Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2022 December ; 2022: 3274–3279. doi:10.1109/bibm55620.2022.9995614.

Identifying Missing IS-A Relations in Orphanet Rare Disease Ontology

Maryamsadat Mohtashamian[†],
The University of Texas Health, Science Center at Houston, School of Biomedical Informatics, Houston, TX

Rashmie Abeysinghe[†],
The University of Texas Health, Science Center at Houston, McGovern Medical School, Houston, TX

Xubing Hao,
The University of Texas Health, Science Center at Houston, School of Biomedical Informatics, Houston, TX

Licong Cui[*]
The University of Texas Health, Science Center at Houston, School of Biomedical Informatics, Houston, TX

Abstract

The Orphanet Rare Disease Ontology (ORDO) provides a structured vocabulary encapsulating rare diseases. Downstream applications of ORDO depend on its accuracy to effectively perform their tasks. In this paper, we implement an automated quality assurance pipeline to identify missing *is-a* relations in ORDO. We first obtain lexical features from concept names. Then we generate related and unrelated feature sharing concept-pairs, where a feature sharing concept-pair can further generate derived term-pairs. If an unrelated and related feature sharing concept-pair generate the same derived term-pair, then we suggest a potential missing *is-a* relation between the unrelated feature sharing concept-pair. Applying this approach on the 2022-06-27 release of ORDO, we obtained 705 potential missing *is-a* relations. Leveraging external ontological information in the Unified Medical Language System, we validated 164 missing *is-a* relations. This indicates that our approach is a promising way to audit *is-a* relations in ORDO, even though further domain expert evaluation is still needed to validate the remaining potential missing *is-a* relations identified.

Keywords

Rare diseases; Orphanet; Orphanet rare disease ontology; ontology quality assurance

[*]Corresponding author: licong.cui@uth.tmc.edu.
[†]Contributed equally

I. Introduction

According to the Orphan Drug Act, a rare disease is a disease or a condition that affects less than 200,000 people in the United States [1]. Over 30 million individuals in the US are affected by more than 7,000 rare diseases. Many can be life-threatening and without any treatments. Developing treatments strategies is a challenge due to various reasons such as insufficient information and inability to conduct clinical trials due to smaller number of patients [2].

In 1997, Orphanet was established in France to collect scarce information on rare disease in order to improve diagnosis, care, and treatment. Orphanet together with the European Bioinformatics Institute jointly developed the Orphanet Rare Disease Ontology (ORDO) capturing the relationships between rare diseases, genes and other related information. ORDO also contains links to other biomedical ontologies, databases, and classification systems. ORDO is updated and released every six months [3]. Figure 1 shows the general hierarchy of ORDO.

The number of investigations leveraging big data in biomedicine is increasing in a rapid manner due to the easy-to-use tools available and the reduced computational costs associated with these analyses. However, most biomedical data are heterogeneous spread across different systems. This heterogeneity makes it difficult to obtain valuable insights from this biomedical data. Biomedical ontologies like ORDO address this issue by playing a vital role in data integration, retrieval, reasoning and decision support by providing a common language enabling effective use of biomedical data [4]. However, errors existing in biomedical ontologies could be problematic in their effective use and may bring about questions about their trustworthiness. Quality control pipelines are generally included as part of their management lifecycle to identify and fix errors. However, similar to Software Quality Assurance, it is impossible to identify all errors at the time of a release. Many biomedical ontologies rely on user feedback as part of its quality assurance effort. The size of modern biomedical ontologies and their complexity has become a barrier in utilizing manual strategies to identify errors. Hence, the development of automated or semi-automated methods has become a pressing need in Ontology Quality Assurance.

Identifying errors in a biomedical ontology is a discovery-oriented task. Methods developed attempt to discover different types of quality issues in an ontology. Different strategies have been explored for this purpose [5]. For example, abstraction networks are a widely used quality assurance technique based on graph summarization [6]–[9]. An abstraction of the graph structure is obtained by grouping terms in an ontology based on certain criteria. Abstraction network-defined characteristics together with manual review is used to identify errors. He et al. has investigated differences between hierarchies of two ontologies to import concept from one ontology to another [10], [11]. Agrawal et al. has explored rule-based and machine learning-based strategies to identify lexically similar concepts that should be modeled similarly. Different modelling strategies among such lexically concepts may potentially denote errors [12]–[14]. In previous work, we have investigated non-lattice subgraphs (graph fragments that violate the desirable lattice property) to uncover missing hierarchical relations and concepts in SNOMED CT, the National Cancer Institute (NCI)

thesaurus, and Gene Ontology [15]–[24]. In addition, we have introduced an automated, lexical-based quality assurance pipeline where hierarchically related and unrelated concept-pairs with the same difference are leveraged to identify missing and erroneous hierarchical relations in an ontology [25]–[28].

