Rare diseases knowledge curation in an ontology-based architecture in Orphanet

Ferdinand Dhombres, Pierre-Yves Vandenbussche, Rémy Choquet, Jos de Roo, Ana Rath, Annie Olry, Marc Hanauer, Bruno Urbero, Ségolène Aymé, Jean Charlet

From terminologies to ontologies and back

Rare diseases knowledge curation in an ontology-based architecture in Orphanet

Ferdinand Dhombres [a,b,c,d,e,*], Pierre-Yves Vandenbussche [a,c,f], Rémy Choquet [a,c], Joss de Roo [g], Ana Rath [c], Marc Hanauer [c], Annie Olry [c], Bruno Urbero [c], Aymé Ségolène [c] and Jean Charlet [a,h]

[a] *INSERM UMRS 872 éq.20, Knowledge engineering for Healthcare, Paris, France.*
[b] *INSERM SC11, Orphanet, Rare Diseases Center, Hôpital Broussais, Paris, France.*
[c] *Pierre et Marie Curie University, Paris, France.*
[d] *Paris Descartes Medical School, Paris, France.*
[e] *Obstetrical and Gynecological Unit, Hôpitaux Saint-Antoine et A. Trousseau, AP-HP, Paris, France.*
[f] *MONDECA, Paris, France.*
[g] *AGFA Healthcare, Sint-Martens-Latem, Belgium.*
[h] *AP-HP – Assistance Publique Hôpitaux de Paris, Paris, France.*

Abstract.
ORPHANET is a reference information portal on rare diseases and orphan drugs for healthcare professionals and general audiences. After ten years of evolution, the current ORPHANET tools and knowledge representation cannot represent appropriately the constantly evolving knowledge on rare diseases. ORPHANET knowledge base has evolved from a simple thesaurus to a multi-classification terminology over the years, without refactoring the knowledge organization from the top (using a meta-model or/and an ontology). We propose in this paper to review the knowledge organization of ORPHANET by introducing a core ontology for rare diseases that has the specificity to also model classifications. 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. Using a specific knowledge cycle, we propose new edition, validation and sharing processes for rare diseases knowledge in ORPHANET. We show in this paper that ontologies are designed to manage the generation of multi-classifications into complex knowledge organization systems (KOS). We also demonstrate that the introduction of formal knowledge representation systems (OWL) improved ORPHANET knowledge base quality. This experiment highlights a continuity in the use of different knowledge organization systems. Nevertheless, the complex knowledge curation of this domain involves a formalization that can be appropriately supported only by the use of rules and an ontology.

Keywords: ontology, terminology, knowledge organisation system, knowledge base curation, rare diseases, OntoOrpha

1. Introduction

1.1. Background

ORPHANET is the reference information portal on rare diseases and orphan drugs[1] funded jointly by the European Commission, INSERM (the French National Institute of Health and Medical Research) and the French Directorate General for Health. It provides healthcare professionals and general public with information on rare diseases[2] in order to improve rare diseases diagnosis and care. In this respect, a multilingual information portal was created, comprising classifications of rare diseases, an online encyclopedia as well as registries of specialised clinics, medical laboratories, ongoing research projects and patient or-

*Corresponding author: Ferdinand Dhombres, Inserm SC11 Orphanet, Plateforme Maladies Rares de l'hôpital Broussais, Bât. Maurice Raynaud, 96 rue Didot, 75014, Paris FRANCE. E-mail: ferdinand.dhombres@inserm.fr.

[1] ORPHANET Website : `http://www.orpha.net`

[2] In Europe, the admitted prevalence threshold specifying a disease as rare is one affected person out of 2000.

ganisations (Aymé, 2002). In 2010, on average, 10,000 visitors were registered per day, 1/3 of them were doctors, 1/5 were other healthcare professionals and 1/3 were patients[3].

After ten years of evolution, ORPHANET knowledge base became too complex for scientist managers to support efficiently the curation with the current tools. The domain of rare diseases is wide (nearly 6,000 rare diseases are referenced by ORPHANET) and rare diseases knowledge was quickly growing. Visualization and validation tools used for knowledge editing and curation came down to spreadsheets allowing views from the database to be modified. Scientific editors could not have a global view of structured knowledge; access to the hierarchies visualization was not provided during the editing process. This latter point was a crucial issue as experts maintained more than one hundred overlapping poly-parental classifications. Moreover, ORPHANET had growing difficulties in answering more and more data extraction requests. Despite 12 standardised data sets, there were between 30 and 50 requests of different custom-built data sets per year (for research projects, institutions and industry). Several communication documents (ORPHANET reports series, directories..) had to be regularly updated according to the database content and adapted to the 36 ORPHANET partner countries. Data extraction was linked to data representation in the database which, still today, prevents the use of models and tools more appropriate to knowledge representation. The issue were thus twofold, involving both the tools and the methods of knowledge representation.

1.2. Objectives

Our general objective as part of the ORPHAONTO project was the integration of ORPHANET knowledge on rare diseases, contained in a relational database (RDB), within an architecture allowing improved knowledge edition and use. Our main constraint was the respect of the current system architecture that supported the production of ORPHANET Website and Web services. Our hypothesis was that an ontology-based platform would be able to overcome the current limits of RDB representation, content edition and validation. The use of such formalized content would also support Semantic Web data publishing. Hence, this architecture must comply with the W3C standards of the Semantic Web (OWL, RDF, SPARQL and SKOS) in order to ensure a long term support of the knowledge representation formalisms.

From this perspective, we developed a specific life cycle of domain knowledge (fig. 1) ; the ontology was built from a set of terminologies stored in ORPHANET RDB. Different uses were based on this ontology such as new terminologies edition, validation and generation. These generated terminologies can be shared or can update the current terminologies in the RDB.

2. State of the Art

2.1. Knowledge organisation systems

Knowledge representation can be supported by a wide range of structuring systems : controlled list, classification, thesaurus, terminology, ontology, etc.. We remind that we preferentially use the term Knowledge Organisation System (KOS) to designate all types of frame of reference previously stated (Hodge, 2000; Binding and Tudhope, 2006). This pluralism is explained by different characteristics of these KOSs linked to their use. In his work, Bodenreider (2008) distinguished three main KOS use categories : (i) knowledge management, (ii) semantic interoperability and (iii) decision support and reasoning. The practical use of KOSs often put at stake several of these themes to reach a more complex and complete goal.

Knowledge management One of the main KOS functions is to gather a domain vocabulary, *i.e.* a list of the wording of entities described within the sysem of reference. This terminological part of KOSs is an essential component for the automated treatment of languages (Bodenreider, 2006) and for tasks linked to knowledge management such as annotation, indexing of resources or access to information and research. Linguistic expressivity will depend on primitives defined in KOSs such as lan-

[3] All the statistics about the portal activity on the last 12 months are available on `http://www.orpha.net/stat/orphanet/`

