8 Appendix

8.1 Pre-processing

The set of ontology classes C_O of each ontology O, is composed solely of classes that share the ontology namespace – defined as internal classes – meaning that reused entities are not included in our approach.

All labels from each internal class are stored after a set of standardization steps, such as string normalization and identification of formulas. Additionally, we assign a confidence score that reflects the semantic of the label property: local names as 1.0, labels as 0.95, exact synonyms and internal synonyms as 0.9, other synonyms and external synonyms as 0.85, and formulas as 0.8. This weight is further corrected by the number of names of the same origin associated with the entity.

8.2 Analysis of Logical Definitions in Published Ontologies

An analysis of the logical definitions was conducted in the following ontologies: the Human Phenotype Ontology (HP), the Mammalian Phenotype Ontology (MP), and the Worm Phenotype Ontology (WBP). HP uses 12 ontologies in the construction of its 6140 logical definitions, while MP and WBP use 11 for their 10135 and 964 logical definitions respectively.

After the analysis of the logical definitions, the Relation Ontology and the Basic Formal Ontology were considered to yield little added value, as they provide very few labels and with minimal meaning, such as *part of* or *characteristic of*, and therefore were not used as part of the target ontologies set and removed from the reference mappings.

Moreover, using all possible target ontologies would sharply increase the runtime, leading to the decision of using only a smaller subset. These subsets were chosen in order to maximise the number of unique entities of each ontology used by the logical definitions, up to five target ontologies. The list of target ontologies can be consulted in Table 1, along with how many unique entities from each are used in the construction of the logical definitions.

The reference alignments were built from the logical definitions that contained only entities from the smaller subset of selected target ontologies, with each complex mapping being constructed as an intersection of the target entities.

In HP, a particularity that had to be addressed was the class *abnormal* (PATO 0000460) as it is used in all logical definitions but is only truly relevant for the definition of some of the complex concepts. An approximation was adopted, where this specific class was only added to reference mappings when the source entity included at least one label that used the word "abnormal" or "abnormality".

8.3 Equations

$$m\_prec(m_i) = \frac{\sum_{i=1}^{n} correctness(e_i)}{n} \tag{13}$$

$$m\_rec(m_i) = \frac{\sum_{i=1}^{n} correctness(e_i)}{completeness(E_t, e_r)} \tag{14}$$

8.4 Algorithms

Algorithm 1 Recursive function used in the lexical candidate mapping generation for a single source name.

Input:source name words s_words;
set of contained target names t_names;
set of fixed target names $solution$;
map of target names by source word $wordNames$;
map of words by target name $nameWords$;
map of overlapping target names $overlap$;
Output: set of candidate mappings M

repeat
  $s\_word \leftarrow$ CHOOSE NEXT WORD(s_words)
  $selected \leftarrow$ FIND NAMES WITH SOURCE WORD$(wordNames, s_word)$
  for $t\_name \in selected$ **do**
    $s\_words \leftarrow$ REMOVE WORDS$(nameWords, t_name)$
    $t\_names \leftarrow$ REMOVE NAMES$(overlap, t_name)$
    $solution.append(t\_name)$
until $s\_words == 0$
$M.append(solution)$
return M

Algorithm 2 Large Language Model-based candidate generation

Input:source ontology s;
set of target ontologies T;
Output: set of mappings M

$s\_names \leftarrow$ GET NAMES(s)
$t\_names \leftarrow$ GET NAMES(T)
$embeddings \leftarrow$ GET ALL EMBEDDINGS(s_names, t_names)
for $s\_name \in s\_names$ **do**
  $s\_embedding \leftarrow$ GET NAME EMBEDDING(s_name)
  $best\_combination \leftarrow$ FIND BEST MAPPING$(s_embedding)$
  $m \leftarrow$ CREATE MAPPING$(best_combination)$
  $M.append(m)$
return M