Projet OrphaOnto – Première étape de l'ontologisation des bases de connaissances d'ORPHANET

Ferdinand Dhombres[1,2,3,4,5], Pierre-Yves Vandenbussche[1,3,6], Ana Rath[2], Marc Hanauer[2], Annie Olry[2], Bruno Urbero[2,7], Rémy Choquet[1,3] & Jean Charlet[1,3,8]

[1] INSERM UMRS 872 ÉQ.20, Ingénierie des connaissances en santé, Paris, France
[2] INSERM SC11, ORPHANET, Plateforme Maladies Rares, Hôpital Broussais, Paris, France
[3] UPMC – Université Pierre et Marie Curie, Paris, France
[4] Faculté de médecine Paris Descartes, Université Paris Descartes, Paris, France
[5] Service de Gynécologie-Obstétrique et Centre de Diagnostic Prénatal de l'Est Parisien, Hôpital Armand Trousseau, AP-HP, Paris, France
[6] MONDECA, Paris, France
[7] INSERM DSI – DR Langedoc-Roussillon, Montpellier, France
[8] AP-HP – Assistance Publique Hôpitaux de Paris, Paris, France.
{ferdinand.dhombres, jean.charlet}@inserm.fr

Résumé : La première étape du projet ORPHAONTO comprend la construction d'une ontologie de domaine des maladies rares à partir de bases de données relationnelles et son utilisation comme support de l'édition des connaissances au sein d'ORPHANET. Une procédure fonctionnelle de construction de l'ontologie est mise en place et des résultats sont obtenus pour améliorer les procédures d'édition et de validation des connaissances.

Mots-clés : Ontologie de domaine, Maladies rares, OWL-DL, SPARQL

1. Introduction

ORPHANET est un service de documentation financé par la Commission Européenne et, en France, par l'INSERM et la Direction Générale de la Santé, qui a pour mission de fournir des informations sur les maladies rares [1] et les médicaments orphelins aux professionnels de santé et au grand public, afin

1. En Europe, le seuil de prévalence admis définissant une maladie comme rare est d'une personne atteinte pour 2000.

d'améliorer le diagnostic et la prise en charge des maladies rares. À cet égard, ce service a mis en place un portail d'information multilingue [2] constitué de classifications de maladies rares, d'une encyclopédie en ligne ainsi que de registres de consultations expertes, de laboratoires de recherche, de projets de recherche en cours et d'associations de malades (Aymé, 2002). En 2010, le site a été consulté en moyenne par 10 000 visiteurs par jour dont 1/3 de médecins, 1/5 d'autres professionnels de santé, 1/3 de malades. [3]

Après une évolution sur près d'une dizaine d'années, les bases de connaissances d'ORPHANET sont devenues si complexes qu'il est devenu très difficile pour les responsables scientifiques d'en assurer la maintenance avec les outils dont ils disposent actuellement. Le domaine des maladies rares est vaste (près de 6 000 maladies rares sont référencées par ORPHANET) et l'évolution des connaissances y est rapide. Les outils de représentation et de validation utilisés pour l'édition et la maintenance des connaissances se résument à des tableurs permettant de modifier des vues de la base de données. Ces outils ne permettent pas aux éditeurs scientifiques d'avoir une vue correcte des connaissances structurées ; une navigation dans les hiérarchies en cours d'édition n'y est par exemple pas possible. De plus, ORPHANET connait des difficultés croissantes pour répondre aux sollicitations de plus en plus nombreuses d'extractions de données. Malgré 12 jeux de données standardisés, il y a 30 à 50 demandes par an de données « sur mesure » différentes (pour des projets de recherche, des institutions ou des industriels). Et de nombreux documents de communication (Les cahiers d'ORPHANET, les annuaires, ...) doivent être actualisés régulièrement selon des données de la base, et adaptés aux 36 pays partenaires d'ORPHANET. L'extraction des données est liée à leur représentation en base et celle-ci empêche, à ce jour, l'utilisation de modèles et d'outils plus adéquats à la représentation de connaissances.

2. Objectifs

Notre objectif général dans le cadre du projet ORPHAONTO est la migration des données d'ORPHANET à partir d'une base de données relationnelle vers une plateforme d'édition et de mise à disposition des connaissances reposant sur une ontologie. Nous posons l'hypothèse que cette plateforme fondée sur une ontologie sera capable de pallier les limites actuelles de représentation

2. Accès en ligne à l'adresse : `http://www.orpha.net`

3. Les statistiques complètes de la fréquentation du site au cours des 12 derniers mois sont disponibles en ligne à l'adresse `http://www.orpha.net/stat/orphanet/`

de la base, d'édition et de validation de contenu, de recherche d'information et de mise à disposition d'un contenu formalisé, adapté aux utilisations actuelles et futures sur le web de données.