In this paper, we adapt (and implement) this automated quality assurance pipeline to ORDO for identification of missing *is-a* relations. According to our knowledge, this work is the first effort towards developing a systematic automated approach for quality assurance of ORDO.

The rest of the paper is organized as follows. Section II discusses the detailed steps of the automated quality assurance pipeline. Section III shows the results obtained applying the approach on ORDO. Section IV contains a discussion of the results, limitation of the approach and future directions. Section V concludes this paper.

II. Methods

In this work, we use the OWL (Web Ontology Language) release file of the 2022-06-27 release of ORDO. We first extract the concept names and relations from the OWL file with Owlready2: a python package for ontology-oriented programming [29]. Our fully automated pipeline is based on lexical features of concept names to suggest missing *is-a* relations in ORDO. Our method leverages derived term-pairs; which denotes the lexical differences between a pair of concepts. If the same derived term pair is observed among a hierarchically related and unrelated concept-pairs that share lexical feature(s), we suggest a potential missing *is-a* relation between the unrelated concept-pair. We use the Unified Medical Language System (UMLS) as a source to validate the identified missing *is-a* relations. The detailed steps of the approach are as follows.

A. Obtaining lexical features of concept

Each concept in ORDO has a name. For example, the concept with the ORDO identifier *Orphanet:98497* has the name "*Genetic peripheral neuropathy*". We obtain a set of lexical features from these concept names by performing the following operations on the concept names:

- converting the name to lower case
- tokenizing the name to words
- stemming the words
- removing duplicated stemmed words.

For instance, for the concept "*Genetic peripheral neuropathy*" (Orphanet:98497), the lexical features would be {'genet', 'peripher', 'neuropathi'}. Note that as can be seen from the example, word stems are not always real words. Stemming is performed in this work to normalize different variations of the same word. We use the Snowball Stemmer with the python natural language processing library Natural Language Toolkit (NLTK) for stemming [30].

B. Extracting feature sharing concept-pairs

A pair of concepts is considered to be a feature sharing concept-pair if they have at least a one common lexical feature. For instance, consider the concepts *"Genetic eye tumor"* (Orphanet:183619) with lexical features {'genet', 'eye', 'tumor'} and *"Rare genetic eye disease"* (Orphanet:101435) with lexical features {'rare', 'genet', 'eye', 'diseas'}. These two concepts will form a feature sharing concept-pair as they have the common lexical features 'genet' and 'eye'.

Feature sharing concept-pairs are further divided into two categories: related and unrelated. If a feature sharing concept-pair is connected by a direct or indirect *is-a* relation, it would be considered as related and if not it would be considered as unrelated. For example, the feature sharing concept-pair *"Genetic eye tumor"* (Orphanet:183619) and *"Rare genetic eye disease"* (Orphanet:101435) is related since there exists a relation such that:

"Genetic eye tumor" is-a "Rare genetic eye disease".

However, the feature sharing concept-pair "*Neuroendocrine tumor of the colon*" (Orphanet:100080) and "*Rare epithelial tumor of colon*" (Orphanet:423991) is considered to be unrelated since there does not exists a direct or indirect *is-a* relation between them.

C. Extracting derived term pairs from feature sharing concept-pairs

Let $L(A)$ and $L(B)$ represent lexical features of feature sharing concepts A and B respectively. A Derived Term Pair (DTP) obtained by this feature sharing concept-pair is defined as:

$$DTP(A, B) = (\{L(A) - L(B)\}, \{L(B) - L(A)\})$$

In other words, a DTP is constructed by removing the common lexical features from the lexical features of each concept. For instance, the related feature sharing concept-pair "*Aggressive primary cutaneous B-cell lymphoma*" (Orphanet:178554), with lexical features: {'aggress' 'primari' 'cutan' 'b-cell' 'lymphoma'} and "*Aggressive B-cell Non-Hodgkin lymphoma*" (Orphanet:300846) with lexical features {'aggress' 'b-cell' 'non-hodgkin' 'lymphoma'} in Figure 2 has the common lexical features 'aggress, 'b-cell', and 'lymphoma'. Removing these from both sets of lexical features would yield the DTP: ({'cutan', 'primari'}, {'non-hodgkin'}).

