Integrating phenotype ontologies with PhenomeNET

Miguel Angel Rodríguez García[1], Georgios V Gkoutos[2], Paul N Schofield[3], and Robert Hoehndorf[1]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, KSA
{miguel.rodriguezgarcia,robert.hoehndorf}@kaust.edu.sa
[2] College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, Birmingham, B15 2TT, UK,
g.gkoutos@bham.ac.uk
[3] Department of Physiology, Development & Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3EG, UK
pns12@hermes.cam.ac.uk

Abstract. PhenomeNET is a system for disease gene prioritization that includes as one of its components an ontology designed to integrate phenotype ontologies. While not applicable to matching arbitrary ontologies, PhenomeNET can be used to identify related phenotypes in different species, including human, mouse, zebrafish, nematode worm, fruit fly, and yeast. Here, we apply the PhenomeNET to identify related classes from four phenotype and disease ontologies using automated reasoning. We demonstrate that we can identify a large number of mappings, some of which require automated reasoning and cannot easily be identified through lexical approaches alone.

Keywords: PhenomeNET, phenotype ontology

1 System Presentation

1.1 State, purpose, general statement

PhenomeNET [1] was built in 2011 as a system for disease gene discovery and prioritization. PhenomeNET consists of an ontology integrating species-specific phenotype ontologies based on the PATO ontology [2] and relations between anatomical structures and physiological processes, a database of gene-to-phenotype associations, and a measure of similarity between sets of phenotypes. Within PhenomeNET, species-specific phenotype ontologies are combined so that phenotypes observed in different species can be compared directly. The main application of PhenomeNET is the prioritization of candidate genes for human diseases by comparing human disease phenotypes to existing gene-phenotype associations derived from model organisms. In particular, human phenotypes associated with a disease can be compared to phenotypes observed in mouse or other model organisms using the integrated PhenomeNET ontology, and similarity between phenotypes can then be used to indicate the genetic basis of a disease. PhenomeNET has been successfully used to find candidate genes for diseases [1, 3], identify novel pathways [4], and repurpose drugs using mouse model phenotypes [5, 6].

Here, we use the PhenomeNET ontology to identify alignments between phenotypes in different species. We present three versions of the PhenomeNET ontology; the first version consists of the plain ontology using only the axioms provided in the Human Phenotype Ontology (HPO) [7] and the Mammalian Phenotype Ontology (MP) [8]; the second version uses additional lexical mappings and represents them as equivalent class axioms in the ontology; the third version further uses mappings generated by the AgreementMakerLight [9] to generate equivalent class axioms between classes in the PhenomeNET ontology and the Disease Ontology (DO) [10] and the Orphanet Rare Disease Ontology (ORDO) [11].

1.2 Specific techniques used

Phenotype classes in the HP and MP ontologies are formally defined using the Entity-Quality (EQ) pattern [2, 12]. Based on the EQ patterns, a phenotype is decomposed into an affected entity and a quality that specifies how the entity is affected. The Entity will usually be a class taken either from an anatomy ontology or a physiology ontology. For example, the phenotype class *macroglossia* (`HP:0000158`) describes an anatomical abnormality and is defined as equivalent to `'has part' some ('increased size' and ('inheres in' some tongue) and ('has modifier' some abnormal))`, relying on the entity *tongue* (from the UBERON anatomy ontology) and the quality *increased size* (from PATO) in its definition. The class *abnormality of salivation* (`HP:0100755`) is a physiological abnormality and is defined as equivalent to `'has part' some (quality and ('inheres in' some 'saliva secretion') and ('has modifier' some abnormal))`, where *saliva secretion* is a class from the biological process branch of the GO.

The general pattern for defining a phenotype class in both the HP and MP ontologies, given Entity E and Quality Q, is to declare them equivalent to `'has part' some (Q and 'inheres in' some E)`. In some cases, the Entity E is further constrained, e.g., by a location in which a certain process may happen. The "E" classes are generally taken either from the UBERON cross-species anatomy ontology [13] or from the GO. As the use of anatomy and physiology ontologies (UBERON and GO) is shared between MP and HP, it should be possible to integrate both ontologies directly, based on the axiom patterns used to constrain their classes. However, the type of axiom pattern used in both ontologies results in a classification that is primarily based on the PATO ontology, as the Quality Q is the main feature that distinguishes different classes.