Nous traitons spécifiquement dans cet article de notre méthologie de construction de l'ontologie ONTOORPHA et à la lumière de nos premiers résultats, nous discutons la capacité de notre modèle, d'une part à améliorer les processus éditoriaux, et d'autre part à représenter les connaissances du domaine. Ce travail débute donc par une présentation de travaux de construction d'ontologies, en particulier à partir de bases de données relationnelles (section 3.). Nous présentons ensuite la base de données d'ORPHANET et nos méthodes de construction de l'ontologie (section 4.), puis détaillons nos premiers résultats : l'ontologie elle-même ainsi que ses premiers cas d'utilisation au sein de notre architecture cible (section 5.). Nous discutons enfin le bilan de cette première étape, évocant les limites de l'expressivité d'OWL-DL pour nos besoins, l'apport possible de l'utilisation de règles ou d'une formalisation en OWL-*full* (section 6.). Notre conclusion introduit les prochaines étapes importantes de notre projet (section 7.).

3. État de l'art

La construction d'ontologies à partir de bases de systèmes d'informations a été principalement étudiée dans deux directions, la reprise de données de bases de données XML sous forme d'ontologies et la reprise des tables de système de gestion de base de données (SGBD) relationnelles, là aussi pour faire des ontologies. Dans la première direction, on trouve, dans le monde biomédical, différents travaux — *e.g.* O'Connor & Das 2010 – qui développent des connecteurs pour transformer des fichiers XML en fichiers OWL dans le but de faire du raisonnement. Dans la seconde direction, on trouve des travaux qui visent à attaquer directement un SGBD relationnel et à spécifier des motifs de nommage des classes et des instances. Un outil comme RDBToOnto (Krivine *et al.*, 2009) propose l'implémentation de méthodes de transformations développées ou permet le recours à une nouvelle méthode obtenue en spécialisant un convertisseur déjà intégré dans la plateforme.

Étant donnée la complexité de notre tâche et l'hétérogénéité des nos bases, aucune méthode n'était satisfaisante et chacune aurait demandé trop de reprises. Nous nous sommes alors orientés vers un environnement de gestion de données (ETL) libre pour la production de l'ontologie.

En ce qui concerne la réutilisation d'autres ressources comme développé

dans (Dhombres *et al.*, 2010a,b), le choix était relativement simple dans la mesure où notre base de travail est un thésaurus déjà existant que nous cherchons à réorganiser en même temps que les bases de connaissances qui lui sont associées : dans un premier temps, nous disposons donc de tout le matériel nécessaire à notre projet.

4. Matériel et méthodes

4.1. Matériel : les bases de données d'ORPHANET

Le SGBD utilisé à ORPHANET est Sybase (version 15.5 pour Solaris 10). Un schéma simplifié (*i.e.* comportant uniquement les données correspondants aux maladies proprement dites) des tables de la version 4.1.0 de la base de données est représenté figure 1. Ce modèle de données a été contruit dans un objectif de diffusion de l'information sur les maladies rares à travers le site internet d'ORPHANET. Notre matériel de travail est constitué d'une extraction brute des tables de la figure 1.

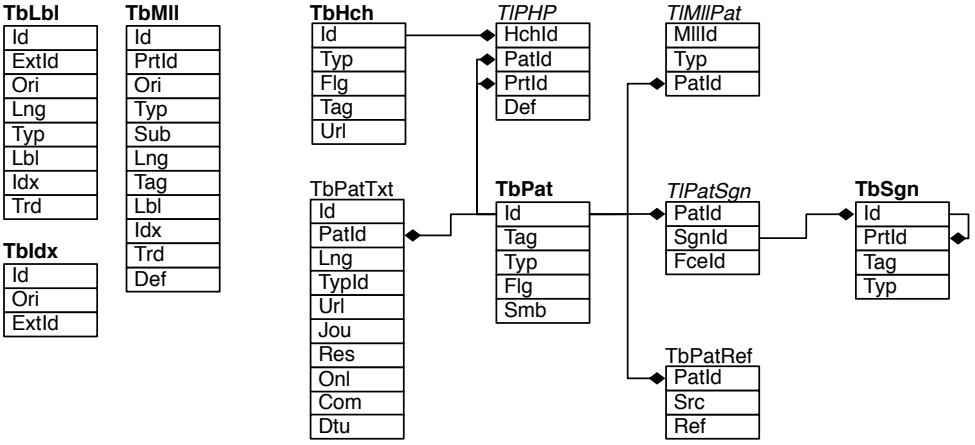

FIGURE 1: Extrait du schéma de la base de données relationnelle d'ORPHANET concernant les maladies (version 4.1) [a].