Note that the DTP is directional, i.e., $DTP(A, B) \neq DTP(B,A)$. In addition, in situations where the lexical features of one concept is a subset of another, i.e., $L(A) \subset L(B)$, then one set of the DTP would be an empty set. For example, the unrelated feature sharing concept-pair "*Pure mitochondrial myopathy*" (Orphanet:254854), "*Mitochondrial myopathy*" (Orphanet:206966) in Figure 3 would generate the DTP: ({'pure'}, {}) since the lexical features of Orphanet:206966 is a subset of Orphanet:254854.

If $L(A) = L(B)$, then, the DTP would be two empty sets. We would not consider such DTPs in this work. In addition, if both the sets in a DTP are all stop words, such DTPs are ignored as well. The stop words considered in this work are: 'with', 'of', 'and', 'or', 'and/

or', 'no', 'not', 'without', 'due to', 'secondary to', 'except', 'by', 'after', 'able', 'removal', 'replacement', 'NOS'.

D. Identifying missing is-a relations

If a related feature sharing concept-pair (*A, B*) and an unrelated feature sharing concept-pair (*C, D*) generate the same DTP, i.e.,

$$DTP(A, B) = DTP(C, D)$$

then, this is considered to be denoting a missing *is-a* relation between *C* and *D* such that *C is-a D*.

For example, in Figure 2, the related feature sharing concept-pair "*Aggressive primary cutaneous B-cell lymphoma*" (Orphanet:178554), "*Aggressive B-cell non-Hodgkin lymphoma"* (Orphanet:300846) generate the DTP: ({'cutan', 'primari'}, {'non-hodgkin'}). The same DTP is also obtained by the unrelated feature sharing concept-pair sharing concept-pair "*Primary cutaneous lymphoma*" (Orphanet:542), *"Non-Hodgkin lymphoma"* (Orphanet:547). Therefore, our approach suggests the missing *is-a* relation:

Orphanet_542 *is-a* Orphanet_547

between the unrelated feature sharing concept-pair.

Similarly, in Figure 3, the DTP ({'pure'}, {}) is obtained from the related feature sharing concept-pair "*Pure hereditary spastic paraplegia*" (Orphanet:102012) and "*Hereditary spastic paraplegia*" (Orphanet:685) as well as the unrelated feature sharing concept-pair "*Pure mitochondrial myopathy*" (Orphanet:254854), "*Mitochondrial myopathy*" (Orphanet_206966). Therefore, in this instance, we suggest:

Orphanet:254854 *is-a* Orphanet_206966

E. Validating identified missing is-a relations

The identified cases are potentially missing *is-a* relations needing to be further validated to confirm their correctness. We leverage the Unified Medical Language System (UMLS) which integrates and links term from many biomedical terminologies. The basic building blocks of the UMLS are atoms which are concept names from different source terminologies. Each atom will have an atom unique identifier (AUI). A UMLS concept with a Concept Unique Identifier (CUI) aggregates all the atoms that represent a single meaning [31], [32]. For example, the UMLS concept "*Fracture of carpal bone*" (with CUI C0016644) is linked to atom "*Fracture of carpal bone*" (with AUI A3023601) from SNOMEDCT and atom "*Fractured carpal bone*" (with AUI A32452940) from Human Phenotype Ontology.

To validate a potential missing *is-a* relation, we first try to map the two concepts to UMLS atoms. Note that what we perform is a normalized map where the ORDO concepts in the potential missing *is-a* relations and the UMLS atoms are normalized. The normalization process includes lowercase conversion, lemmatization, stop word removal, and synonym replacement as performed in one of our previous works [33].

After obtaining mappings, we further check if UMLS records a direct or indirect hierarchical relation between the mapped atoms. If so, the missing *is-a* relation is considered to be validated.

III. Results

The 2022-06-27 release of ORDO contains 15,302 concepts. Applying the above discussed automated pipeline on this version of ORDO, we obtained 705 potential missing *is-a* relations. Table 1 shows the top 10 DTPs that identified the most number of potential missing *is-a* relations. For example, the DTP ({'genet'}, {}) identified 154 potential missing *is-a* relations.

A. Validation of potential missing is-a relations

Out of the 705 potential missing *is-a* relations identified by the approach, 164 were validated through UMLS. These 164 were validated through 210 distinct UMLS atom-pairs meaning the concepts can be mapped to and relations validated by multiple atom-pairs. For example, our method suggested a missing *is-a* relation between the concepts "*Chronic endophthalmitis*" (Orphanet:279891) and "*Endophthalmitis*" (Orphanet:199323). Orphanet:279891 was mapped to atom "*Chronic endophthalmitis*" (with AUI A2892272) from SNOMED CT and the atom "*chronic endophthalmitis*" (with AUI A14149447) from MEDCIN. In addition, "*Endophthalmitis*" (Orphanet:199323) was mapped to both the atoms "*Endophthalmitis*" (with AUI A2881177) from SNOMED CT and the atom "*endophthalmitis*" (with AUI A13899807) from MEDCIN. UMLS records *is-a* relations: A2892272 *is-a* A2881177 in the SNOMED CT and A14149447 *is-a* A13899807 in MEDCIN. Therefore, this missing *is-a* relations has been validated through both SNOMED CT and MEDCIN.

