The Orphanet Ontology – Formalising Knowledge on Rare and Genetic Diseases

Maja Miličić
National Genetics Reference Laboratory Manchester
maja.milicic@manchester.ac.uk

Michael Cornell
National Genetics Reference Laboratory Manchester
michael.cornell@cmft.nhs.uk

Andrew Devereau
National Genetics Reference Laboratory Manchester
andrew.deverau@cmft.nhs.uk

ABSTRACT

This poster presents an ongoing project to formalise knowledge on rare and genetic diseases from Orphanet [1] database. Orphanet data on rare diseases, including their classification, related clinical signs, genes, mode of inheritance and epidemiological data, was translated into an OWL ontology. This translation revealed a large number of inconsistencies, mainly in the classification of diseases. As a result the Orphanet classifications have been revised and improved. In addition, Orphanet rare disease terms are being mapped to established medical terminologies including Snomed CT. The revised edition of Orphanet data, together with mappings of Orphanet rare diseases into Snomed CT, will be used to revise Snomed CT's data on rare diseases.

1. INTRODUCTION

Rare and genetic diseases are under-represented in established medical ontologies/terminologies such as Snomed CT [2] and MeSH [3]. This poses a problem in situations where encoding of rare and genetic diseases is necessary, for example in genetic labs. Although comprehensive general medical ontologies, such as Snomed CT, contain some data on genetic diseases, as well as on genes and clinical signs, it is out of their scope to encode relations between them.

The Orphanet portal (http://www.orpha.net) and database on rare diseases were founded by Inserm in 1997 and contain comprehensive data on approximately 7000 rare, mostly genetic, diseases. The database is frequently revised and updated. The data includes classification of rare diseases, the association of clinical signs and genes with diseases, their inheritance modes and epidemiological data. Orphanet data is becoming a standard for the classification of rare and genetic diseases and it is going to be incorporated into the next edition of ICD (International Classification of Diseases) [4] – ICD-11.

Our goal is to improve the representation and traceability of rare diseases in medical terminologies, health records and systems. The ongoing work comprises several parts: (i) developing OWL model of Orphanet data and using automated reasoning to detect inconsistencies; (ii) mapping Orphanet diseases into established medical ontologies/terminologies such as Snomed CT and MeSH. The Orphanet OWL Ontology is the first ontology of this kind, and the inconsistencies we detected prompted revision of Orphanet data by the Orphanet group. Once the Orphanet data becomes publicly available, the same will hold for the corresponding OWL ontology. The lexical mappings we produced in part (ii) have revealed that only 1/3 of Orphanet rare diseases are present in Snomed CT. We plan to use these mappings in order to accomplish part (iii) of our project: revise/import missing rare diseases into Snomed CT and possibly MeSH. Cross-referencing Snomed CT with Orphanet rare diseases would provide a rich set of rare/genetic diseases for coding purposes and facilitate access to the unique Orphanet data on relations of rare diseases and clinical signs/genes and possibly enable development of sophisticated clinical systems where one can e.g. insert clinical signs of a disease and obtain a list of genes to be tested to confirm this disease.

2. ORPHANET OWL ONTOLOGY

We developed an OWL model of Orphanet data, stored in a relational database. To this end we introduced several general classes, such as Disease, ClinicalSign, Gene, InheritanceMode, AgeOfOnset, AgeOfDeath, which correspond to the main entities from the Orphanet database. Moreover, we used properties such as hasGene, hasInheritanceMode, hasAgeOfOnset etc. to relate diseases to genes, inheritance mode, and their age of onset respectively, and properties such as is_frequently_sign_of to relate clinical signs to diseases.

Modelling Orphanet data in OWL required additional background knowledge on rare and genetic diseases, as the data stored in the database has ambiguous meaning. For example, in the table relating diseases with their inheritance mode there are two rows with the following data:

Juvenile myoclonic epilepsy	Autosomal dominant
Juvenile myoclonic epilepsy	Autosomal recessive

[1] http://www.orpha.net
[2] http://www.ihtsdo.org/snomed-ct/
[3] http://www.nlm.nih.gov/mesh/
[4] http://www.who.int/classifications/icd/en/

This can be interpreted in several ways, including these: (1) each instance of Juvenile myoclonic epilepsy is inherited in both autosomal dominant and autosomal recessive way; (2) each instance of Juvenile myoclonic epilepsy is inherited in autosomal dominant or in autosomal recessive way; (3) there are instances of Juvenile myoclonic epilepsy that are inherited in autosomal dominant and those that are inherited in autosomal recessive way; (4) conjunction of (2) and (3). In this case, the intended semantics of the table rows above is (4); (1) can be eliminated since each instance of a disease can have only one inheritance mode. Thus the corresponding OWL axioms look as follows:

$$\begin{aligned} \mathsf{hasInheritanceMode} &\ \text{is} \ \text{functional} \\ \mathsf{Juvenile\_myoclonic\_epilepsy} &\sqsubseteq \forall \mathsf{hasInheritanceMode}. \\ &\quad (\mathsf{Dominant} \sqcup \mathsf{Recessive}) \end{aligned}$$

$$\begin{aligned} \mathsf{J\_m\_epilepsy\_d} &\sqsubseteq \mathsf{Juvenile\_myoclonic\_epilepsy} \\ \mathsf{J\_m\_epilepsy\_d} &\sqsubseteq \exists \mathsf{hasInheritanceMode.Dominant} \\ \mathsf{J\_m\_epilepsy\_r} &\sqsubseteq \mathsf{Juvenile\_myoclonic\_epilepsy} \\ \mathsf{J\_m\_epilepsy\_r} &\sqsubseteq \exists \mathsf{hasInheritanceMode.Recessive} \\ \mathsf{Dominant} \ &\text{and} \ \mathsf{Recessive} \ \text{are disjoint} \end{aligned}$$