In the PhenomeNET ontology, we rewrite all axioms in HP and MP using a pattern-based approach that allows us to utilize axioms from anatomy and physiology ontologies and enrich the classification of phenotype classes [14]. In general, we declare phenotype classes defined using an Entity E and Quality

Q as equivalent to `'has part' some (E and has-quality some Q)` and we further add grouping classes that are defined as equivalent to `'has part' some (('part of' some E) and has-quality some Q)`. The aim of rewriting the axioms is to base the classification of phenotype classes primarily on anatomical or physiological entities instead of the quality, and to utilize the axioms involving parthood in anatomy and physiology ontologies. Crucially, all axioms we generate fall in the OWL 2 EL profile [15]. The first version of the PhenomeNET ontology (PhenomeNET-Plain) consists only of these axioms and no additional mappings.

In addition to this knowledge-based approach to linking the HP and MP ontologies, we also add lexical mappings, mappings derived from cross-references in the ontologies [3], and mappings between HP and MP from BioPortal [16]. Each mapping is added as a single equivalent classes axiom to the first version of the ontology (PhenomeNET-Plain) to generate a version of the PhenomeNET ontology with mappings (PhenomeNET-Map).

Neither version of these ontologies contains the DO or ORDO ontologies, despite there being a significant overlap between the four ontologies. Since neither DO nor ORDO contain axioms that follow a similar pattern to the axioms in HP and MP, we rely exclusively on lexical mappings to integrate DO and ORDO. We use the AgreementMaker Light (AML) [9] in its default settings to generate mappings between HP and DO, HP and ORDO, MP and DO, MP and ORDO, and DO and ORDO. We then add an equivalent class axiom for each mapping AML identifies and that has a score by AML over greater than 0.7. The resulting ontology contains HP, MP, ORDO, and DO, and can be used to generate mappings between these ontologies.

All versions of the PhenomeNET ontology contain the classes from the HP and MP ontologies as well as the subclass axioms between named classes asserted in these ontologies. Furthermore, the PhenomeNET ontology imports the ChEBI [17] and Mouse Pathology [18] ontologies using an OWL import statement. Additionally, PhenomeNET includes all classes from the UBERON anatomy ontology [13], the Gene Ontology [19], the BioSpatial Ontology [20], the Zebrafish Anatomy ontology [21], the PATO ontology [2], the Cell Ontology [22], and the Neuro-Behavior Ontology [23]. However, these ontologies are not directly imported but rather pre-processed so that all disjointness axioms from these ontologies are excluded while all other axioms contained within them are included in the PhenomeNET ontology. The aim of this pre-processing step is to avoid unsatisfiable classes due to different conceptualizations between anatomy and phenotype ontologies, or within anatomy ontologies (Zebrafish Anatomy and UBERON).

Mappings between ontologies included in PhenomeNET are generated using the ELK reasoner [24]. We use ELK to classify the PhenomeNET ontology and identify pairs of equivalent classes C_1 and C_2 that belong to the ontologies to be aligned. These constitute equivalent class mappings. Furthermore, subclass and superclass mappings are generated through queries for sub- and superclasses using ELK.

Ontology	Number of classes	Number of axioms	Mappings added
HP-MP	219,423	1,399,411	0
HP-MP+mappings	219,423	1,400,570	1,160(AML), 639(BioPortal)
HP-MP+DO-ORDO	241,817	1,631,543	1489(AML), 1018(BioPortal) HP-MP: 1,160 (AML), 639(BioPortal); DO-MP: 423 (AML); DO-HP: 1,074; ORDO-MP: 151; ORDO-HP: 531;

Table 1. Number of classes, axioms and mappings in the PhenomeNET ontologies

1.3 Adaptations made for the evaluation

Within PhenomeNET, we use an ontology consisting only of the (rewritten) axioms in MP and HP as well as equivalent class axioms derived from explicit mappings between HP and MP (expressed as `xref` annotation properties). For the evaluation, we further used the AML [9] to generate additional mappings. The AML mappings were generated using the default settings of AML with a confidence cutoff of 0.7. In the case of DOID and ORDO mappings we additionally included 18 mappings derived from BioPortal. Our systems relying on these mappings were submitted as separate submissions.

Initially, we developed our matching system to take into account not only the direct sub- and super-classes, but also all inferred classes. We modified our system to output only the most specific mappings instead for the evaluation; Table 2 shows both the number of direct and inferred mappings.

