



HAL
open science

OntoOrpha: an ontology to support edition and audit of rare diseases knowledge in Orphanet

Ferdinand Dhombres, Pierre-Yves Vandebussche, Ana Rath, Marc Hanauer, Annie Olry, Bruno Urbero, Rémy Choquet, Jean Charlet

► **To cite this version:**

Ferdinand Dhombres, Pierre-Yves Vandebussche, Ana Rath, Marc Hanauer, Annie Olry, et al.. OntoOrpha: an ontology to support edition and audit of rare diseases knowledge in Orphanet. 2nd International Conference on Biomedical Ontology, Jul 2011, Buffalo, United States. hal-03912146

HAL Id: hal-03912146

<https://hal.science/hal-03912146v1>

Submitted on 23 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

OntoOrpha : an ontology to support edition and audit of rare diseases knowledge in Orphanet

Ferdinand Dhombres^{1,2,3,4}, Pierre-Yves Vandenbussche^{1,2}, Ana Rath², Marc Hanauer², Annie Olry², Bruno Urbero^{2,5}, Rémy Choquet^{1,2}, and Jean Charlet^{1,2,4}

¹ INSERM UMRS 872 ÉQ.20, Paris, France

² INSERM SC11, ORPHANET, Paris, France

³ Sorbonne Universités, UPMC, Paris, France

⁴ Service de Gynécologie-Obstétrique et Centre de Diagnostic Prénatal de l'Est Parisien, Hôpital Armand Trousseau, AP-HP, Paris, France

⁵ Mondeca, Paris, France

⁶ INSERM DSI – DR Langedoc-Roussillon, Montpellier, France

⁷ AP-HP – Assistance Publique, Hôpitaux de Paris, Paris, France
{ferdinand.dhombres, jean.charlet}@inserm.fr

Abstract. ORPHANET is the reference information portal on rare diseases and orphan drugs, for healthcare professionals and for general audience. After ten years of evolution, current Orphanet tools cannot support efficiently the edition, update and data sharing processes of a constantly growing rare diseases knowledge. In order to improve the edition workflow, we conduct research to build and use a rare diseases knowledge base in an *Ontology-based architecture* that complies with the W3C standards of the semantic web : OWL, RDF, SPARQL and SKOS. Our ontology design approach is based on both domain expertise (in rare diseases and in knowledge engineering) and knowledge extraction from our relational database. The current version of OntoOrpha comprises over 11,000 classes and 190,000 annotations organized under a *Rare Diseases Core Ontology*.

In comparison with current Orphanet edition tools, our preliminary experiments are consistent with : (1) better visualization of the knowledge base (2) improved classification edition procedures (3) improved annotation edition procedures (4) valid semantic validation procedures.

1 Background

ORPHANET is the reference information portal on rare diseases and orphan drugs, for healthcare professionals and for general audience. This unit is led by a European consortium of around 40 countries, coordinated by the French team which is responsible for the infrastructure of Orphanet, management tools, quality control, rare disease inventory, classifications and production of the encyclopaedia. After ten years of evolution, current Orphanet tools cannot support efficiently the edition, update and data sharing processes of a constantly growing rare diseases knowledge (6000 rare diseases with annotations and more than a hundred of overlapping classifications).

2 Methods

In order to improve the edition workflow, we conduct research to build and use a rare diseases knowledge base in an *Ontology-based architecture*. This architecture complies with the W3C standards of the semantic web : OWL, RDF, SPARQL and SKOS.

Our ontology design methodology is multidisciplinary as it involves both domain expertise in rare diseases and in knowledge engineering. Our methodology also involves automatic knowledge extraction from Orphanet relational database. The current version of OntoOrpha comprises over 11,000 classes, 72,000 sub-ClassOf axioms (with a high level of polyparentality) and 190,000 annotations, organized under a *Rare Diseases Core Ontology* (fig. 1).