a. Détails des tables : **TbPat** : table des maladies et des gènes, **TbHch** : table des hiérarchies, **TbSgn** : table des signes, **TbPatTxt** : table des textes (résumés et textes longs) des maladies, **TbLbl** : table des libellés, **TbMll** : table des libellés multilingues et des données épidémiologiques, **TbIdx** : table des index, **TlPHP** : table des liens hiérarchiques entre maladies, **TlPatSgn** : table des liens entre maladies et signes, **TlPatRef** : table des liens entre maladies et référence externe.

L'ensemble des 5781 maladies rares référencées par ORPHANET sont dans la table TbPat, dans des tuples identifiables à partir de la colonne Typ par la valeur "Pat". Des jointures avec les tables TlMllPat et TlPatSgn permettent d'obtenir respectivement les données épidémiologiques (par exemple la prévalence, l'âge d'apparition, ...) et les signes associés à un identifiant unique de maladie (id_maladie). Des jointures avec la table TbLbl permettent d'obtenir dans toutes les langues de la base les libellés et les synonymes des maladies sous forme de tuples [id_maladie–libellé–langue] et [id_maladie–synonyme–langue].

Il existe 106 hiérarchies différentes dans la base de données. 32 classifications rassemblent la totalité des maladies rares des classifications ORPHANET (par exemple la classification des maladies génétiques rares, la classification des anomalies du développement, ou encore celle des anomalies rares de l'œil, de la peau *etc*...). La structure de ces hiérarchies de subsomption, avec un important contingent de polyparentalité, est dans la table TlPHP sous la forme de tuples [id_hiérarchie–id_maladie–id_maladie mère]. Pour une hiérarchie donnée, l'arborescence est dont reconstruite de manière récursive à partir de ces données. Les 74 autres classifications de maladies représentent soit des vues spécifiques à une activité de service [4] dans le domaine des maladies rares, soit des classifications qui ne sont pas produites par ORPHANET mais issues de la littérature médicale [5] et présentent un intérêt pour les maladies rares, soit encore une classification de maladies non rares qui est nécessaire à la représentation de certains liens entre maladies. Il existe également des hiérarchies correspondant à des projets de recherche internes.

Les 1360 signes ou groupes de signes présents dans la base de données sont dans la table TbSgn. Un ensemble de jointures permet d'obtenir l'identifiant de la maladie ainsi que l'identifiant de la fréquence du signe dans la maladie sous la forme de tuple [id_signe–id_maladie–id_fréquence]. Il existe trois types de fréquence liant un signe à une maladie (occasionel, fréquent et très fréquent), prémices de trois sous-types de relations sémantiques "signe de" (*signOf*).

L'ensemble des gènes (2415) sont dans la table TbPat, dans des tuples identifiables à partir de la colonne Typ par la valeur "Gen". Comme pour les

4. Ces classifications sont dites "classifications fonctionnelles" ; nous chercherons à les dériver des 32 premières selon des règles "métier" ou des règles "applicatives". La formalisation de ces règles fait partie des étapes suivantes de notre projet.

5. Ces classifications sont dites "classifications expertes". Il existe des classifications à la fois fonctionnelles et expertes.

signes, la contruction des paires [id_gène–id_maladie] à partir de jointures permet de formaliser la relation entre gène et maladie : “est un gène de” (*geneOf*).

4.2. Méthodes

Chacune des étapes de notre méthodologie a été établie par un groupe de travail composé de médecins experts du domaines, de documentalistes, d'informaticiens et d'ingénieurs ontologues. Au sein de ce groupe transdisciplinaire, repenser les méthodes de travail éditoriales pour se défaire des contraintes opérationnelles initiales (en particulier l'utilisation de tableur) a constitué une étape clé dans notre approche. Le modèle relationnel ne correspond pas au modèle de connaissance implicite des experts du domaine : nous avons donc choisi de définir les concepts du haut de notre ontologie (par exemple les concepts de maladie, de signe et de gène) dans une première version du modèle qui est présentée dans ce travail. Les relations identifiées précédemment (*signOf* et *geneOf*) permettent de définir les propriétés de ces classes.

4.2.1. De la base de donnée relationnelle au formalisme OWL

Afin de contruire la ressource OWL respectant la syntaxe RDF/XML à partir de la base de données en accord avec le modèle ontologique évoqué précédemment, nous avons choisi d'utiliser dans un premier temps un environnement de gestion de données (ETL) libre, adapté à notre démarche de recherche [6].