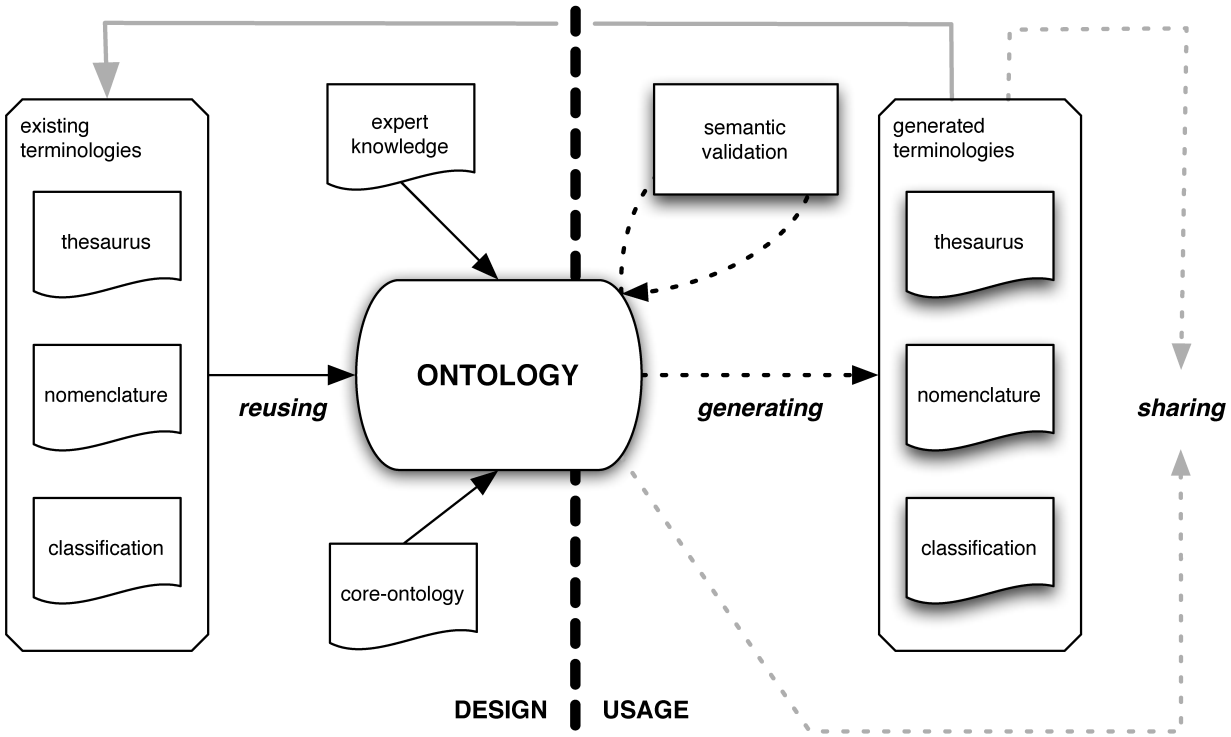


Fig. 1. Knowledge cycle in ORPHANET.

guage management for term, summary, synonymy, metonymy and translation relations. Annotation consists in affixing elements of a controlled vocabulary on a document. These annotations used for indexing helped for information research: for example, the annotation and indexing of a biomedical article with MeSH (Medical Subject Headings). This process applied to healthcare documents is often named *coding* (Giannangelo, 2006). The main function of document indexing is the information research support. By conducting a textual research, it is possible to access to a document thanks to linguistic entries and their variations (synonyms, abbreviations, etc..) contained in the KOS used for the annotation. Researches could be extended to more specific concepts than the one originally found thanks to the KOS hierarchy (Greenberg, 2001). For example, the search query "Upper limb anomalies" will provide results annotated by the concept with the term "Upper limb anomalies", but also by the concept with the term "Anomalies of hands" or the concept with the term "Radio-ulnar synostosis" (considering that the first concept is described as more general than the two others in the KOS).

Interoperability and data integration KOSs are also used to centralise, federate and share potentially heterogeneous data coming from several sources. Let us detail two different approaches of the use of these systems of reference: semantic interoperability and data integration. The first one aimed at sharing inter-system information on the basis of a common interpretation context: KOS. The second one aimed at gathering information in a homogeneous group to make data analysis, for example. KOSs provide a community of users with a shared conceptualisation of a specific part of the world in order to facilitate the efficient communication of a complex knowledge. This need to exchange information and to share a conceptualisation conveys the notion of interoperability that Miller (2000) defined by: "process of ensuring that the systems, procedures and culture of an organisation are managed in such a way as to maximise opportunities for exchange and re-use of information, whether internally or externally". Interoperability in Knowledge Engineering can be seen as the ability of two persons or information systems to communicate information (syntactic interoperability) and to share a common context to interpret this piece of information with no ambiguity and in a similar way (semantic interoperability) (Degoulet et al., 1997).

Decision support and reasoning From the simplest terminologies allowing us to reason (whose structure is limited to a subsumption hierarchy, for example) to ontologies rich in formal descriptions, KOSs offer a representation of a domain knowledge with a formal expression which can be interpreted by machines. In addition to classic data search tools, the formalization of information improves decision support. Thanks to explicit type conversion and KOS heuristics reasoning, the system displays correlations or alerts that help with decision making. In Medicine, we can quote the example of the drug interaction detection for drug monitoring (Amardeilh et al., 2009). From formal axioms of a domain ontology, some information can be automatically deduced : for example, an ontology for emergency Medicine allows to deduce that a patient affected by appendicitis is likely to contract peritonitis. Reasoning engines also validate whether information is in compliance with the rules clearly described in the system of reference. For example, a system can give us information about the applicability of a law article to individuals according to its proprieties and defined constraints.

ORPHANET set of terminologies is effective for the first and the second point. It is used to regulate a common vocabulary between two different healthcare actors, thus making interoperability easier. On the other hand, its semantic formalization mostly remains implicit: the relationship between two concepts either means "is a variant of the disease" or "belongs to the group of disease" or else "affects the system". This representation is efficient for information research, mental route from the system to the disease, but it does not allow to have logical reasoning. That is why we want to formalize, in this project, this thesaurus semantic within an ontology. This ontology opens to very new uses: reasoning, formal validation of business knowledge and monitoring of knowledge integrity.

2.2. Ontology design from terminologies and databases

The construction of ontologies from information system databases was mainly studied from two approaches: the re-use of data from XML databases in the form of ontologies and the re-use of relational database management system tables (RDBMS) to build ontologies. With the first approach, we found several works in the biomedical field (for example O'Connor and Das 2010) which developed connectors to transform XML files into OWL files in order to perform reasoning. With the second approach, we found works aiming at launching directly a relational DBMS and specifying name class and request patterns. A tool such as RDBToOnto (Krivine et al., 2009; Cerbah, 2008) offered the implementation of developed transformation methods or allowed the use of a new method acquired by specializing a converter which was already integrated into the platform.

Another tool, D2RQ, influenced by the Semantic Web offered an integrated environment in order to access to relational data via standard APIs (Application Programming Interface) such as JENA or via the creation of SPARQL endpoints (Bizer C., 2007). Since a couple of years, a group coming from W3C, the *RDB2RDF Incubator Group* (2009), has been specifically studying this matter and is standardizing a language[4] for relational data transformation into RDF, called the *RDB2RDF Mapping Language* (R2RML).

Given the complexity of our task (needs for natural language processing (NLP), reports generation, etc..) and the heterogeneity of our bases, none of the methods were satisfactory and each and every one of them would have required too many re-uses. So we chose an open data management and integration environment for the production of ontology. Regarding the re-use of other sources as developed previously (Dhombres et al., 2010a,b) the choice was quite easy to take insofar as our working base was a current thesaurus that we wanted to reorganize at the same time as the knowledge bases which were related to it. Initially, we had all the required material for our project

[4] http://www.w3.org/TR/r2rml/

2.3. Terminology generation from ontology

SPARQL language is a query language (used in validation process) for knowledge expressed in RDF format which also allows the generation of a new knowledge graph. We chose to use this language because it was widely adopted by the Semantic Web community, its implementations were available and its functionalities were rich, especially the non-monotonic function *OPTIONAL* and the ability to create a new graph by model transformation thanks to *CONSTRUCT* operator (Polleres et al., 2007).

3. Material

3.1. Orphanet relational database overview

As part of this work, we used ORPHANET databases. ORPHANET current database management system was Sybase (v15.5 for Solaris 10). A simplified schema (comprising data corresponding to diseases in the strict sense of the word only) of the tables of version *4.1.0* of the database is represented fig. 2. This data model was built in order to allow information on rare diseases circulation via ORPHANET Website. We worked with a gross extraction of tables from fig. 2 in the exploration phase, then with a copy of the complete database, weekly updated, on a server dedicated to our research project.