1.4 Link to the system, parameters file, alignments

Our submission consists of two modules: PhenomeNetBridge and PhenomeNetMatcher. The PhenomeNetBridge module wraps the SEALS infrastructure for the evaluation, and the PhenomeNetMatcher module performs the mappings, using one of three ontologies. Source code for the matching system, including parameter files, and the generated alignments, are available at http://github.com/bio-ontology-research-group/OAEI2016. Code to generate the PhenomeNET ontology is available at
https://github.com/bio-ontology-research-group/phenomeblast/tree/master/fixphenotypes.

2 Results

2.1 Phenotype ontologies: HP and MP

The PhenomeNET ontology is primarily intended to integrate the HP and MP ontologies. Using the axioms in the ontology alone (PhenomeNET-Plain submission), we identify 745 equivalent classes between the HP and MP ontologies

Ontology	HP-MP (≡)	HP-MP (⊑)	DO-ORDO (≡)	DO-ORDO (⊑)
HP-MP	745	2,707 (96,278)	0	0
HP-MP+mappings	1,536	3,999 (107,268)	0	0
HP-MP+DO-ORDO	1,582	4,144 (112,366)	1,527	4,576 (16,838)

Table 2. Equivalent and sub-equivalent classes found in the experiments

Ontology	Precision	Recall	F-Measure	Found	Correct	Reference
HP-MP task						
HP-MP	3.90 %	40.80%	7.10%	6,730	261	639
HP-MP+mappings	6 %	100 %	11.30%	10,698	639	639
HP-MP+DO-ORDO	5.80 %	100 %	10.90%	11,086	639	639
DOID-ORDO task						
HP-MP	0 %	0 %	0 %	0	0	1,018
HP-MP+mappings	0 %	0 %	0 %	0	0	1,018
HP-MP+DO-ORDO	12.70 %	99.90 %	22.50 %	8,036	1,017	1,018

Table 3. Precision, Recall, F-measure in HP-MP and DOID-ORDO experiments

(see Table 2). These correspond to a recall of 40.8% with respect to the reference mappings provided (see Table 3). Additionally, a large number of sub- and super-class mappings can be identified based on querying the ontology using the ELK reasoner [24] for sub- or super-classes in the two ontologies.

The number of pairs of equivalent classes identified increases to 1,536 when adding explicit mappings derived from AML. Of these, 370 are generated both by automated reasoning and are included in AML, 791 are generated from the AML-derived equivalent classes axioms, and 375 could only be derived through the automated reasoning. Total recall with respect to the reference mappings is 100% in this version of PhenomeNET. Additionally, we observe an improvement in the number of equivalent class mappings when adding the ORDO and DO ontologies to the PhenomeNET ontology. The increase in mappings (from 1,536 to 1,582 classes) is a result of additional inferences obtained from adding the mappings from HP and MP to ORDO and DO, and combining them with the axioms in the PhenomeNET ontology. For example, we infer a new mapping between *decreased IgG level* (`MP:0001805`) and *agammaglobulinemia* (`HP:0004432`) based on the equivalence axioms between both classes and *agammaglobulinemia* (`DOID:2583`) generated by AML (based on the shared synonym "hypogammaglobulinemia" between the class in DO and MP). Table 3 summarizes our results with respect to the reference mappings provided in the challenge.

2.2 Disease ontologies: ORDO and DO

PhenomeNET is primarily designed for ontologies that follow the Entity-Quality definition pattern based on the PATO ontology. Neither ORDO nor DO follow this pattern, and ORDO and DO are primarily included in the PhenomeNET ontology through equivalent class axioms based on lexical mappings generated by AML. We achieve a recall of 99.9% with the PhenomeNET-Full ontology. Notably, the mappings we generate are increased by including HP and MP. For example, we identify a mapping between *mandibulofacial dysostosis* (`ORPHANET:155899`) and *treacher collins syndrome* (`DOID:2908`), based on common AML-generated mappings to *mandibulofacial dysostosis* (`HP:0005321`).

2.3 OAEI evaluation

In order to carry out the final evaluation, the OAEI utilized the SEALS infrastructure executed in a Ubuntu Laptop with an Intel Core i7-4600U CPU @ 2.10GHz x 4 and allocating 15Gb of RAM. The system carried out the evaluation according to following criteria:

- Precision and Recall with respect to a voted reference alignment automatically generated by merging/voting the outputs of the participating systems.
- Recall with respect to alignment manually generated.
- Manual assesment of a subset of generating mappings.
- Performance in other tracks.

