# OntoOrpha : an ontology to support the editing and audit of rare diseases knowledge in Orphanet

Ferdinand DHOMBRES[1,2], Pierre-Yves VANDENBUSSCHE[1], Ana RATH[2], Marc HANAUER[2], Annie OLRY[2], Bruno URBERO[2], Rémy CHOQUET[2] & Jean CHARLET[1]

(1) INSERM U872 EQ20 - Knowledge Engineering for Healthcare & UPMC, Paris, France (2) INSERM SC11 - Orphanet, Paris, France Contact : ferdinand.dhombres@inserm.fr

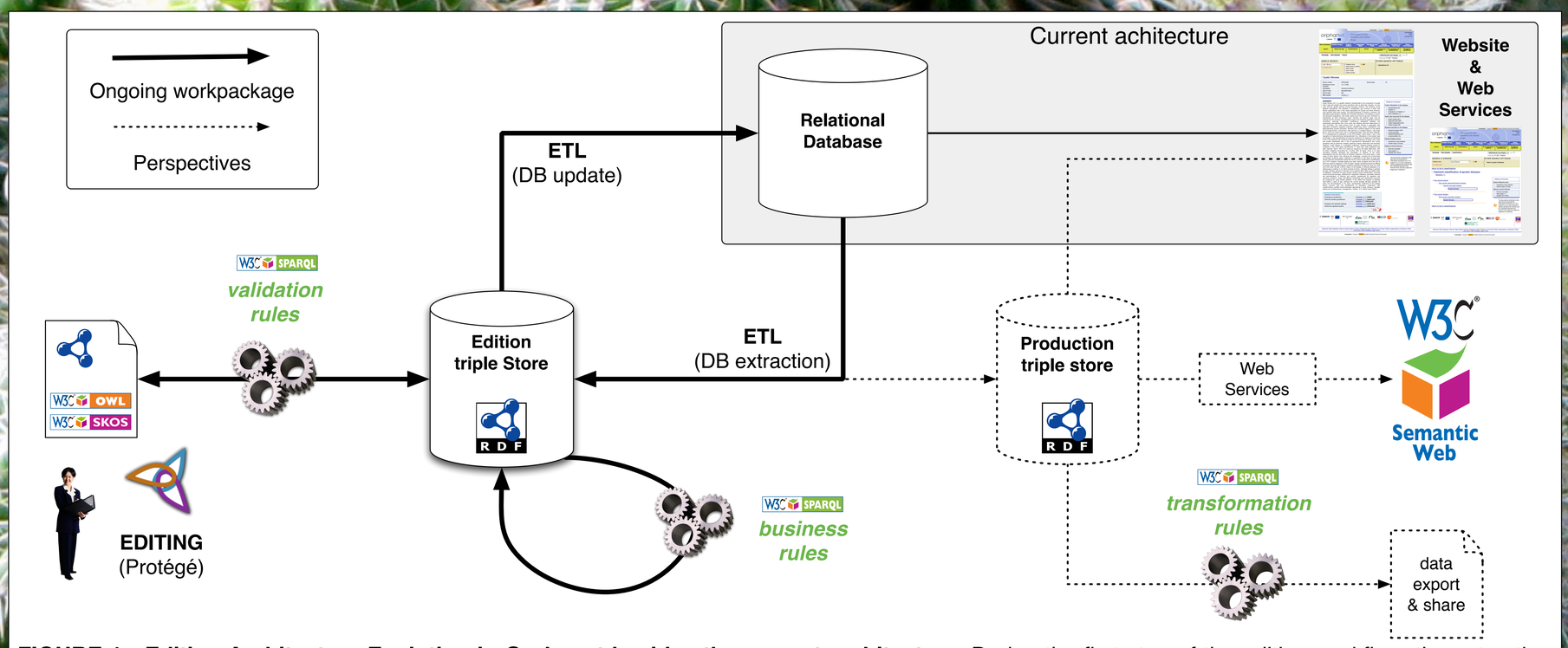


**FIGURE 1** - **Editing Architecture Evolution in Orphanet besides the current architecture.** During the first step of the editing workflow, the extraction from the database, the building of the ontology and the uploading to a triplestore (semantic database) are performed. This extract-transform-load (ETL) step is automated. The second step is the editing of the ontology by the expert in association with rule-based validation procedures. The final step is the relational database update with an ETL process from the triplestore.

## BACKGROUND

Orphanet is the reference information portal on rare diseases and orphan drugs for all audience (for both healthcare professionals and general public). 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, the management tools, the quality control, the rare diseases inventory, the classifications and the edition 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).