3.2. Rare diseases normenclature

All the 5,781 rare diseases referenced by ORPHANET were in the table TbPat, in tuples identifiable from the column Typ by the value “Pat”. Inner joins between the tables TlMllPat and TlPatSgn allowed to have respectively epidemiological data (for example, prevalence, age of onset, etc..) and signs related to a single disease identification (disease ID). Inner join with the table TbLbl allow to have the main labels and the synonyms of diseases in the form of tuples [disease_ID–label–language] and [disease_ID–synonym–language], in all the languages of the database (English, French, German, Spanish, Italian and Portuguese).

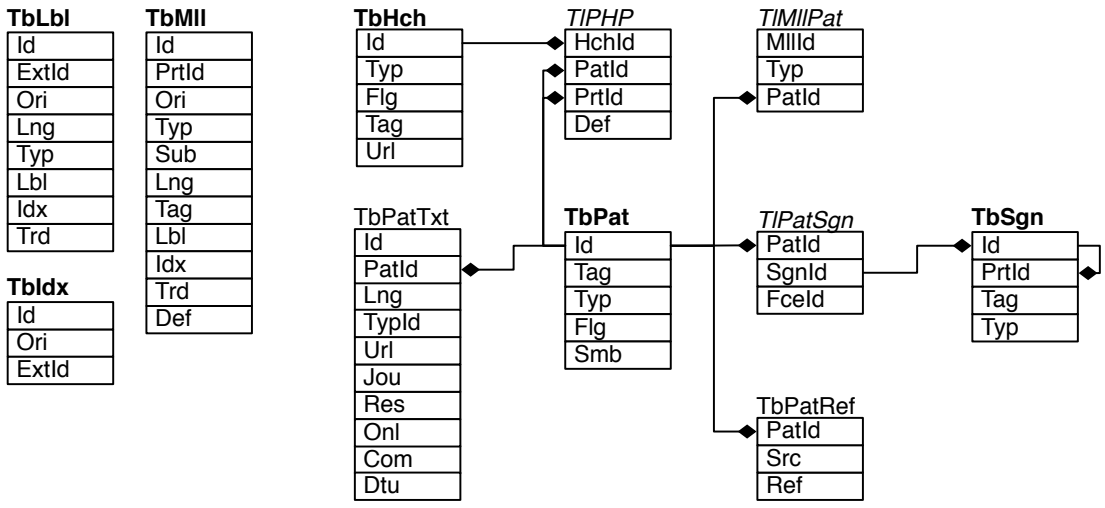


Fig. 2. Orphanet relational database schema (v4.1) : partial view of the tables [a] involved in the project.

[a]Details of the tables: **TbPat**: diseases and genes table, **TbHch**: hierarchies table, **TbSgn**: clinical signs table, **TbPatTxt**: texts of diseases (abstracts and long texts) table, **TbLbl**: labels table, **TbMll**: multilingual labels and epidemiological data table, **TbIdx**: index table, **TlPHP**: hierarchical relashionships between diseases table, **TlPatSgn**: relashionships between diseases and signs table, **TlPatRef**: relashionships between diseases and external references table.

3.3. Rare diseases classifications

There were 106 distinct hierarchies in the database. From a semantic point of view, hierarchical links were *broader than/narrower than* (BTNT). 32 classifications gathered all the rare diseases in ORPHANET (for example, rare genetic diseases classification, congenital abnormalities classification, or else rare eye abnormalities and rare skin abnormalities classifications). The structure of these subsumption hierarchies, with a significant contingent of polyparentality, is stored in the table TlPHP in the form of tuples [hierarchy_ID-disease_ID-parent disease_ID]. For a given hierarchy, the arborescence was built in a recursive way from these data. The other 74 classifications of diseases represented either views specific to an online service[5] in rare diseases domain, or classifications that were not produced by ORPHANET but resulted from healthcare literature and had an interest in rare diseases, or else a non-rare diseases classification required for the representation of some links between diseases. There were also hierarchies corresponding to internal research projects.

3.4. Clinical signs thesaurus

The 1,360 clinical signs (or groups of signs) present in the database were in the table TbSgn. Thanks to a set of inner joins, the disease identification and the identification of the frequency of the sign in the disease were extracted as tuple [sign_ID–disease_ID–frequency_ID]. There were three types of frequency linking a sign to a disease (occasional, frequent, very frequent), beginnings of three subtypes of semantic relation "sign of" (*signOf*).

3.5. Genes nomenclature

All the genes (2,415) were in the table TbPat, in the tuples identifiable from the column Typ by the value "Gen". Like the signs, the building of the pairs [gene_ID–disease_ID] from inner joins allow to formalize the relation between gene and disease: "is a gene of" (*geneOf*).

3.6. External references

Genes and diseases in ORPHANET database were annotated from reference terminologies of the domain. These external references were stored in a specific table (TbPatRef). All these annotations were produced by ORPHANET experts or during cross-validation initiatives involving a validation by an expert on the occasion of cooperation with several organizations. Diseases were annotated by ICD-10 codes (International Classification of Diseases, 10th ed., WHO), by OMIM numbers (Online Mendelian Inheritance in Man, Johns Hopkins University[6]) and by queries on PubMed website. External references based on other terminologies are being validated by ORPHANET experts until November 2011 and are not included in this working material yet (MESH terms, SNOMED-CT code, UMLS[7]). Genes were annotated by corresponding identifications in GENEATLAS, UniProtKB/Swiss-Prot, HGNC as well as by OMIM numbers.

4. Methods

Every steps of our methods were drawn up by a working group comprising domain experts (MDs and biologists), documentalists, computer scientists and knowledge engineers. Within this transdisciplinary group, reconsidering working editorial methods to get rid of the initial operational constraints (in particular the use of spreadsheets) was a key step in our approach. Choices of ontological modeling evolved during the project, leading to knowledge rewording by experts themselves.

[5] These classifications are called *functional classifications* ; we will try to derive them from the 32 first, using business and applicative rules. These rules formalization is part of the following steps of the project.

[6] http://www.ncbi.nlm.nih.gov/omim

[7] Medical Subject Headings, Systematized Nomenclature of Medicine – Clinical Terms, Unified Medical Language System.

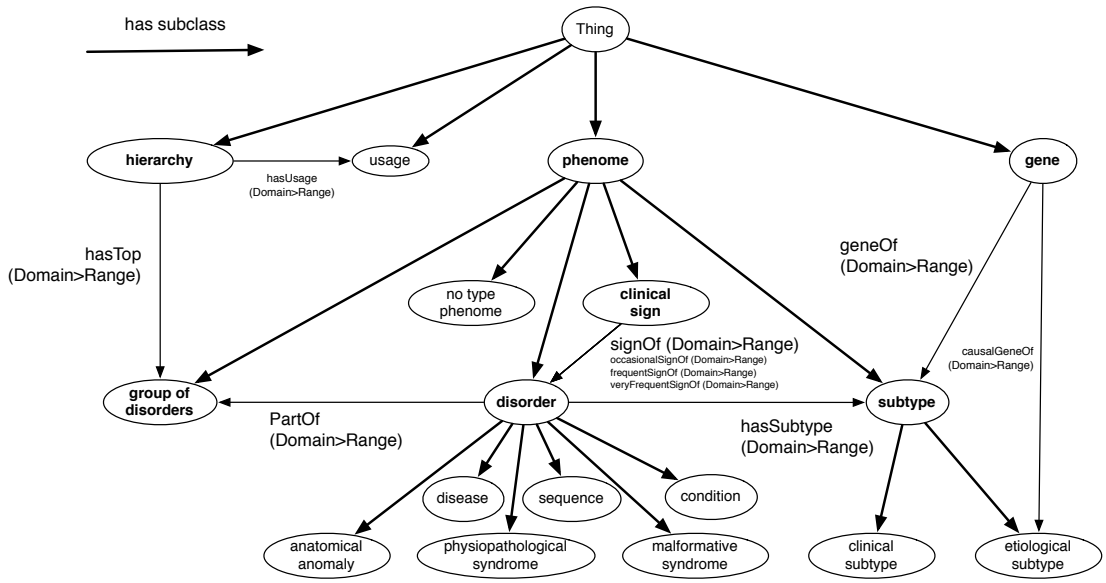


Fig. 3. Rare Diseases Core Ontology (partial view of OntoOrpha, version 2011-09-19).