4.2.1.1. Sélection des maladies à partir des classifications

Les maladies rares devant figurer dans l'ontologie sont sélectionnées selon leur appartenance à une des 32 classifications ORPHANET, à la classification des maladies en cours d'édition ou à la classification des maladies non rares. Cette sélection de maladies est nécessaire car la base de données contient d'anciennes nomenclatures (versions antérieures d'ORPHANET) qui ne sont plus pertinentes pour représenter les maladies rares dans l'état actuel des connaissances. Cette procédure, relativement complexe, met en jeu un certain nombre de *flags* utilisés dans la base de données.

6. Talend Open Studio v4.0, *open integration solutions.* (`http://www.talend.com/`)

Déclaration* de la classe “Syndrome de Marfan” et de ses labels

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

Déclaration* de la classe “Hyperlaxité ligamentaire/articulation hyperlaxe”

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

Déclaration* de la classe “Transforming growth factor, beta receptor II”

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

Déclaration de la propriété “signOf”

```
<owl:ObjectProperty rdf:about="&orpha;signOf">
 <rdfs:range rdf:resource="&orpha;pat_id_0"/>
 <rdfs:domain rdf:resource="&orpha;sgn_id_1"/></owl:ObjectProperty>
```

Déclaration* d’une restriction

```
<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>
```

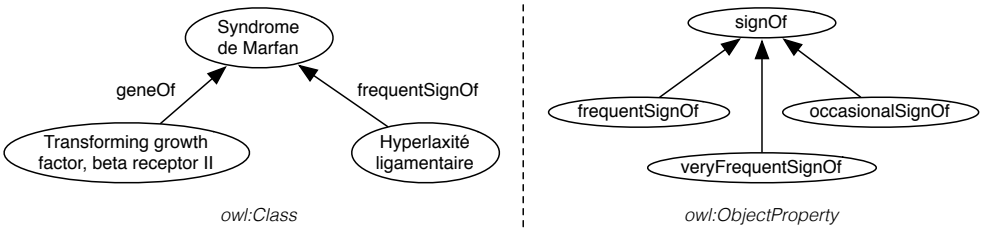

TABLEAU 1: Exemples : concepts de gène, maladie et signe autour du Syndrome de Marfan. (* code généré automatiquement par procédure ETL)

4.2.1.2. Définition des classes – OWL

L’URI [7] des concepts de maladie est construite à partir de la clé de la table pour la classe déclarée, assurant l’unicité de celle-ci. Cette clé est concaténée à un préfixe défini pour chacune des classes [8]. Les classes correspondantes aux maladies et aux gènes sont ainsi extraites de la table TbPat, les signes de TbSgn, et les hérarchies de TbHch (*cf.* “Déclaration des classes” dans le tableau 1).

7. *Uniform Resource Identifier* : identifiant unique d’une ressource web respectant la syntaxe définie par Berners-Lee *et al.* (2005)

8. Par exemple : `http://www.orphanet.org/rdfns#pat_id_` est le préfixe utilisé pour construire l’URI de la classe maladie.

4.2.1.3. Définition des labels des classes – SKOS

Le langage SKOS [9] se définit comme un langage de représentation de systèmes d'organisation de connaissances tels que thésaurus, taxonomies, ou tout autre type de vocabulaire contrôlé ou structuré. Ce standard met à disposition certaines primitives dédiées à la terminologie avec pour chaque langue, un terme préféré *prefLabel*, des synonymes *altLabel* et une définition *definition*. Ces primitives appartenant à un standard courament utilisée sont donc appropriées pour la représentation des noms et synonymes des concepts de l'ontologie (*cf.* "Déclaration de la classe Syndrome de Marfan et de ses labels" dans le tableau 1).

4.2.1.4. Définition des propriétés – RDFS

Les propriétés des classes de l'ontologie (*owl :ObjetProperty*) sont déclarées dans l'en-tête ainsi que leurs domaine et co-domaine (*cf.* "Déclaration de la propriété signOf" tableau 1). Les restrictions sont générées automatiquement par une procédure ETL à partir des tables de liens existantes (*cf.* "Déclaration d'une restriction" tableau 1).

4.2.1.5. Définition des annotations – OWL

Certaines caractérisques des concepts de l'ontologie ne sont pas héritables le long de l'arbre de subsomption ; nous les avons liées au concept par des annotations. Cela semble évident pour les labels des concepts ou encore pour leurs définitions. Comme dans d'autres ontologies médicales (Dhombres *et al.*, 2010a), nous avons également utilisé des annotations pour attacher les identifiants de références externes (classifications de maladies, bases de données de génétique, ...) aux classes *maladie* et *gène*.

