ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

OntoOrpha: An Ontology to Support the Editing and Audit of Knowledge of Rare Diseases in ORPHANET

Ferdinand Dhombres[1,2,3,4], Pierre-Yves Vandenbussche[1,5], Ana Rath[2], Marc Hanauer[2], Annie Olry[2], Bruno Urbero[2,6], Rémy Choquet[1,2], Jean Charlet[1,2,3,7]

[1]INSERM UMRS 872 éq.20, Paris, France
[2]INSERM SC11, ORPHANET, Paris, France
[3]Sorbonne Universitiés, 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 – Langedoc-Roussillon, Montpellier, France
[7]AP-HP – Assistance Publique, Hôpitaux de Paris, Paris, France

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 demanded by a constantly growing rare diseases knowledge. In order to improve the editing workflow, we are conducting 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 the current ORPHANET editing tools, our preliminary experiments are consistent with: (1) better visualization of the knowledge base (2) improved classification editing procedures (3) improved annotation editing 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 the general public [3]. ORPHANET is led by a large European consortium of around 40 countries, coordinated by the French INSERM team which is responsible for the infrastructure of ORPHANET, management tools, quality control, rare diseases inventory, classifications and production of the encyclopedia. After ten years of evolution, current ORPHANET tools are limited in efficiently supporting the editing, update and data sharing processes of a constantly growing rare diseases knowledge (6000 rare diseases with annotations and more than one hundred overlapping classifications).

2 Methods

In order to improve the editing workflow, we are conducting 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 [1], RDF [4], SPARQL [9] and SKOS [7].

Our ontology design methodology is multidisciplinary as it involves both domains of expertise: rare diseases and knowledge engineering. Our methodology also involves automatic knowledge extraction from the ORPHANET relational database [5,8]. The current version of OntoOrpha comprises over 11,000 classes and 190,000 annotations, organized under a *Rare Diseases Core Ontology* (fig. 1).

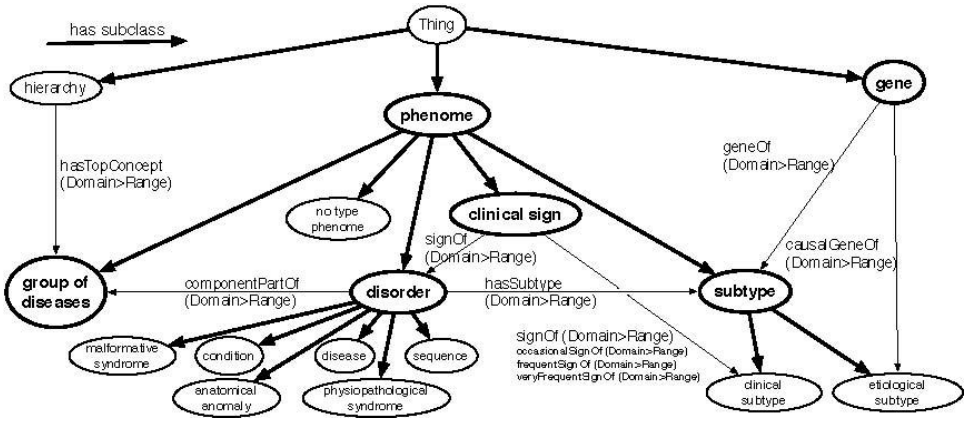

Figure1. Rare Diseases Core Ontology (view of OntoOrpha, version 2011-06-08).