4.1. Designing OntoOrpha, the Orphanet Ontology of Rare Diseases

4.1.1. The need for a core ontology of rare diseases

The relational model did not match the implicit knowledge model from the domain experts. So we chose to define the concepts of our ontology (for example, the concepts of disease, sign and gene) in a first version of the model. The previously identified relations (signOf and geneOf) allowed us to define the main proprieties of these classes. All the hierarchical relations of classification, transformed in *is-a* relations, were used to build the subsumption tree with a very high polyparentality rate.

This first model supported the nomenclature exploration and edition while offering a visualization of concepts subclasses and superclasses. Yet, this model was not kept because of the limits of specific representation of the various classifications. Moreover, the use of validation rules was not possible. And the strong semantic of the *is-a* relation was not adapted to these hierarchies content, made up of groups of diseases, diseases and specific subtypes of these diseases. Indeed, proprieties inheritance led to inconsistencies.

Confronted with these limits, we decided to build a meta-model of the domain allowing to represent diseases concepts (with their genes and signs) and classifications. This model, in the form of a core ontology (fig. 3) was designed from extra expert knowledge such as phenome types definition and classification specifications. Concepts related to diseases were organized under the general *phenome* concept from clinical genetics. Concepts of *clinical sign*, *disorder*, *group of disorders* and *subtype* were types of *phenomes*. The *disorder* concept matched all the types of rare affectations listed by ORPHANET (for example, malformative syndromes, associations, sequences, etc..). The *group of disorders* concept was a grouping of disorders according to a specific logic (by anatomic system for example). Likewise, classifications were organized under the general *hierarchy* concept and had two particular attributes: *usage* (context of use) and design logic (by anatomic system, biological mechanism, etc..). And each classification was related to a single *group of disorders* which was its top (head of classification).

4.1.2. From relational database to OWL formalism

In order to build the OWL resource with RDF/XML syntax from the database in line with the previously-mentioned ontological model, we chose to use an open data management environment (supporting extraction-transform-load (ETL) processes), adapted to our research step[8]. This tool, written in JAVA, al-

[8]Talend Open Studio v4.0, open integration solutions (`http://www.talend.com/`)

Table 1
Shared classifications (hierarchies) supported by the ontology.

ORPHANET classifications of rare diseases (32)	
Orphanet classification of developmental anomalies during embryogenesis	Orphanet classification of rare abdominal surgical diseases
Orphanet classification of rare allergic disease	Orphanet classification of rare bone diseases
Orphanet classification of rare cardiac diseases	Orphanet classification of rare cardiac malformations
Orphanet classification of rare circulatory system diseases	Orphanet classification of rare endocrine diseases
Orphanet classification of rare eye diseases	Orphanet classification of rare gastroenterological diseases
Orphanet classification of rare genetic diseases	Orphanet classification of rare gynecological and obstetric diseases
Orphanet classification of rare hematological diseases	Orphanet classification of rare hepatic diseases
Orphanet classification of rare immunological diseases	Orphanet classification of rare inborn errors of metabolism
Orphanet classification of rare infectious diseases	Orphanet classification of rare infertility disorders
Orphanet classification of rare intoxications	Orphanet classification of rare neurological diseases
Orphanet classification of rare odontological diseases	Orphanet classification of rare otorhinolaryngological diseases
Orphanet classification of rare psychiatric diseases	Orphanet classification of rare renal diseases
Orphanet classification of rare respiratory diseases	Orphanet classification of rare skin diseases
Orphanet classification of rare surgical maxillo-facial diseases	Orphanet classification of rare surgical thoracic diseases
Orphanet classification of rare systemic and rhumatological diseases	Orphanet classification of rare tumors
Orphanet classification of rare urogenital diseases	Orphanet classification of teratologic disorders
Other ORPHANET classifications (2)	
Classification of elsewhere unclassified rare diseases	Classification of Orphanet non-rare diseases

Table 2
URI prefixes for OWL Classes in ONTOORPHA.

Class	URI prefix
phenome	`http://www.orphanet.org/rdfns#pat_id_`
sign	`http://www.orphanet.org/rdfns#sgn_id_`
gene	`http://www.orphanet.org/rdfns#gen_id_`
hierarchy	`http://www.orphanet.org/rdfns#hch_id_`

lowed data extraction from the relational database, their transformation (NLP processes, XML serialization), their loading in a triplestore by JAVA connectors, the execution of SPARQL queries on this store and the generation of reports.

Diseases concepts selection Rare diseases which should be in the ontology were selected according to the knowing of their type ; typed *phenomes* were selected only. The knowledge of each *phenome* type was an expert knowledge, absent from the database and specifically produced following the study of the need for a core ontology. This selection of *phenome* was required for the database contained old nomenclatures (previous versions of ORPHANET) which were no longer relevant to represent rare diseases according to current knowledge. This process, quite simple, used a filter in our ETL processes.

OWL Classes definition The URI[9] of each concepts were built from the key of the table for the declared class, providing its uniqueness. This key was concatenated to a prefix defined for each class (table 2). Classes corresponding to *phenome* and *gene* subclasses were extracted from the table TbPat. Subclasses of *clinical sign* were extracted from TbSgn and subclasses of *hierarchy* from table TbHch (see "Declaration of classes" in table 3).

Terms definition SKOS[10] language is defined as a language of representation of knowledge organization systems such as thesauri, taxonomies or any other type of controlled and structured vocabulary. This standard make available some primitives dedicated to terminology with a preferred term (*skos:prefLabel*), synonyms (*skos:altLabel*) and a definition (*skos:definition*) for every languages. These primitives belong-

[9] *Uniform Resource Identifier* : unique identifier of a web resource respecting the syntax defined by Berners-Lee et al. (2005)
[10] Le *Simple Knowledge Organization System* (SKOS) est développé dans le cadre du W3C depuis 2003 (Miles and Bechhofer, 2009).

Table 3
Examples : concepts of gene, disease and sign for the Marfan Syndrome. (* automatically generated code by ETL process)

Declaration* of the owl:Class : *Mafran Syndrome* and its labels

```
<owl:Class rdf:about="&orpha;pat_id_109">
 <skos:prefLabel xml:lang="en">Marfan Syndrome</skos:prefLabel>
 <skos:prefLabel xml:lang="fr">Syndrome de Marfan</skos:prefLabel></owl:Class>
```

Declaration* of the owl:Class : *Hyperextensible joints/articular hyperlaxity*>

```
<owl:Class rdf:about="&orpha;sgn_id_46360">
```

Declaration* of the owl:Class : *Transforming growth factor, beta receptor II*

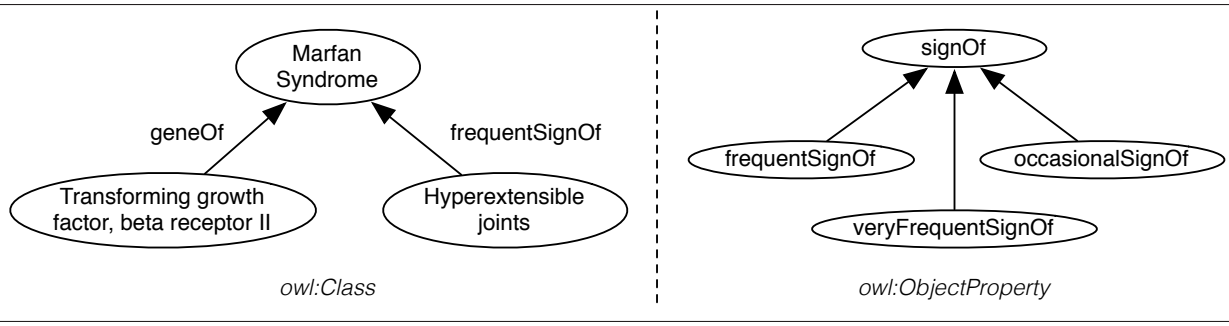
```
<owl:Class rdf:about="&orpha;gen_id_15611">
```

