Name)	**Input**: "Plant Size" **Ontology**: "Plant Size" **Score**: 1.0 (Perfect Match)	**Input**: "Plant Size" **Ontology**: "Regulation of Plant Size" **Score**: 0.67 (Partial Overlap)	**Input**: "Plant Size" **Ontology**: Definition: "Control of height and width of a plant" **Score**: 0.85 (Contextual Match)
When to Use	- Ideal for identical matches. - Low-complexity tasks.	- Useful for loosely matching names. - Detects abbreviations and truncated terms.	- Effective for capturing synonyms, nuanced definitions, or semantic relationships.
Complexity	Low	Moderate	High

The final metric combines string similarity and semantic similarity, providing a comprehensive evaluation by considering both the structural alignment and the contextual meaning of the trait name.

Final correctness score = max(String Similarity, Semantic Similarity)

This approach ensures that both the textual form and the deeper semantic relationships of the terms are equally accounted for, achieving a balanced and accurate assessment.

2. Task 2

The goal of this task was to aggregate the correctness scores across rows in multiple CSV files and rank the CSV files based on the overall quality of their trait ontology mappings. The correctness of each row was evaluated based on how well the trait name and ID matched against a predefined ontology (using a combination of exact, partial, and semantic similarity techniques).

The total correctness score for a CSV file was computed by summing up the individual scores of all rows. The average score for the CSV file was then obtained by dividing the total score by the number of rows.

Average score=∑ (correctness score of all rows)/number of rows

After evaluating all the CSV files, they were ranked based on their average scores, as shown below. Figure 2 illustrates this ranking:

```
Ranked CSV Files:
[('response_model1.csv', 0.9012692196179282),
('response_model2.csv', 0.8741346153846151),
('response_model3.csv', 0.531364413373014)]
```

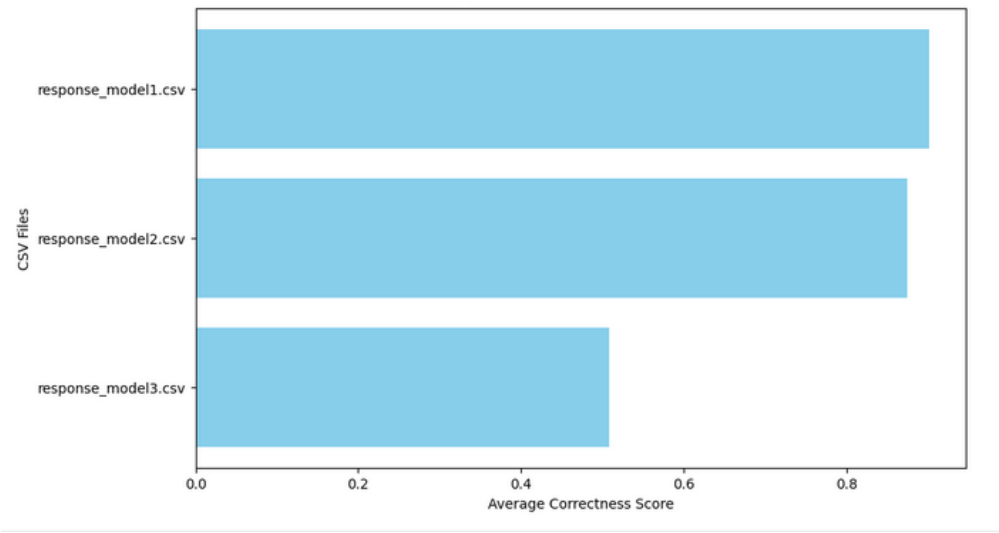

Figure 2: Ranking of CSV Files Based on Correctness

This task allowed us to assess and rank the overall quality of the mappings in the CSV files, based on both their structural accuracy (exact/partial string matching) and their contextual accuracy (semantic matching). The CSV files that achieved higher scores indicated better overall accuracy in aligning the trait names and IDs with the ontology terms, which is critical for ensuring consistency and correctness in the dataset.

3. Task 3

The primary objective of Task 3 is to enhance the mapping of trait names to their corresponding trait IDs in the dataset by systematically identifying and correcting mismatches. This process ensures a more accurate representation of traits, thereby improving the overall quality of the dataset for subsequent analyses.

A critical step in this approach involves constructing a directed graph from the trait ontology. In this graph, nodes represent trait IDs, and edges depict hierarchical relationships (e.g., *"is_a"*). This structure facilitates the computation of hierarchical similarities between traits, which is vital for understanding their relational context within the ontology. The hierarchical similarity is as follows which means that when two nodes have shorter path between them the hierarchical similarities increases.

hierarchical similarities=1/(1+Shortest path)

For traits with correctness scores below a predefined threshold (e.g., 0.7), potential corrections are generated by comparing semantic embeddings of trait definitions, and synonyms. Additionally, hierarchical similarities derived from the ontology graph are incorporated into the evaluation. The combined scores from these measures are used to rank the most suitable alternative trait IDs, with the highest-scoring ID being suggested as a correction.