Different mappings were used to evaluate the participating systems: i) Silver standard with vote 2, ii) Silver standard with vote 3, iii) manually dataset and manual assessment. In the first dataset, PhenomeNET including all mappings reached an F-measure of 0.82 in the HP-MP task, and 0.89 in the DO-ORDO task. In the second evaluation, although the system PhenoMP was able to find the largest number of mappings in HP-MP task, it reached an F-measure of 0.76 in the HP-MP task and 0.94 in the DO-ORDO task. When evaluating against manually created mappings, PhenomeNET achieved a recall of 0.897 in the HP-MP task but could not generate any new mappings between DO and ORDO. For this task, PhenomeNET achieved a precision of 1.0 in the manual assessment of a subset of the generated mappings.

3 General comments

3.1 Comments on the results

PhenomeNET is a system to match phenotypes; as such, it is not a system that can be applied to match ontologies in general. The axiom-based approach in PhenomeNET can be applied to any ontologies that utilize PATO and the Entity-Quality definition patterns [2]. In particular, PhenomeNET can not only be used to integrate MP and HPO, but also has been used to further integrate yeast, fly, worm, slime mold, and fish phenotypes [1, 25]. Furthermore, the combination of semantic matching (using automated reasoning) and lexical matching in PhenomeNET mitigates some of the limitations of using lexical approaches alone, and we demonstrate this by inferring several hundred mappings between HP and MP that cannot be inferred using AML.

However, relying on manually created axioms also has several limitations. In particular, the axioms are created by domain experts, and only about half the classes in MP and HP are constrained by an Entity-Quality based axiom. Furthermore, the quality of the axioms is difficult to assess, and there are distinct differences between HP and MP in how the classes are constrained.

3.2 Discussions on the way to improve the proposed system

One of the main limitations in PhenomeNET is the need for manually created axioms that constrain classes in phenotype ontologies. A possible solution to this approach would be to generate phenotype ontologies fully automatically using anatomy and physiology ontologies as templates and applying the axiom patterns we use in the PhenomeNET [26].

Another limitation of PhenomeNET is the reliance on OWL 2 EL which limits the expressivity of axiom patterns. The choice is mainly due to the size of the PhenomeNET ontology and the complexity of reasoning. However, more complex axiom patterns would enable more comprehensive classification of phenotypes involving absences and abnormalities [14]; experiments with an updated ontology will likely require improvement in OWL reasoning technologies.

4 Conclusions

We have developed an ontology matching system for disease and phenotype ontologies. We generated three different version of the PhenomeNet ontology, each with different information and ontologies included. PhenomeNET is primarily based on deductive inference and automated reasoning, and while it can utilize lexically derived mappings in the ontology generation process, it does not on its own include any lexical matching algorithms. Our results demonstrate that a combination of lexical and semantic approaches may improve upon mappings between ontologies generated using only one of these methods.

References

1. Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: Phenomenet: a whole-phenome approach to disease gene discovery. Nucleic Acids Res **39**(18) (2011) e119
2. Gkoutos, G.V., Green, E.C., Mallon, A.M.M., Hancock, J.M., Davidson, D.: Using ontologies to describe mouse phenotypes. Genome biology **6**(1) (2005) R5
3. Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: An integrative, translational approach to understanding rare and orphan genetically based diseases. Interface Focus **3**(2) (2013) 20120055
4. Hoehndorf, R., Dumontier, M., Gkoutos, G.V.: Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics **28**(16) (2012) 2169–2175
5. Hoehndorf, R., Oellrich, A., Rebholz-Schuhmann, D., Schofield, P.N., Gkoutos, G.V.: Linking PharmGKB to phenotype studies and animal models of disease for drug repurposing. Pacific Symposium on Biocomputing (PSB) (2012) 388–399
6. Hoehndorf, R., Hiebert, T., Hardy, N.W., Schofield, P.N., Gkoutos, G.V., Dumontier, M.: Mouse model phenotypes provide information about human drug targets. Bioinformatics **30**(5) (2014) 719–725

7. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., FitzPatrick, D.R., Eppig, J.T., Jackson, A.P., Freson, K., Girdea, M., Helbig, I., Hurst, J.A., Jähn, J., Jackson, L.G., Kelly, A.M., Ledbetter, D.H., Mansour, S., Martin, C.L., Moss, C., Mumford, A., Ouwehand, W.H., Park, S.M., Riggs, E.R., Scott, R.H., Sisodiya, S., Vooren, S.V., Wapner, R.J., Wilkie, A.O.M., Wright, C.F., Vulto-van Silfhout, A.T., Leeuw, N.d., de Vries, B.B.A., Washingthon, N.L., Smith, C.L., Westerfield, M., Schofield, P., Ruef, B.J., Gkoutos, G.V., Haendel, M., Smedley, D., Lewis, S.E., Robinson, P.N.: The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res **42**(D1) (2014) D966–D974
8. Smith, C.L., Goldsmith, C.A.W., Eppig, J.T.: The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol **6**(1) (2004) R7 DOI:10.1186/gb-2004-6-1-r7.
9. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M. In: The AgreementMakerLight Ontology Matching System. Springer Berlin Heidelberg, Berlin, Heidelberg (2013) 527–541
10. Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H., Schriml, L.M.: Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res **43** (2014) D1071–D1078
11. Sarntivijai, S., Vasant, D., Jupp, S., Saunders, G., Bento, A.P., Gonzalez, D., Betts, J., Hasan, S., Koscielny, G., Dunham, I., Parkinson, H., Malone, J.: Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. Journal of Biomedical Semantics **7**(1) (2016) 1–11
12. Mungall, C., Gkoutos, G., Smith, C., Haendel, M., Lewis, S., Ashburner, M.: Integrating phenotype ontologies across multiple species. Genome Biol **11**(1) (2010) R2+
13. Mungall, C., Torniai, C., Gkoutos, G., Lewis, S., Haendel, M.: Uberon, an integrative multi-species anatomy ontology. Genome Biology **13**(1) (2012) R5
14. Hoehndorf, R., Oellrich, A., Rebholz-Schuhmann, D.: Interoperability between phenotype and anatomy ontologies. Bioinformatics **26**(24) (10 2010) 3112 – 3118
15. Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C.: Owl 2 web ontology language: Profiles. Recommendation, World Wide Web Consortium (W3C) (2009)
16. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A.A., Chute, C.G., Musen, M.A.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research **37**(Web Server issue) (July 2009) W170–173
17. Degtyarenko, K., Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research (2007)
18. Schofield, P.N., Sundberg, J.P., Sundberg, B.A., McKerlie, C., Gkoutos, G.V.: The mouse pathology ontology, mpath; structure and applications. Journal of Biomedical Semantics **4**(1) (2013) 1–8
19. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, M.J., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Tarver, L.I., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1) (May 2000) 25–29

20. Balhoff, J.P., Mik, I., Yoder, M.J., Mullins, P.L., Deans, A.R.: A semantic model for species description applied to the ensign wasps (hymenoptera: Evaniidae) of new caledonia. Systematic Biology **62**(5) (2013) 639–659
21. Dahdul, W.M., Balhoff, J.P., Engeman, J., Grande, T., Hilton, E.J., Kothari, C., Lapp, H., Lundberg, J.G., Midford, P.E., Vision, T.J., Westerfield, M., Mabee, P.M.: Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS One **5**(5) (2010) e10708
22. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. Genome Biology **6**(2) (2005)
23. Hoehndorf, R., Hancock, J.M., Hardy, N.W., Mallon, A.M., Schofield, P.N., Gkoutos, G.V.: Analyzing gene expression data in mice with the Neuro Behavior Ontology. Mamm Genome **25**(1-2) (2014) 32–40
24. Kazakov, Y., Krötzsch, M., Simancik, F.: The incredible elk. Journal of Automated Reasoning **53**(1) (2014) 1–61
25. Hoehndorf, R., Hardy, N.W., Osumi-Sutherland, D., Tweedie, S., Schofield, P.N., Gkoutos, G.V.: Systematic analysis of experimental phenotype data reveals gene functions. PLoS ONE **8**(4) (04 2013) e60847
26. Hoehndorf, R., Harris, M.A., Herre, H., Rustici, G., Gkoutos, G.V.: Semantic integration of physiology phenotypes with an application to the cellular phenotype ontology. Bioinformatics **28**(13) (2012) 1783–1789