Declaration of the owl:ObjectProperty (domain and range) : *signOf*

```
<owl:ObjectProperty rdf:about="&orpha;signOf">
 <rdfs:range rdf:resource="&orpha;typ_id_4"/>
 <rdfs:domain rdf:resource="&orpha;sgn_id_1"/></owl:ObjectProperty>
```

Declaration of a owl:Restriction on Property : *signOf*

```
<owl:Class rdf:about="&orpha;sgn_id_46360">
 <rdfs:subClassOf><owl:Restriction>
  <owl:onProperty rdf:resource="&orpha;frequentSignOf"/>
  <owl:someValuesFrom rdf:resource="&orpha;pat_id_109"/>
 </owl:Restriction></rdfs:subClassOf></owl:Class>
```



ing to a widely-used standard were appropriate for the representation of the labels (and synonyms) of the ontology concepts (see "Declaration of the owl:Class *Marfan Syndrome* and its labels" in table 3).

Subsumbtion definition (rdfs:subclassOf) *Group of disorders* subclasses formed a subsumption hierarchy determined from BTNT relations between this type of phenomes only. *Disorder* subclasses (malformative syndrome, etc..) had a single subclass level. *Subtype* subclasses formed a subsumption hierarchy also determined from BTNT relations between this type of phenomes only.

Properties definition Class properties in the ontology (*owl:ObjetProperty*) were declared in the header-block, so were their domain and range (see *Declaration of the owl:ObjectProperty signOf* in table 3). Restrictions on these properties were automatically generated by ETL processes from current link tables (see *Declaration of a owl:Restriction*, table 3), apart from elements of the core ontology. Restrictions on the property *partOf* (linking a *disorder* to a *group*) were automatically generated for each one of the 32 classifications, from the relation *broader than* between the last *group* of the hierarchy and the first *disorder*. Likewise, restrictions on the property *hasSubtype* (linking a *subtype* to a *disorder*) were automatically generated from the relation *narrower than* between the first *subtype* of the hierarchy and the last *disorder*.

OWL anotations definition Some attributes of the ontology concepts were not transmissible along the subsumption tree. We linked them to the concept by annotations. This seemed obvious for the concepts labels or else for their definitions. Like in other medical ontologies (Dhombres et al., 2010a), we also used annotations to link external references identifications (diseases classifications, genetics databases,

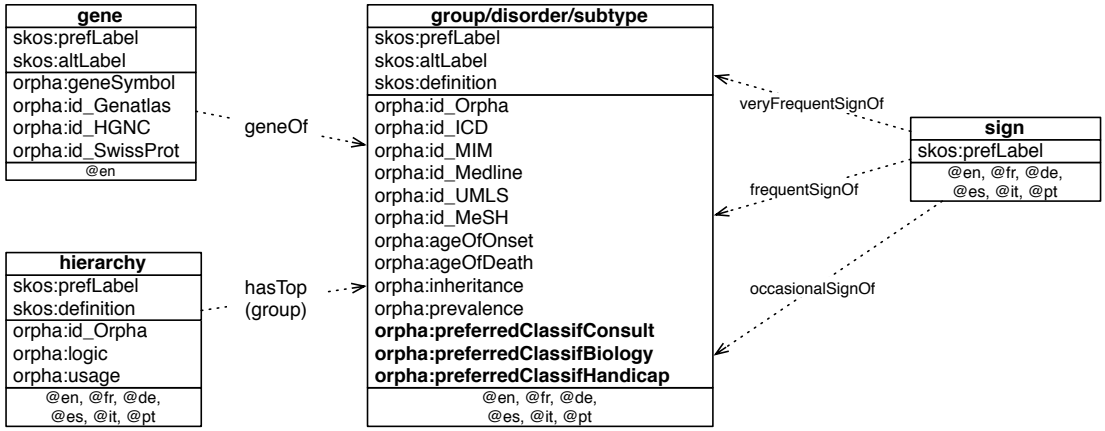


Fig. 4. OWL annotations of the main ONTOORPHA classes.

etc..) to *phenome* and *gene* subclasses. The "age of onset", the "age of death", the "prevalence" and the "inheritance mode" of a *disorder* (or of a *subtype*) were characteristics also linked to disorder and subtype concepts in the form of annotations[11] (fig. 4).

Group of disorders did not have this type of characteristics.

RDF/XML header definition The RDF/XML file header of ONTOORPHA was generated by an ETL process from manually-curated files and from the database. The ontology version was determined by an ETL process from the creation date of the ontology. The header-block had enrichment elements for knowledge formalization. These elements were not in the database and were extracted from a manually-edited files containing:

- the core ontology class definitions (using a dedicated prefix[12]), labels and subsumption hierarchy,
- the object properties definitions (domain and range),
- the ontology meta-data definitions (using Dublin Core[13])

Triplestore uploading and SPARQL *querying* The file generated by our processes was uploaded in a JAVA triplestore (during a testing phase, we chose the Sesame server, developed as part of *Aduna Open Source project*[14]). SPARQL queries were captured in the Web interface or automated for iterative needs (for example, validation and classifications generation rules).

4.2. Knowledge curation : editing and validation

In our targeted information system, the ontology is a continuous editable piece of knowledge. To keep a certain level of quality and structure in order to fulfill its initial requirements (generate classifications) we proposed a validation architecture based on queries and rules. Since the ontology was structuring both the rare diseases and the classifications, it was easier to build validation rules or queries: we can take advantage of real sub-classes and typed relations. We have defined a set of anti-patterns, which is composed of SPARQL queries and N3 rules. Whilst the SPARQL query statements represent an implicit formulation of anti-patterns (the false pattern is not stated as such in the query), the N3 rules explicitly states the false patterns. The N3 rules can also profit from the inferences the reasoner can do at runtime.

As in example fig. 5, we designed a set of rules to detect false triples in the ontology. We can consider this type of rule also as a integrity rules, we cannot implement this directly in OWL. As in example fig. 6, we looked for triples that match our query pattern. Whilst the N3 rules required to be resolved using a

[11] This temporary choice should evolve into the definition of properties in a future version of the ontology.

[12] http://www.orphanet.org/rdfns\#typ_id_

[13] http://dublincore.org/documents/2010/10/11/dcmi-terms/

[14] http://www.openrdf.org/

```
 {  ?DISEASE rdfs:subClassOf orpha:typ_ID_4.
    ?S e:findall ( ?LABEL
      {  ?DISEASE skos:prefLabel ?LABEL.
         (?LABEL) func:lang-from-PlainLiteral "en".
      }  ?RESULT ).
    ?RESULT math:memberCount ?COUNT.
    ?COUNT math:greaterThan 1.
 }
=> false.

 {  ?DISEASE rdfs:subClassOf orpha:typ_ID_4.
    ?S e:findall (?LABEL
      {  ?DISEASE skos:prefLabel ?LABEL.
         (?LABEL) func:lang-from-PlainLiteral "en".
      } ?RESULT).
    ?RESULT math:memberCount ?COUNT.
    ?COUNT math:equalTo 0.
 }
=> false.
```

Fig. 5. N3 rule stating that a *disorder* (*orpha:typ_ID_4*) should have 1 and only 1 english main label (*skos:prefLabel*).

```
PREFIX orpha:<http://www.orphanet.org/rdfns#>
SELECT DISTINCT ?classOnto ?prefLabel ?patType
WHERE{
  ?subclass rdfs:subClassOf orpha:typ_ID_4.
  ?classOnto rdfs:subClassOf ?subclass.
  ?classOnto skos:prefLabel ?prefLabel.
  ?classOnto orpha:patType ?patType.
  FILTER(Lang(?prefLabel)=\"fr\")
  OPTIONAL{?classOnto rdfs:subClassOf ?restriction.
    ?restriction owl:onProperty orpha:partOf. }
  FILTER(!bound(?restriction))
 }
```

Fig. 6. SPARQL query looking for the number of sub-sub-classes of *disorder* (*orpha:typ_ID_4*) that do not have a *partOf* relation with a *group of disorders*.

reasoner (Euler[15]), the SPARQL query was simply processed by a query engine. The result of queries or inferences were then treated and notified to the editing user.