Table 2 shows 10 validated cases of missing *is-a* relations out of the 164 that were validated in total. For instance, the relation "*Primary cutaneous lymphoma*" (Orphanet:542) *is-a "Non-Hodgkin lymphoma"* (Orphanet:547) is a valid missing *is-a* relation.

IV. Discussion

In this paper, we introduced an automated pipeline to identify missing *is-a* relations in ORDO. Our approach is based on the Difference Term Pair (DTP) which holds unique lexical features of a pair of concepts. From purely a lexical perspective, in a related feature sharing concept-pair, the DTP holds the information that makes the relation hold. This is because the common lexical features removed from the DTP does not contribute to the relation as they exist in both the concepts. Therefore, when an unrelated concept-pair exhibit the same DTP, we consider this as evidence to the potential existence of a relation.

A. The distance between the concepts in the valid missing is-a relations

For the 164 validated missing *is-a* relations, we further checked the distance between their mapped atoms in the respective source terminologies which they were validated from. The distance was measured as the number of direct *is-a* relations linking the two atoms. Table 3 shows the distribution of the distances. As can be seen, a vast majority of cases are with a

distance of 1 which means that they are direct *is-a* relations in the source terminologies. Direct *is-a* relations are generally easily fixable than indirect ones. To fix an indirect relation, intermediate missing relations need to be identified which could be complicated.

B. Comparison with related work

The approach discussed in this paper is an adaptation of a previously introduced approach to identify missing *is-a* relations in several biomedical terminologies including Gene Ontology and SNOMED CT [25–28]. However, the difference between this approach and the previous approaches is the usage of stemming on the lexical features which further normalized the lexical features.

C. Limitations and future work

As lexical features of concepts, we only considered the stemmed set of words in their names. However, additional attributes of concepts such as lexical features of the synonyms or ancestors could be considered. These can not only be obtained from ORDO, may also be obtainable from external ontological sources in UMLS.

The fully automated UMLS-based validation is quick and efficient since it does not require manual review. In addition, it is also able to validate a considerable number of missing *is-a* relations (23%). However, with this type of validation, we are not able to measure the precision of the method. Therefore, we propose to perform a manual review of a random sample of potential missing *is-a* relations to properly quantify the performance of the approach.

We will submit the 164 validated potential missing *is-a* relations to ORDO developers so that after internal review, they can make necessary changes to a future release of the ontology based on our findings.

V. Conclusion

In this paper, we implemented a fully automated lexical approach to identify missing is-a relations in the Orphanet Rare Disease Ontology. Our method included obtaining lexical features from concept names and generating feature sharing concept-pairs. The feature sharing concept-pairs further generated derived term-pairs. If the same derived term-pair could be generated from both a related and an unrelated feature sharing concept-pair, then we suggested a potential missing is-a relation between the unrelated feature sharing concept-pair. Applying this approach to the 2022-06-27 release of the Orphanet Rare Disease Ontology, we obtained 705 potential missing is-a relations. Leveraging the is-a relations of external ontologies in the Unified Medical Language System, we validated 164 missing is-a relations. This approach seems to show promise in auditing is-a relations in the Orphanet Rare Disease Ontology, though further manual review is needed to confirm and validate the rest of the potential missing is-a relations that could not be validated automatically.

Acknowledgments

This work was supported by the National Science Foundation (NSF) through grant 2047001, National Institutes of Health (NIH) through grant R01LM013335, R21AG068994, and R01NS116287. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

[1]. "Orphan Drug Act - Relevant Excerpts," FDA, Accessed: Oct. 07, 2022. [Online]. Available: https://www.fda.gov/industry/designating-orphan-product-drugs-and-biological-products/orphan-drug-act-relevant-excerpts

[2]. "Rare Diseases at FDA," FDA, https://www.fda.gov/patients/rare-diseases-fda (accessed Oct. 07, 2022).

[3]. "ORDO – Orphadata." https://www.orphadata.com/ordo/ (accessed Oct. 07, 2022).