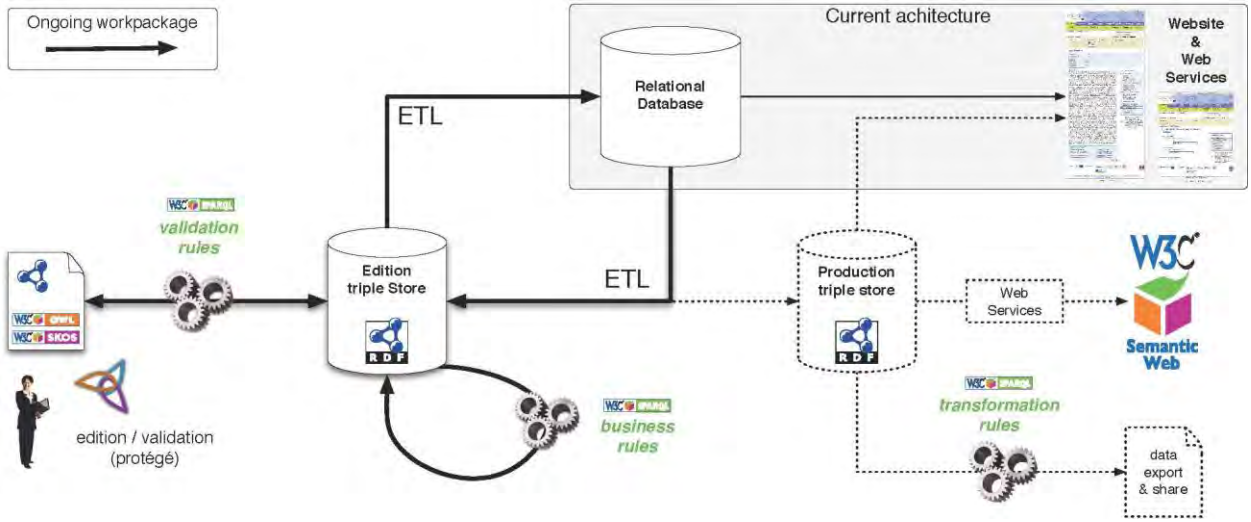

Figure 2. Rare Diseases Knowledge editing workflow (research architecture).

This core ontology was designed as a meta-model for the domain; this abstraction level was mandatory to provide the 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. In addition to this, we take into account that this metamodel should provide all the primitives needed for the description of validation rules.

3 Results

Besides the current architecture, a new experimental architecture is provided by the project to support editing workflow (fig. 2). At the first step of the workflow the extraction from the database, the building of the ontology and the uploading to a triplestore (semantic database) 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 uploaded to the RDF triple store

The second step is editing the ontology by the expert (Classes, Object-Property, 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 database update with an ETL process from the triplestore.

In comparison with current ORPHANET editing tools, our preliminary experiments are consistent with:

- a 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 editing procedures: the experts edit the ontology itself and the rare diseases classifications are automatically generated, using stable SPARQL queries,
- improved annotation editing procedures: we use a lexicalization plugin for PROTÉGÉ developed in our unit [6] that is SKOS compliant and supportive for multilingual editing of labels, synonyms and abstracts,
- 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).

Finally, the core ontology allows us to globally review and reorganize the ORPHANET rare disease knowledge. It provides the necessary coherent top-structure of the knowledge managed into the knowledge base (diseases, classifications, genes, ...). ORPHANET core ontology guarantees a consistent evolution of the ontology going forward.

References

1. OWL 2 web ontology language. W3C recommendation. W3C OWL working group. (2009), http://www.w3.org/TR/2009/REC-owl2-overview-20091027/
2. Protégé 4 - Open Source Ontology Editor. Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine (2011), http://protege.stanford. edu/
3. Aymé, S.: Orphanet: serveur d'information sur les maladies rares et les médicaments orphelins, INSERM SC11. http://www.orpha. net/ (2002).
4. Beckett, D., McBride, B.: RDF/XML syntax specification (Revised). W3C recommendation. (2004), http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/
5. Krivine, S., Nobécourt, J., Soualmia, L., Cerbah, F., Duclos, C.: Construction automatique d'ontologie à partir de bases de données relationnelles: application au médicament dans le domaine de la pharmacovigilance. In: Actes des 20e Journées Francophones d'Ingénierie des Connaissances. pp. 73–84. Fabien Gandon (Ed.), Hammamet, Tunisie (2009)
6. Mazuel, L.: Archonte plugin for protégé. (2010), http://pertomed.spim.jussieu. fr/~lma/doe/fr.spim.archonte.jar
7. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation. (2009), http://www.w3.org/ TR/2009/REC-skos-reference-20090818/
8. O'Connor, M.J., Das, A.: Semantic reasoning with XML-based biomedical information models. Studies in Health Technology and Informatics 160(Pt 2), 986–990 (2010)
9. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C recommendation. (2008), http://www.w3.org/TR/2008/ REC-rdf-sparql-query-20080115/