Additionally we state:

$$\begin{aligned} \mathsf{Juvenile\_myoclonic\_epilepsy} &\sqsubseteq \mathsf{Disease} \\ \mathsf{Disease} &\sqsubseteq \exists \mathsf{hasInheritanceMode.InheritanceMode} \\ \mathsf{Dominant} \sqcup \mathsf{Recessive} &\sqsubseteq \mathsf{InheritanceMode} \end{aligned}$$

We modelled other properties of diseases, as well as relations between diseases and genes in a similar way. Further functional properties are e.g. hasAgeOfOnset, hasAgeOfDeath, while associatedGene is not functional, since one instance of disease can be associated with several genes.

3. ORPHANET INCONSISTENCIES

By means of DL reasoners Fact++ and Pellet we found a large number of inconsistencies in the Orphanet OWL ontology we produced. These inconsistencies correspond to cases where constraints on diseases are not propagated to their descendant diseases. In the following example, we have a different age of onset for a disease and its subcategory (note that hasAgeOfOnset is functional and ages of onset Adulthood and Adolescence are disjoint):

$$\begin{aligned} \mathsf{Hemochromatosis\_juvenile} &\sqsubseteq \mathsf{Hemochromatosis} \\ \mathsf{Hemochromatosis\_juvenile} &\sqsubseteq \exists \mathsf{hasAgeOfOnset.Adolescence} \\ \mathsf{Hemochromatosis} &\sqsubseteq \exists \mathsf{hasAgeOfOnset.Adulthood} \end{aligned}$$

Such inconsistencies show either errors in properties of diseases or in their classification. We forwarded the list of inconsistent classes together with responsible axioms to the Orphanet Group for a revision and it turned out that in a small number of cases inconsistencies were due to wrong properties and in the majority of cases, various relations between diseases are wrongly set to "subclass of". These relations include causal relation (Iron_deficiency_anemia is caused by Schistosomiasis) and, very often, "a variant of" relation (Hemochromatosis, _juvenile is a variant of Hemochromatosis, not a subclass of Hemochromatosis). In order to model "a variant of" relation in OWL, we need to introduce a common parent disease of Hemochromatosis and Hemochromatosis_juvenile (which has a variable age of onset):

$$\begin{aligned} \mathsf{Hemochromatosis\_juvenile} &\sqsubseteq \mathsf{Hemochromatosis\_general} \\ \mathsf{Hemochromatosis} &\sqsubseteq \mathsf{Hemochromatosis\_general} \end{aligned}$$

As a result of our Orphanet OWL ontology inconsistency report, Orphanet Group are revising their disease classification at the moment in order to eliminate detected inconsistencies and encode relations between diseases in a truthful way. The revised Orphanet database will hopefully lead to a consistent next version of Orphanet OWL ontology.

4. MAPPING ORPHANET

In order to establish correspondence between Orphanet rare and genetic diseases and those encoded in well established medical terminologies we are producing mappings of Orphanet diseases to Snomed CT, MeSH and MedDRA terminologies, all of which are a part of the UMLS (Unified Medical Language System)[5]. The mapping process consists of three steps: *1.* Produce **lexical mappings** – exact (corresponding to finding equivalent classes $\mathsf{Orphanet : C1} \equiv \mathsf{Target : C2}$) and partial (finding class inclusions $\mathsf{Orphanet : C1} \sqsubseteq \mathsf{Target : C2}$) by comparing disease names in Orphanet and in the target ontologies. We produced these mappings by adapting mapping results of MetaMap [6], a specialized NLM (the National Library of Medicine) tool for mapping terminologies into UMLS; *2.* Establish quality control of lexical mappings produced in Step 1 by clinical experts; *3.* Produce **structural mappings** – using the set of approved lexical mappings in Step 2 and structure of ontologies (in this case, class hierarchy) to produce further mappings of the form $\mathsf{Orphanet : C1} \sqsubseteq \mathsf{Target : C2}$.

At the moment, we have completed Step 1 (finding lexical mappings), and Step 2 (quality control of these) is taking place. Our preliminary results show that only about 1/3 of Orphanet diseases are present in Snomed CT, and significantly less in MeSH and MedDRA. This confirms our claim that rare and genetic diseases are under-represented in general medical terminologies.

5. REVISING SNOMED CT

Our future work will include revision and update of the Snomed CT data on rare and genetic diseases, using the mappings and the improved version of the Orphanet ontology. In order to revise Snomed CT's rare disease class hierarchy, we will use exact lexical mappings to import this hierarchy into Orphanet ontology, and then perform a consistency check as described in Section 3. If it is ensured that the Orphanet ontology itself is consistent, inconsistencies in the merged ontology will point to possible errors in Snomed CT classification of rare diseases. Once the Snomed CT classification of rare diseases is revised and synchronised with Orphanet, we can use mappings to import missing Orphanet rare diseases as well their hierarchy into Snomed CT.

6. ACKNOWLEDGMENTS

Many thanks to Ana Rath, Ségolène Aymé and the rest of Orphanet Group, to Alan Rector, Jeremy Rogers, Stéfan Darmoni and Tayeb Merabti for their support and advice on Orphanet, Snomed CT, MetaMap and UMLS.

[5]http://www.nlm.nih.gov/research/umls/
[6]http://mmtx.nlm.nih.gov/