[4]. Bodenreider O, "Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support," Yearb. Med. Inform, vol. 17, no. 01, pp. 67–79, Aug. 2008.

[5]. Zhu X, Fan J-W, Baorto DM, Weng C, and Cimino JJ, "A review of auditing methods applied to the content of controlled biomedical terminologies," J. Biomed. Inform, vol. 42, no. 3, pp. 413–425, Jun. 2009. [PubMed: 19285571]

[6]. Halper M, Gu H, Perl Y, and Ochs C, "Abstraction networks for terminologies: supporting management of 'big knowledge,'" Artif. Intell. Med, vol. 64, no. 1, pp. 1–16, May 2015. [PubMed: 25890687]

[7]. Geller J, Ochs C, Perl Y, and Xu J, "New Abstraction Networks and a New Visualization Tool in Support of Auditing the SNOMED CT Content," AMIA. Annu. Symp. Proc, vol. 2012, pp. 237–246, Nov. 2012. [PubMed: 23304293]

[8]. Ochs C, Perl Y, Halper M, Geller J, and Lomax J, "Quality assurance of the gene ontology using abstraction networks," J. Bioinform. Comput. Biol, vol. 14, no. 03, p. 1642001, Jun. 2016. [PubMed: 27301779]

[9]. Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, Hripcsak G, "A Tribal abstraction network for SNOMED CT target hierarchies without attribute relationships," J. Am. Med. Inform. Assoc, vol. 22, no. 3, pp. 628–39, May. 2015, [PubMed: 25332354]

[10]. Keloth VK, He Z, Chen Y, and Geller J, "Leveraging Horizontal Density Differences between Ontologies to Identify Missing Child Concepts: A Proof of Concept," AMIA. Annu. Symp. Proc, vol. 2018, pp. 644–653, Dec. 2018. [PubMed: 30815106]

[11]. He Z, Geller J, and Chen Y, "A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization," Artif. Intell. Med, vol. 64, no. 1, pp. 29–40, May 2015. [PubMed: 25890688]

[12]. Agrawal A, Perl Y, Ochs C, and Elhanan G, "Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators," in Proceedings - 2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, Dec. 2015, pp. 476–483.

[13]. Agrawal A and Qazi K, "Detecting modeling inconsistencies in SNOMED CT using a machine learning technique," Methods, vol. 179, pp. 111–118, Jul. 2020. [PubMed: 32442671]

[14]. Agrawal A, "Evaluating lexical similarity and modeling discrepancies in the procedure hierarchy of SNOMED CT," BMC Med. Inform. Decis. Mak, vol. 18, no. 4, p. 88, Dec. 2018. [PubMed: 30537959]

[15]. Zhang G-Q and Bodenreider O, "Large-scale, Exhaustive Lattice-based Structural Auditing of SNOMED CT," AMIA. Annu. Symp. Proc, vol. 2010, pp. 922–926, 2010. [PubMed: 21347113]

[16]. Abeysinghe R, Zheng F, and Cui L, "A Comparison of Exhaustive and Non-lattice-based Methods for Auditing Hierarchical Relations in Gene Ontology," AMIA. Annu. Symp. Proc, vol. 2021, pp. 177–186, Feb. 2022. [PubMed: 35308995]

[17]. Zheng F, Abeysinghe R, and Cui L, "A Hybrid Method to Detect Missing Hierarchical Relations in NCI Thesaurus," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, Nov. 2019, pp. 1948–1953.

[18]. Cui L, Bodenreider O, Shi J, and Zhang G-Q, “Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs,” J. Biomed. Inform, vol. 78, pp. 177–184, Feb. 2018. [PubMed: 29274386]

[19]. Zheng F, Abeysinghe R, Sioutos N, Whiteman L, Remennik L, and Cui L, “Detecting missing IS-A relations in the NCI Thesaurus using an enhanced hybrid approach,” BMC Med. Inform. Decis. Mak, vol. 20, no. 10, p. 273, Dec. 2020. [PubMed: 33319703]

[20]. Cui L et al. , “Enhancing the Quality of Hierarchic Relations in the National Cancer Institute Thesaurus to Enable Faceted Query of Cancer Registry Data,” JCO Clin. Cancer Inform, vol. 4, pp. 392–398, May 2020. [PubMed: 32374632]

[21]. Hao X, Abeysinghe R, Zheng F, and Cui L, “Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT,” Proc. IEEE Int. Conf. Bioinforma. Biomed, vol. 2021, pp. 1805–1812, Dec. 2021.