* Aymé, S. Orphanet : The portal for rare diseases and orphan drugs, INSERM SC11. website : http://www.orpha.net/

## 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 (fig. 1) 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 (fig. 2). 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 diseases 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, we take into account that this meta-model should provide all the primitives needed for the description of validation rules.

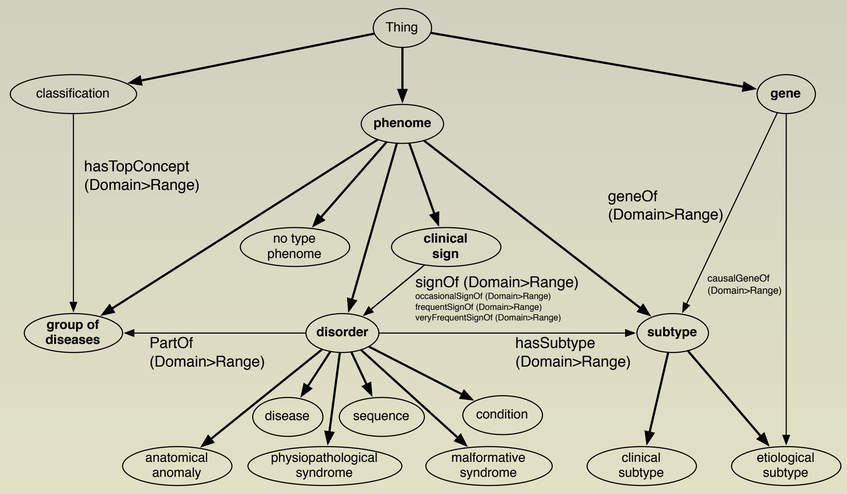


**FIGURE 2** - Rare Diseases Core Ontology for Orphanet Editing (2nd version)
(bold arrows represent hasSubclass properties)

## RARE DISEASES CLASSIFICATIONS CURATION

In comparison with current Orphanet tools, a better visualization of the knowledge base is possible in the new architecture : a global view of the hierarchies and the relations is provided in the ontology editor during the editing process (fig. 3). Moreover, simple features like "drag and drop" and "search with autocompletion of query" are now available. In this architecture,

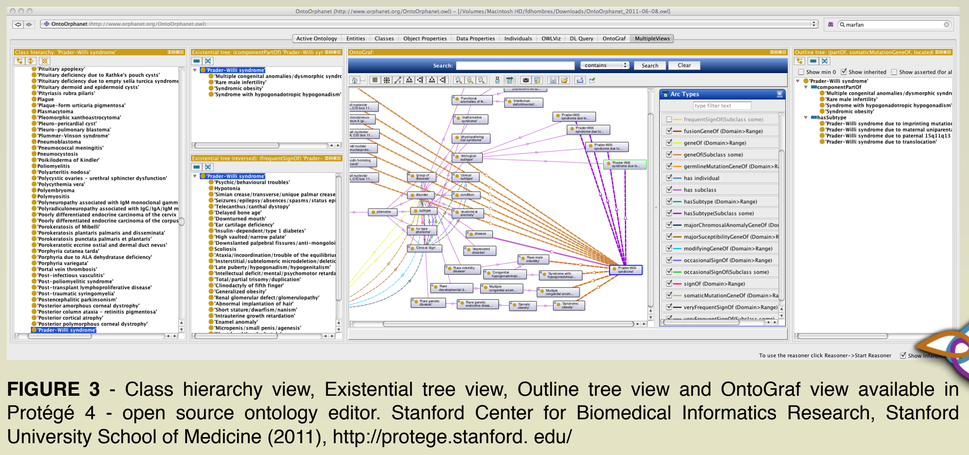

**FIGURE 3** - Class hierarchy view, Existential tree view, Outline tree view and OntoGraf view available in Protégé 4 - open source ontology editor. Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine (2011), http://protege.stanford. edu/