4.3. Classifications generation

The SPARQL language that we used was adapted to graph generation from an ontology stored in a triplestore. Various classifications were generated by stable queries. First of all, regarding the classifications designed to be shared as such (table 1), the generated graph was a BTNT relations hierarchy, in SKOS. The generated classifications can be visualized by experts thanks to various tools. Then, the classifications called "functional", whose usage was information display management on the Website, or specific Web services support, were generated by particular rules requiring extra expert knowledge of phenomes, modeled in the form of annotation. This expert knowledge consisted in allocating a preferred classification for each one of the phenomes, for a definite usage (fig. 4).

5. Results and discussion

In this section, we have distinguished: 1) specifications of the target architecture and its first implementation step 2) the produced ontology 3) the contribution of this ontology in our target architecture

[15]Euler is an inference engine supporting logic based proofs. scriptsize `http://www.agfa.com/w3c/euler/`

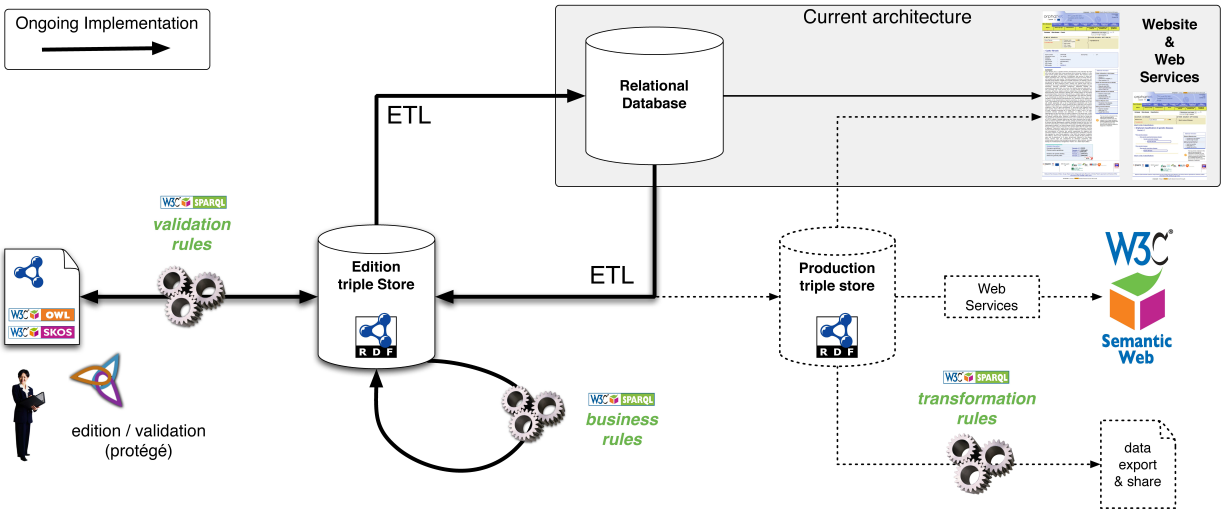


Fig. 7. Architecture proposal for rare diseases knowledge curation, besides the current architecture in ORPHANET.

Table 4
ONTOORPHA metrics (793 779 RDF triples).

owl:Class	11 077	owl:Restriction	
owl:ObjectProperty	10	orpha:occasionalSignOf	10 530
owl:AnnotationProperty	29	orpha:frequentSignOf	12 384
AnnotationAssertion	194 337	orpha:veryFrequentSignOf	21 281
Poly-parental Classes	2 843	orpha:geneOf	3 819

for editorial processes 4) our approaches of knowledge validation 5) the various sharing methods of the knowledge supported by the new architecture and 6) some challenging issues in a production persepective.

5.1. *Respect for the current architecture*

The target architecture that we proposed (fig 7) was built upon a domain ontology. The difficulties of such a setting up can be explained by the fact that this modification must be achieved without troubling the current evolutions and changes. And the use of knowledge by the portal but also by other partners must always be available. Thus, we firstly integrated our ontology-based architecture besides the current functional architecture. Eventually, we will consider the whole process through semantic representations.

5.2. *A domain ontology for rare diseases edition and validation*

Our method allowed us to draw up a design process of a first ontology version. This entirely automated process runnable on demand encouraged us to consider a release: 7 minutes to upload the files extracted from the Sybase, to produce the ontology and to upload in the RDF store (the construction step of the ontology itself during less than a minute). The ontology metrics were determined by SPARQL queries on the RDF triplestore (see table 4). Annotations of the disease class (*prefLabel*, *altLabel*, *definition* and external reference annotations) are detailed in table 5. The verification of ETL exportation processes confirmed that the base entries were all converted to classes (table 6). There were more classes in the ontology than entries in the base. This difference was explained by formalism improvement thanks to the addition of the core ontology whose classes were defined in the file header-block.

Table 5
Les annotations de la classe *maladie* dans ONTOORPHA (parmi les 194 337 annotations de l'ontologie).

	total	fr	en	de	es	it	pt
AnnotationAssertion	153 513	23 203	23 641	20 808	18 625	19 253	14 126
skos:prefLabel	42 970	7 163	7 163	7 161	7 161	7 161	7 161
skos:altLabel	27 984	5 615	5 990	4 836	4 837	4 389	2 317
skos:definition	22 152	6 000	6 063	4 386	2 202	3 278	223

Table 6
Result of ONTOORPHA building from the RDB.

	group of disorders, disorder, subtype	*clinical sign*	*gene*	*hierarchy*
ORPHANET RDB (ID)	7,161	1,360	2 415	106
ONTOORPHA (URI)	7,164	1,361	2 416	123

5.3. Editing methods evolution

Thanks to the separation of knowledge of rare diseases domain and business rules, editing methods were simpler. Any knowledge that can be automatically generated by rules was no longer manually edited.

The ontology editor PROTÉGÉ[16] was used with a specific plugin: ARCHONTE[17]. We compared current processes with experimental processes[18] within our target architecture in two cases of routine use of knowledge updating by domain experts (table 7). Processes for the two studied cases were simplified by experimental tools. The main improvement was the visibility of the modifications being edited (in particular for hierarchies). The decrease in errors during the manual edition of the spreadsheet lines was not quantified in this work but seemed acquired. The terminology management plugin was also easier to use. So the initial choice of PROTÉGÉ (ontology editor) allowed significant progress.

5.4. Knowledge validation approaches

Thanks to formal definitions of the model of rare diseases which were in the ontology, the building of integrity validation rules became easier. Constraints expressed in the ontology were first checks on integrity. For example, the relation *signOf* applied by definition to a *clinical sign* and had an object of the *disorder* type as range. If some facts within the knowledge base did not respect this definition, an error was generated. So validation rules could be based on the ontology, for example its transitive relations, in order to consider all the diseases with a specific sign. Thus, if a new *disorder* appears with this *clinical sign*, the rule will dynamically apply to it.