[22]. Abeysinghe R, Brooks MA, and Cui L, “Leveraging Non-lattice Subgraphs to Audit Hierarchical Relations in NCI Thesaurus,” AMIA. Annu. Symp. Proc, vol. 2019, pp. 982–991, Mar. 2020. [PubMed: 32308895]

[23]. Cui L, Zhu W, Tao S, Case JT, Bodenreider O, and Zhang G-Q, “Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT,” J. Am. Med. Inform. Assoc, vol. 24, no. 4, pp. 788–798, Jul. 2017. [PubMed: 28339775]

[24]. Abeysinghe R, Brooks MA, Talbert J, and Licong C, “Quality Assurance of NCI Thesaurus by Mining Structural-Lexical Patterns,” AMIA. Annu. Symp. Proc, vol. 2017, pp. 364–373, Apr. 2018. [PubMed: 29854100]

[25]. Abeysinghe R, Hinderer EW, Moseley HNB, and Cui L, “Auditing subtype inconsistencies among gene ontology concepts,” in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov. 2017, pp. 1242–1245.

[26]. Abeysinghe R et al. , “Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method,” J. Biomed. Inform, vol. 134, p. 104162, Oct. 2022. [PubMed: 36029954]

[27]. Abeysinghe R, Yang Y, Bartels M, Zheng WJ, Cui L, “An evidence-based lexical pattern approach for quality assurance of Gene Ontology relations.” Briefings in Bioinformatics, vol. 23, no. 3, bbac122, May 2022.

[28]. Manuel W, Abeysinghe R, He Y, Tao C, and Cui L, “Identification of missing hierarchical relations in the vaccine ontology using acquired term pairs,” J. Biomed. Semant, vol. 13, no. 1, p. 22, Aug. 2022.

[29]. “Welcome to Owlready2’s documentation! — Owlready2 0.36 documentation.” https://owlready2.readthedocs.io/en/v0.37/ (accessed Oct. 07, 2022).

[30]. Loper E and Bird S, “NLTK: The Natural Language Toolkit.” arXiv, May 17, 2002. Accessed: Oct. 09, 2022. [Online]. Available: http://arxiv.org/abs/cs/0205028

[31]. Metathesaurus. National Library of Medicine (US), 2021. Accessed: Oct. 10, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9684/

[32]. Bodenreider O, “The Unified Medical Language System (UMLS): integrating biomedical terminology,” Nucleic Acids Res, vol. 32, no. suppl_1, pp. D267–D270, Jan. 2004. [PubMed: 14681409]

[33]. Hao X, Abeysinghe R, Zheng F, and Cui L, “Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT,” Proc. IEEE Int. Conf. Bioinforma. Biomed, vol. 2021, pp. 1805–1812, Dec. 2021.