Suggest score=max (Semantic similarity, hierarchical similarities)

This systematic methodology leverages both semantic similarity and ontology relationships to effectively address mismatches, resulting in a more robust and reliable dataset.

Table 2, 3 and 4 show the differences between the original id and corrected id for reponse_model1, reponse_model2, and reponse_model3.

Table 2: comparison of original and corrected id in response_model1

Trait name	Original id	Original score	New id	New score
cytokinin accumulation	0006016	0.27	0000167	0.78
auxin response	0006016	0.36	0000163	0.92
nitrogen use efficiency	0000506	0.63	000506	0.68
wheat sharp eyespot disease resistance	0000439	0.61	0000673	0.64
root system depth	0000343	0.4	0000227	0.82
grain development	0000396	0.54	0000028	0.67
root biomass	0001041	0.6	0000143	0.64
pre-harvest sprouting resistance	0000697	0.42	0000388	0.54
hybrid necrosis	0002663	0.6	0002663	0.63
cereal cyst nematode resistance	0000384	0.54	0001111	0.61

Table 3: comparison of original and corrected id in response_model2

Trait name	Original id	Original score	New id	New score
nitrogen use efficiency	0000506	0.63	0000506	0.68
abiotic stress tolerance	0000164	0.62	0000168	0.96
Thousand-kernel weight	0000442	0.63	0000919	0.80
Kernel length	0000442	0.48	0000734	0.97
Kernel width	0000442	0.52	000975	0.96
spike length	0006030	0.67	0000152	0.73
spikelet number per spike	0006030	0.64	0000456	0.83
thousand-grain weight	0000442	0.56	0000382	0.95
tiller number per plant	0000441	0.32	0000152	0.89
kernel weight	0000442	0.69	0000919	0.95
kernel thickness	0000442	0.25	0000975	0.75
1000-grain weight	0000442	0.53	0000382	0.97
heading date	0000344	0.61	0000137	0.79

Table 4: comparison of original and corrected id in response_model3

Trait name	Original id	Original score	New id	New score
plant regeneration	0006003	0.34	0000096	0.69
cytokinin accumulation	0000270	0.3	0000167	0.78
nitrogen use efficiency	0000439	0.35	0000506	0.68
abiotic stress tolerance	0000165	0.46	0000168	0.96
thousand kernel weight	0000382	0.53	0000919	0.81
spike architecture	0001110	0.56	0020077	0.71
pollen exine formation	0000727	0.5	0000218	0.64
supernumerary spikelets	0000434	0.31	0000152	0.71

thousand kernel weight	0000382	0.53	0000919	0.81
disease resistance	0000111	0.39	0000112	0.81
male sterility	0000016	0.57	0000037	0.83
salt tolerance	0006003	0.36	0006001	0.71
JA-responsiveness	0006004	0.35	0000172	0.53
resistance to Rhizoctonia cerealis	0002636	0.36	0000548	0.55
thousand kernel weight	0000382	0.53	0000919	0.81
resistance to powdery mildew	0000133	0.42	0000054	0.50
spikelet number	0000749	0.33	0000456	0.79
thousand kernel weight	0000382	0.53	0000919	0.81
pollen development	0000499	0.43	0000245	0.75
seed dormancy	0000659	0.27	0000619	0.68
powdery mildew resistance	0006003	0.35	0000054	0.48
drought stress tolerance	0006004	0.39	0000188	0.74
phosphate starvation tolerance	0006005	0.44	0020112	0.59
grain size	0002734	0.5	0002626	0.85
glume pubescence	0006003	0.31	0020035	0.76
salt tolerance	0006002	0.38	0006001	0.71
flowering time	0000134	0.36	0002616	0.91
thousand kernel weight	0000382	0.53	0000919	0.81
grain development	0006004	0.53	0000028	0.67
resistance to Pratylenchus neglectus	0000260	0.38	0006067	0.51
Fusarium head blight susceptibility	0006053	0.33	0000673	0.68
meiotic crossover frequency	0000728	0.51	0000729	0.55
seed germination	0000391	0.67	0000430	0.85
seed germination	0000391	0.67	0000430	0.85
root biomass	0000043	0.5	0000143	0.64
specific root length	0000229	0.4	0000227	0.84
powdery mildew resistance	0006002	0.35	0000054	0.48
leaf senescence	0000375	0.27	0000055	0.74
Puccinia triticina resistance	0000271	0.4	0020057	0.58
salt tolerance	0006046	0.43	0006001	0.71
Fusarium head blight resistance	0000259	0.57	0000673	0.80
resistance to stripe rust	0000358	0.44	0020056	0.46
pre-harvest sprouting resistance	0000619	0.38	0000388	0.54
photoperiod sensitivity	0000652	0.38	0000158	0.65
freezing tolerance	0000207	0.33	0002662	0.57
seed aging	0000357	0.42	0000667	0.67
seed vigor	0000358	0.31	0000355	0.69
thousand kernel weight	0000382	0.53	0000919	0.81
hybrid necrosis	0000439	0.33	0002663	0.63