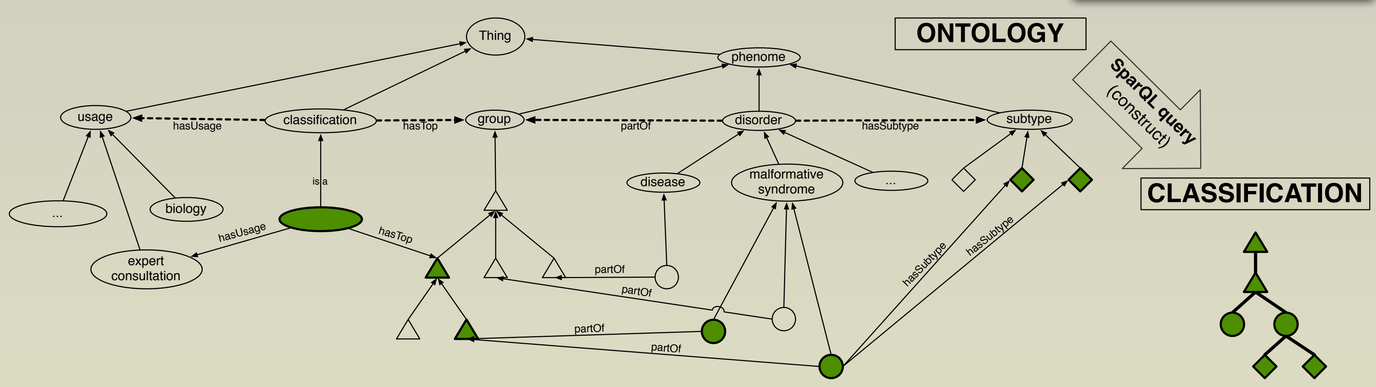


**FIGURE 4** - Automated Classification Generation Principles. SparQL queries (using the *CONSTRUCT* method) are used to extract a specific classification from the ontology stored in an RDF repository.

the experts edit the ontology. They can modify the subsumption hierarchies (for groups and subtypes) and the restrictions on properties (partOf and hasSubtype). This does improve classification production procedures : the experts only edit the ontology and rare diseases classifications are automatically generated, using stable SparQL queries. In current architecture, the editing of each classification is manual, must be done for each classification and is performed line by line in tabular files (each "broader than-narrower than" (BTNT) relations is defined on one line).

## ANNOTATIONS EDITING

Improved annotation editing procedures are based on a lexicalization plugin for Protégé developed in our unit (fig. 5). This plugin is Skos compliant and supportive for multilingual editing of labels (prefLabel), synonyms (altLabel) and encyclopedic definition (definition).

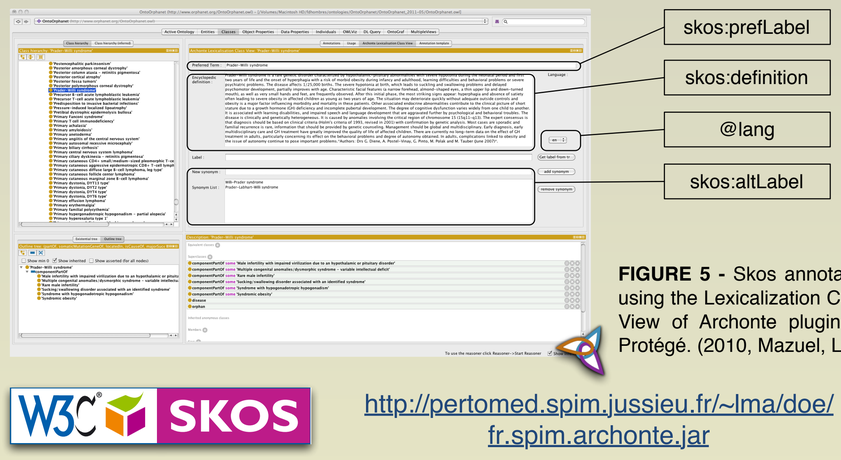


**FIGURE 5 -** Skos annotation using the Lexicalization Class View of Archonte plugin for Protégé. (2010, Mazuel, L.)

http://pertomed.spim.jussieu.fr/~lma/doe/fr.spim.archonte.jar

## KNOWLEDGE BASE VALIDATION & AUDIT

Semantic validation procedures are performed at different levels of the architecture :

1) in the ontology editor, *a priori* validation by one of the build-in reasoner (HermiT*)
2) in the triplestore validation by an external reasoner (Euler**) and by SparQL queries.

Rule-based procedures implemented with iterative SparQL queries on the triplestore are used for audit. Automated report generation for the experts provides lists of Classes that do not comply with a given set of business rules like "all the disorders are part of a group", "all the genetic disorders have an inheritance mode" or "all the phenomes have a label in english".

* http://hermit-reasoner.com/
** http://www.agfa.com/w3c/euler/

## PERSPECTIVES

The core ontology permits to globally review and reorganize 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 and allow dynamic rare diseases knowledge sharing perspectives :

- public access in Bioportal*
- publication of mapping with other resources
- SparQL Endpoint for rare diseases, ...

* http://purl.bioontology.org/ontology/OntoOrpha