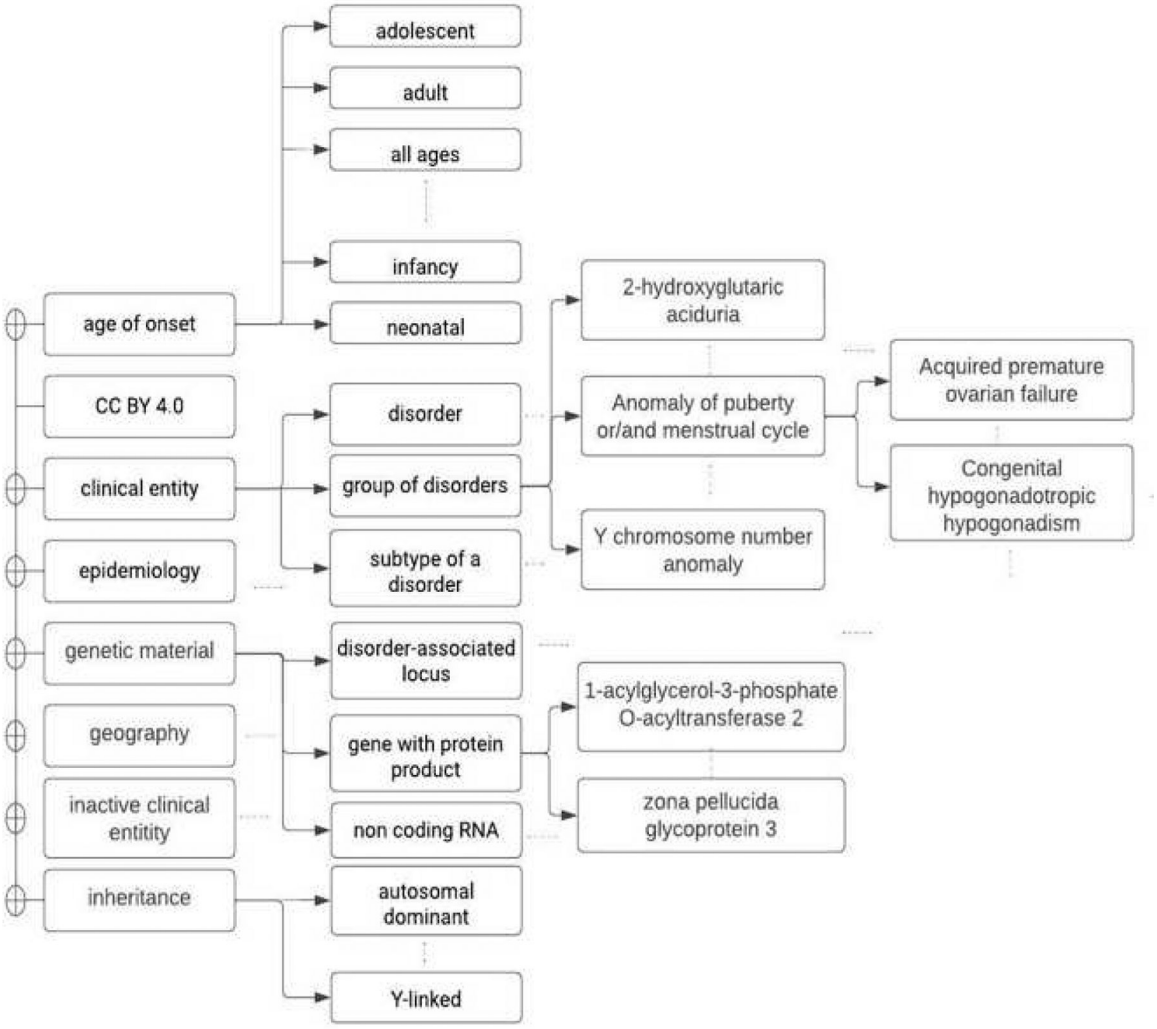

Figure 1.
The hierarchy of the 2022-06-27 release of ORDO

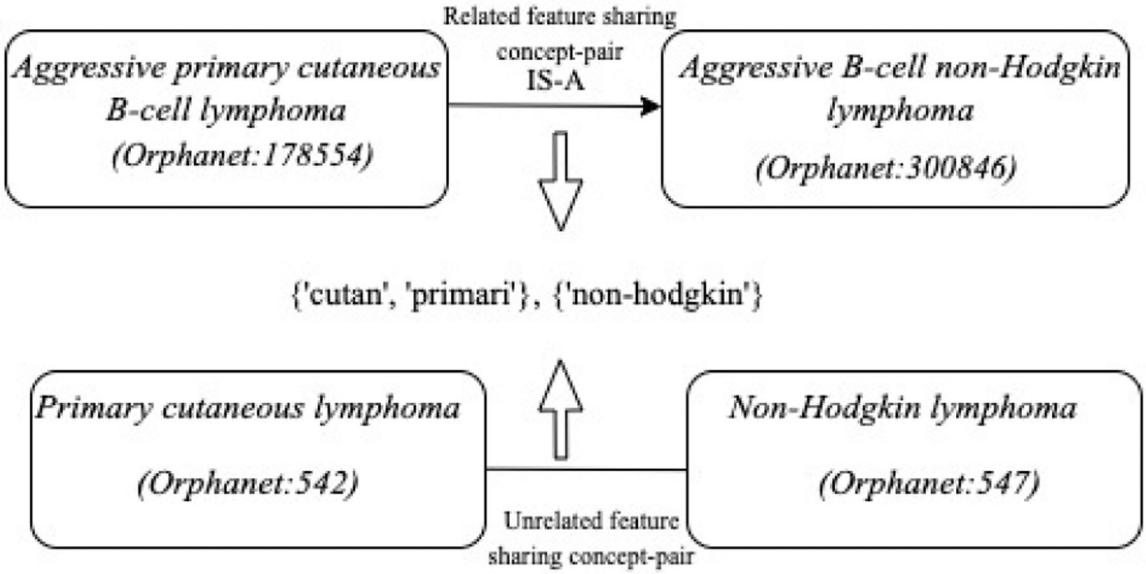

Figure 2.

Related feature sharing concept-pair “Aggressive primary cutaneous B-cell lymphoma” (Orphanet:178554), “Aggressive B-cell non-Hodgkin lymphoma” (Orphanet:300846) and unrelated feature sharing concept-pair “Primary cutaneous lymphoma” (Orphanet:542), “Non-Hodgkin lymphoma” (Orphanet:547). Both the concept-pairs derive the same DTP: ({‘cutan’, ‘primari’}, {‘non-hodgkin’}).