powdery mildew resistance	0006033	0.4	0000054	0.48
grain length	0002734	0.42	0002626	0.92
stripe rust resistance	0000398	0.37	0020056	0.45
cereal cyst nematode resistance	0006002	0.47	0001111	0.61
heat stress tolerance	0000274	0.39	0000164	0.70
cell wall organization	0006033	0.27	0000728	0.51
pollen development	0000254	0.33	0000245	0.75
anther development	0000254	0.44	0000214	0.53

After evaluating all the corrected CSV files, they were ranked based on their average scores, as shown below. Figure 3 illustrates this ranking and we can see the average corrected score increased.

```
Ranked CSV Files:
[('corrected_response_model1.csv', 0.9147224098391238),
('corrected_response_model2.csv', 0.9107692307692302),
('corrected_response_model3.csv', 0.7316264229886074)]
```

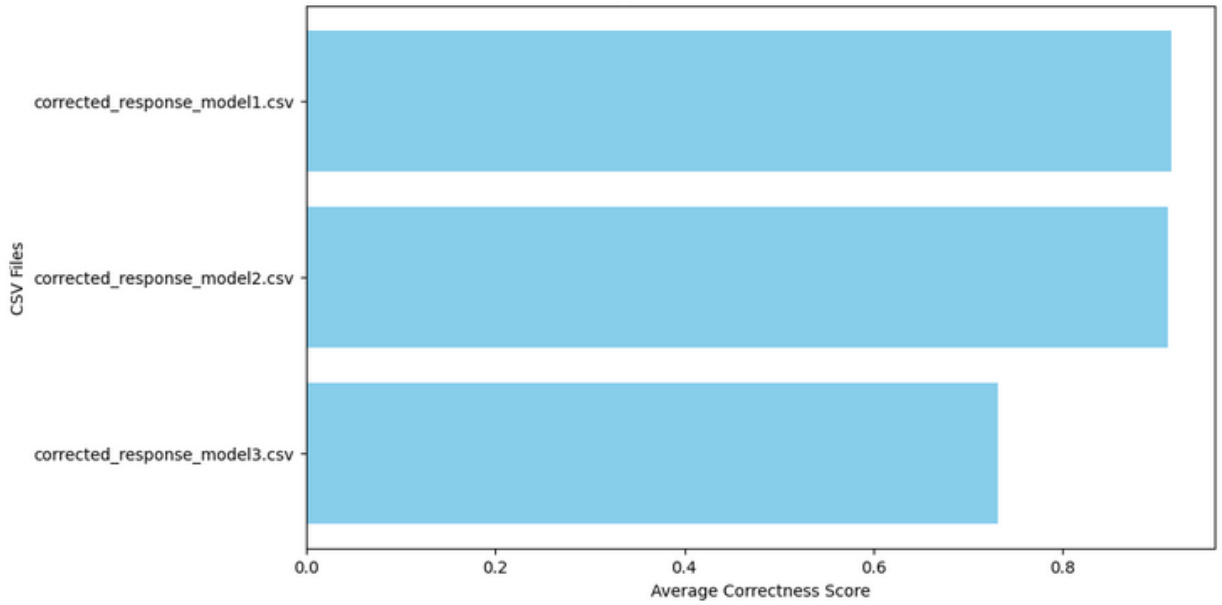

Figure 3: Ranking of corrected CSV Files Based on semantic and hierarchical similarity

Challenges

One of the CSV files, named response_model3, contains some trait_id entries that do not exist in the trait_ontology_details dataset. Before proceeding with the evaluation, it is crucial to address these discrepancies.

We considered two potential solutions to resolve this issue:

1. **Deleting the Mismatched Rows**: While straightforward, this approach risks losing valuable trait information that might be essential for the research.
2. **Correcting the Incorrect IDs**: Instead of removal, we can replace the incorrect IDs with better alternatives. This can be done either randomly or by leveraging the partial similarity of trait names to suggest more appropriate matches.

We opted for the second solution, using partial similarity of trait names to correct the IDs. This method not only preserves the integrity of the dataset but also enhances the quality of the results, ensuring a more meaningful and accurate evaluation. The table 5 show the incorrect and new trait id.

Table 5: the incorrect and new trait id in response_model3

Incorrect trait id	New trait id
0006071	0020055
0006080	0020065
0006081	0000619
0006301	0000254