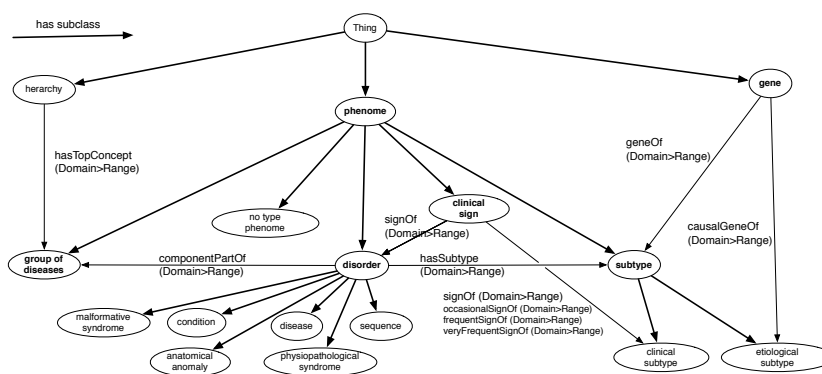


Fig. 1. Rare Diseases Core Ontology (view of OntoOrpha, version 2011-06-08).

This core ontology was designed as meta-model for the domain ; this level of model was mandatory to provide appropriate representation of the whole disease inventory extracted from the database as classes, and to represent the classifications as classes as well. Domain and range of relationships between classifications, disorders (disease, malformative syndrome,...), groups of disorders (by anatomical system, by physiopathological mechanism,...), subtypes (clinical subtypes, etiological subtypes), clinical signs and genes are therefore represented in the core ontology. Moreover, we take into account that this meta-model should provide all the primitive for validation rules description.

3 Results

Beside current architecture, a new experimental architecture is provided by the project to support edition workflow (fig. 2). At the first step of the workflow extraction from the database, building of the ontology and uploading it in a triplestore are performed. This extract-transform-load (ETL) step is automated. The generation process of the ontology is threefold :

1. generation of the header of the final RDF file (including ontology metadata, classes and object properties of the core ontology),
2. generation of the body parts by extraction/transformation of the data (URI construction, skos:prefLabels/altLabels/definition including the management of 6 languages, owl:Restriction,...)
3. merging those parts to produce an XML/RDF file that will be upload in the triple store.

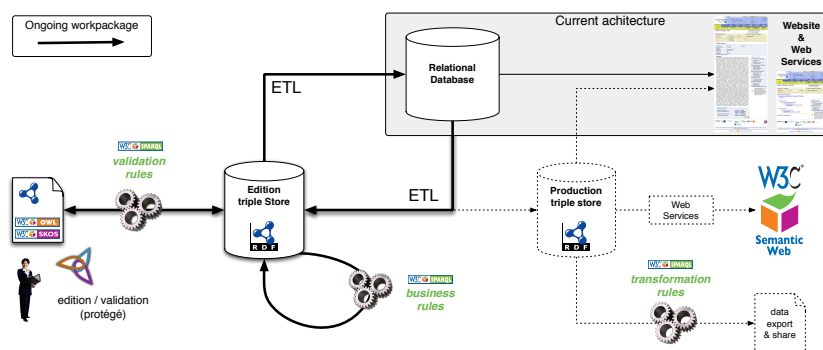


Fig. 2. Rare Diseases Knowledge edition workflow (research architecture).

The second step is consistent with expert edition of the ontology (Classes, ObjectProperty, Annotations) and rule-based procedures implemented with iterative SPARQL queries on the triplestore (for validation, classification generation, audit report generation). The final step is the relational data base update with an ETL process from the triplestore.

In comparison with current Orphanet edition tools, our preliminary experiments are consistent with :

(1) better visualization of the knowledge base : a global view of the hierarchies is provided in the ontology editor during the edition process (*Class hierarchy view, Existential tree view, Outline tree view and OntoGraf view* in PROTÉGÉ),

(2) improved classification edition procedures : the experts edit the ontology itself and the *rare diseases classifications are automatically generated*, using stable SPARQL queries,

(3) improved annotation edition procedures : we use a lexicalisation plugin for PROTÉGÉ developed in our unit that is SKOS compliant and supportive for multilingual edition of labels, synonyms and abstracts,

(4) semantic validation procedures : both in the ontology editor (a priori validation by the build-in reasoner) and in the triplestore (validation and audit by reasoner and SPARQL queries).