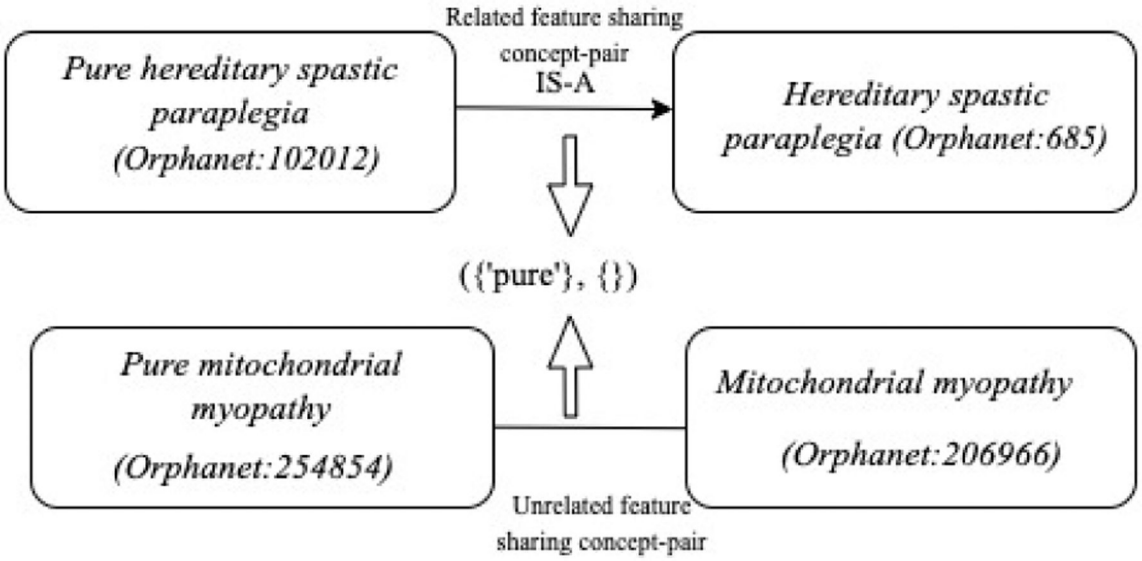

Figure 3.

Related feature sharing concept-pair “Pure hereditary spastic paraplegia” (Orphanet:102012). “Hereditary spastic paraplegia” (Orphanet:685) and unrelated feature sharing concept-pair “Pure mitochondrial myopathy” (Orphanet:254854), “Mitochondrial myopathy” (Orphanet:206966). Both the concept-pairs derive the same DTP: ({‘pure’}, {}).

Table 1.

The 10 DTPs that identified the most number of potential missing is-a relations.

DTP	Number of potential missing *is-a* obtained
({genet}, {})	154
({type 1}, {})	58
({type, 2}, {})	57
({type 3}, {})	32
({genet}, {rare}	26
({autosom, domin}, {})	23
({x-link}, {})	17
({recess, autosom}, {})	17
({isol}, {})	10
({juvenil}, {})	10

Table 2.

Ten examples of valid missing is-a relations identified by our method.

Descendant	Ancestor
Primary cutaneous lymphoma (Orphanet:542)	*Non-Hodgkin lymphoma* (Orphanet:547)
Von Willebrand disease type 3 (Orphanet:166096)	*Von Willebrand disease* (Orphanet:903)
Autosomal dominant Robinow syndrome (Orphanet:3107)	*Robinow syndrome* (Orphanet:97360)
Bilateral generalized polymicrogyria (Orphanet:208447)	*Bilateral polymicrogyria* (0rphanet:268940)
IgG4-related systemic disease (Orphanet:596448)	*IgG4-related disease* (Orphanet:284264)
Peeling skin syndrome type C (Orphanet:263558)	*Peeling skin syndrome* (Orphanet:817)
Pseudohypoaldosteronism type 2A (0rphanet:88938)	*Pseudohypoaldosteronism type 2* (0rphanet:757)
Congenital stromal corneal Dystrophy (Orphanet:101068)	*Stromal corneal dystrophy* (Orphanet_98626)
Acquired motor neuron disease (Orphanet:98506)	*Motor neuron disease* (Orphanet:98503)
Transient congenital hypothyroidism (Orphanet:178045/)	*Congenital hypothyroidism* (Orphanet:442)

Table 3.

The distance between the mapped atoms for the valid missing is-a relations.

Distance	Number of valid missing *is-a* relations
1	160
2	2
3	2