As stated in 4.2, we have implemented a validation architecture based on anti-patterns that we can consider as integrity rules. A first set of anti-patterns have been implemented, five SPARQL queries and three N3 rules. The table 8 depicts the result of our experiment. This first experiment has validated our overall approach. SPARQL queries were sometimes more convenient to write and use whereas N3 rules combined with a proof engine were more explicit and could be generalized.

Table 8 provides a global view of the knowledge base: all the phenomes had a label, there were clinical signs, genes and a transmission mode which were not linked to a *disorder* nor to a *subtype*. And the transmission mode of half of rare genetic *disorder* or *subtype* was not listed. The results of table 5 come from SPARQL queries and provide an interesting view of the progress of diseases annotations translations.

[16] PROTÉGÉ, Version 4.1.0, build 209 (`http://protege.stanford.edu/`)

[17] This plugin was developed in our research department by L. Mazuel. It results from the integration of a part of DOE software functionalities into PROTÉGÉ combined with an annotation interface managing multilinguism and SKOS labels (*definition*, *prefLabel* and *altLabel*).

[18] Secure access to the data and right management are not detailed here because they are not implemented in our experimental processes.

Table 7
Current and research processes for knowledge update.

Case 1 : disorder label update – CURRENT PROCESS
– access to the proprietary update interface (MAJOR tool)
– select editable data (button labelled *disease*)
– search and select the disease by numeric idenfier
– edit the field *main label* FR/EN (synonyms editing 2 screens further)
– validate and run BD update (by a dedicated button labelled *update*)
Case 1 : disorder label update – RESEARCH PROCESS
– open PROTÉGÉ tool and the generated RDF/XML file : ontoOrphanet.owl
– search the disorder to edit (search field by label, with autocompletion)
– edit and visualize all the labels (main and synonyms) for each language (ARCHONTE *plugin*)
– save file
Case 2 : moving a disease in a hierarchy – CURRENT PROCESS
– open the proprietary DB view extraction tool (PLATOR)
– select (4 screens) et download the csv file containing the diseases hierarchies (PatPrt.Txt)
– run the proprietary visualization application (ARBOR tool generate a static tree of all the hierarchy from the table PatPrt)
– edit the file PatPrt.Txt line by line (over 40,000 lines) in a spreadsheet : (1) invalidate lines to be deleted, (2) new lines creation [hierarchy-disease-broader disease] for each hierarchy containing the disease
– save the updated lines in a new file
– access to the proprietary update interface by file (INJECTOR tool), and upload this new file and wait for an update report (Email send by the system)
– perform a new extraction of PatPrt.Txt (PLATOR) and relaunch ARBOR in order to visualize the effective updates.
Case 2 : moving a disease in a hierarchy – RESEARCH PROCESS
– open PROTÉGÉ tool and the generated RDF/XML file : ontoOrphanet.owl
– search the disorder to edit (search field by label, with autocompletion)
– *drag'n drop* the disease in the hierarchy (with a dynamic visualization of the herarchies)
– save file

Table 8
Knowledge validation based on rules (N3 and SPARQL).

0	phenome with more than one main label (for each language)
0	phenome with more than one PubMed link
0	phenome without any label
1	inheritance mode without at least one related disorder
4	signs without at least one related disorder
20	genes without at least one related disorder
2,494	genetic disorder(s) without at least one inheritance mode (in comparison with the full ORPHANET collection of 5,272 rare genetic disorders)

5.5. Sharing and distribution of formalized knowledge

Automated classifications generation We used SPARQL query language in order to describe our validation rules (section 5.4). The use of the *OPTIONAL* function and the *CONSTRUCT* operator allowed the generation from the ontology of 32 classifications necessary for knowledge distribution, such as the rare genetic diseases classification (Dhombres et al., 2011).

From the model point of view, a classification was defined in a generic way by a top, a logic and a usage (fig. 8). The head (top) of classification was a *group of disorders*, for which an harmonization work with other international classifications is being carried out. The logic of a hierarchy were *owl:annotation* : "by system", "by mechanism", "by deficit", etc... The usage (edition, consultation, biology, disability) determined the links between *phenomes* and ORPHANET resources. The presence of the classification

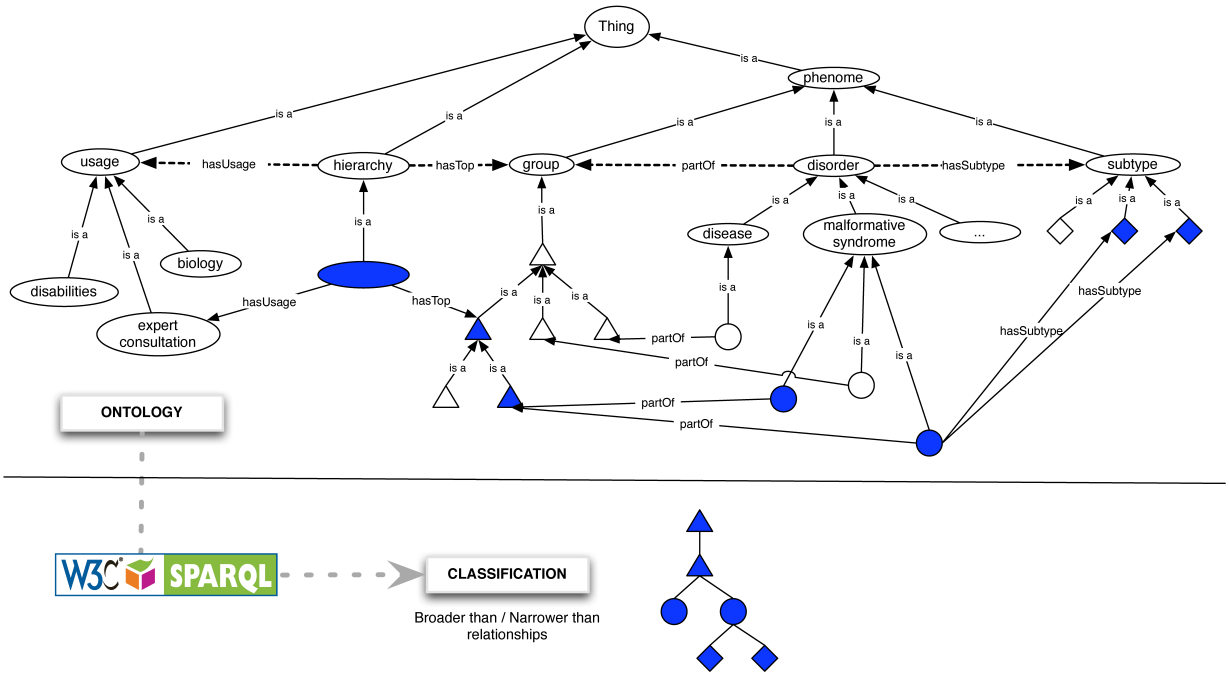


Fig. 8. From ontology to terminology : general principles in BTNT classification generation using SPARQL CONSTRUCT.

usage information (property) and design logic (annotation) determined the classification generation rule (SPARQL query). For each *phenome* (*group of disorder*, *disorder*, *subtype*), two pieces of information are required, in addition to its type, for the generation of the numerous ORPHANET classifications (in particular classifications used within the portal, allowing to link services information to phenomes): its preferred classification and the context (usage) of this preference. For now, only 32 classifications with a clinical usage (consultations) have been successfully generated, allowing the complete updating cycle of the RDB according to the knowledge life cycle that we initially proposed (fig. 1).

Distribution by SPARQL *endpoints* We previously mentioned the classification distribution as a significant modality of rare diseases knowledge sharing. Semantic Web tools allowed to consider the sharing of the ontology itself (and not of the mere terminology) via SPARQL endpoints. ORPHANET rare diseases knowledge will be stored in a new RDF triplestore, opening to new Web services. By transformation and SPARQL export rules, it was possible to generate on demand export files in Semantic Web formats. These transformation rules, using the ontology formal definition, had advantages: easy curation and scalabity. Indeed, these rules based on semantic will dynamically evolve as the content changes. Another perspective is also considered: the direct use of RDF triples for ORPHANET Website supply.

Ontology sharing and portals integration It was possible to share directly generated classifications. But the automation of alignment processes from annotations of the ontology concepts by external references contributed to the integration into biomedical ontologies sharing portals. For example, *BioPortal*[19] offered many Web services for the ontologies that it hosted (access to the content, term research, alignments, annotation of documents, etc..). A possible alternative, logically requiring a strong harmonization of the candidate ontology with the top-ontology BFO[20] was proposed by *The OBO Foundry*[21]. ONTOORPHA ontology was initially put online on *BioPortal* in order to offer an original resource on rare diseases that could be used in other research projects (Adamusiak et al., 2011).

[19] http://bioportal.bioontology.org/
[20] http://www.ifomis.org/bfo/
[21] http://www.obofoundry.org/

5.6. Operational stakes

Thanks to the contribution of more expressive languages and formalisms, we intended to improve the quality of data (knowledge) provided by ORPHANET in France and in Europe, but also the flexibility of the current production tool (spreadsheets + interfaces + relational databases). We think that it is necessary to reorganize the information architecture within ORPHANET by using for example appropriate representation languages according to nature and processed data. The separation of the information linked to concepts and classifications (OWL), terms management (SKOS) and business rules management (N3) will provide more flexibility, interpretation and make data extraction and curation easier. We must offer an information architecture compatible with a production service where the functioning range of the service is continuous and where several users across Europe can work on the ORPHANET database content. We already observed the limits of tools such as Sesame and PROTÉGÉ for user rights management or multi-position accesses. Consequently, within the hypothetical framework of the ORPHANET RDB (Sybase) replacement by a RDF triplestore, the management of access and rights security must be insured. Moreover, scalability measurements consistent with the website activity will be carried out. The first tests on the use of efficient triplestores combined with visualization and edition tools (e.g. ALLEGROGRAPH-GRUFF[22]) were encouraging because they could allow a better integration within our target architecture and one of the highest performance from the perspective of a release.

6. Conclusion

The first step of the ORPHAONTO project produced positive results. A first valid version of the rare diseases ontology formalized in OWL-DL within a dedicated architecture allowed an evolution of editorial methods and a new approach of ORPHANET knowledge bases audit and sharing. The choices of knowledge representation among the distinct KOSs according to the life cycle of knowledge on rare diseases allowed a consistent adaptation of the resource (terminological and ontological) to the considered uses.

KOSs had indeed different uses: the great expressivity capacity of ontologies allowed a tailored representation of domain knowledge. But for some uses (such as the management of data from information systems) where speed is an issue, a terminology will be more appropriate. The formalism of ONTOORPHA ontology allowed us to improve the quality of knowledge organization on rare diseases by its core ontology and its semantic relations that were richer than a mere terminology. Moreover, the use of W3C standards provided KOS setting up durability.

This life cycle illustrated a real *continuum* between the present KOSs. However, a part of expert knowledge, essential to this project, could not be modeled by the usual terminologies and required a semantic modeling based on an ontology and the use of rules. So there was a form of *discontinuity* and terminologies could be considered as particular views of the knowledge without covering it as widely as ontology.

References

Adamusiak, T., Burdett, T., Kurbatova, N., Joeri van der Velde, K., Abeygunawardena, N., Antonakaki, D., Kapushesky, M., Parkinson, H., and Swertz, M. A. (2011). OntoCAT–simple ontology search and integration in java, r and REST/JavaScript. *BMC Bioinformatics*, 12:218.

Amardeilh, F., Bousquet, C., Guillemin-Lanne, S., Wiss-Thébault, M., Guillot, L., Delamarre, D., Louet, L., and Burgun, A. (2009). A knowledge management platform for documentation of case reports in pharmacovigilance. *Medical Informatics Europe*.

Aymé, S. (2002). Orphanet : The portal for rare diseases and orphan drugs, INSERM SC11.

Berners-Lee, T., Fielding, R., and Masinter, L. (2005). RFC 3986/STD 0066 - uniform resource identifier (URI): generic syntax.

Binding, C. and Tudhope, D. (2006). Kos at your service: programmatic access to knowledge organisation systems. *Journal of Digital Information*, 4(4).

Bizer C., Cyganiak, R. (2007). D2rq — lessons learned. Technical report, Cambridge. Position paper for the W3C Workshop on RDF Access to Relational Databases.

[22] http://www.franz.com/agraph/

Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. *Text mining for biology and biomedicine*, pages 43–66.

Bodenreider, O. (2008). Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*, 67:79.

Cerbah, F. (2008). Learning highly structured semantic repositories from relational databases: The rdbtoonto tool. In *In Proc. of ESWC 2008*.

Degoulet, P., Fieschi, M., and Attali, C. (1997). Les enjeux de l'interopérabilité sémantique dans les systèmes d'information de santé. *Volume 9 Springer-Verlag France, Paris*.

Dhombres, F., Aymé, S., Rath, A., Olry, A., Vandenbussche, P., and Charlet, J. (2011). Automated generation of diseases classification supported by orphanet ontology of rare diseases. In *12th International Congress of Human Genetics / 61st Annual Meeting of The American Society of Human Genetics*, Montreal, Canada. [poster][accepted].

Dhombres, F., Charlet, J., Jouannic, J., Mazuel, L., and Jaulent, M. (2010a). Re-use of terminological and ontological resources for the construction of domain ontologies in medicine: a description of two experimental approaches. In *EKAW 2010: Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management.*, pages 1–12, Lisbonne, Portugal. Springer-Verlag. [workshop] [http://www-limbio.smbh.univ-paris13.fr/ReuseOnto-EKAW2010/].

Dhombres, F., Jouannic, J., Jaulent, M., and Charlet, J. (2010b). Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In *Actes des 21e Journées Francophones d'Ingénierie des Connaissances*, pages 1–12, Nîmes, France. [http://hal.archives-ouvertes.fr/hal-00487736/fr/].

Giannangelo, K. (2006). *Healthcare code sets, clinical terminologies, and classification systems*. American Health Information Management Association (AHIMA).

Greenberg, J. (2001). Automatic query expansion via lexical–semantic relationships. *Journal of the American Society for Information Science and Technology*, 52(5):402–415.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries*. Citeseer.

Krivine, S., Nobécourt, J., Soualmia, L., Cerbah, F., and Duclos, C. (2009). Construction automatique d'ontologie à partir de bases de données relationnelles : application au médicament dans le domaine de la pharmacovigilance. In Gandon, F., editor, *20 th French Knowledge Engineering Workshop*, pages 73–84, Hammamet, Tunisie.

Miles, A. and Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C recommendation.

Miller, P. (2000). Interoperability: What is it and why should i want it? *Ariadne*, 24.

O'Connor, M. J. and Das, A. (2010). Semantic reasoning with xml-based biomedical information models. In Safran, C., Marin, H. F., and Reti, S. R., editors, *MEDINFO 2010 - Proceedings of the 13 th World Congress on Medical and Health Informatics - Partnerships for effective e-Health solutions*, volume 160 of *Stud Health Technol Inform*, pages 986–90, Cape Town, South Africa. IOS Press.

Polleres, A., Scharffe, F., and Schindlauer, R. (2007). SPARQL++ for mapping between RDF vocabularies. In *6th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE 2007)*, page 4803, Vilamoura, Portugal. Springer-Verlag.

RDB2RDF Incubator Group, W. (2009). A survey of current approaches for mapping of relational databases to rdf. Technical report, W3C. `http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport_01082009.pdf`.