L'âge d'apparition, l'âge de décès, la prévalence ou encore le mode de transmission d'une maladie fille ne sont pas celui ou celle de sa maladie mère. Les caractéristiques de ce type (*cf.* figure 2) ont donc été liées aux concepts sous forme d'annotations dans un premier temps. Ce point sera discuté au paragraphe 6.2.

9. Le *Simple Knowledge Organisation System* (SKOS) est développé dans le cadre du W3C depuis 2003 (Miles & Bechhofer, 2009).

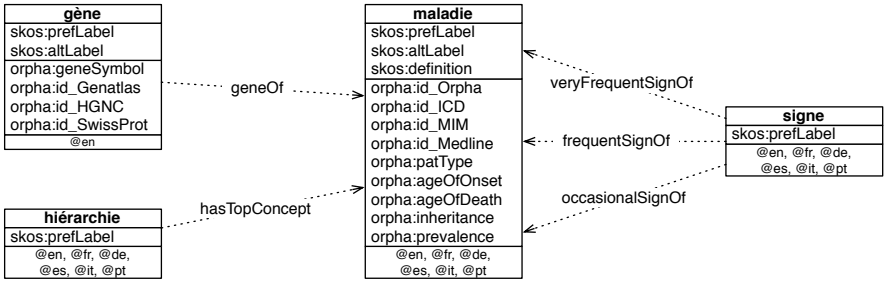

FIGURE 2: Annotations des concepts d'ONTOORPHA.

4.2.1.6. Définition de l'en-tête – RDF/XML

L'en-tête du fichier RDF/XML d'ONTOORPHA est générée par une procédure ETL à partir de fichiers maintenus manuellement et à partir de la base de données. La version de l'ontologie est déterminée par une procédure ETL à partir de la date de création de l'ontologie. L'en-tête contient donc des éléments d'enrichissement de la formalisation des connaissances non présents dans la base de données, qui sont extraits d'un fichier édité manuellement :

- définition des concepts du haut de l'ontologie de domaine (comme le concept de maladie, de maladie rare, de signe ou de hiérarchie),
- définition des propriétés,
- définition des méta-données de l'ontologie.

4.2.1.7. Entreposage des triplets et requêtes – SPARQL

Le fichier généré par nos procédures est téléchargé dans un entrepôt de triplets en JAVA (notre choix s'est porté sur le serveur Sésame développé dans le cadre du projet *Aduna Open Source* [10]). Des requêtes SPARQL peuvent être saisies dans son interface web ou automatisées pour des besoins itératifs (par exemple des règles de validation).

5. Résultats

L'architecture cible que nous proposons (figure 3) repose sur une ontologie du domaine. Les difficultés d'une telle mise en place tiennent au fait que ce changement doit être réalisé sans perturber les évolutions et modifications

10. `http://www.openrdf.org/`

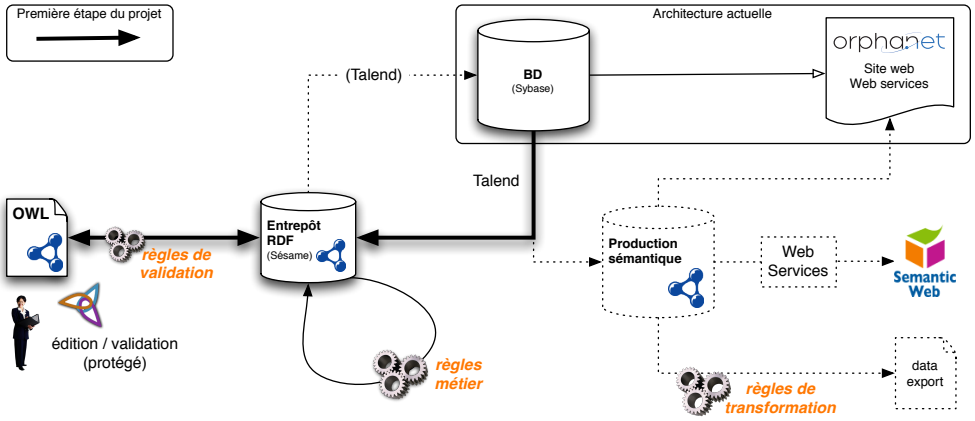

FIGURE 3: Première étape du projet d'intégration d'une architecture sémantique au sein du workflow de production d'ORPHANET.

en cours. Ainsi dans un premier temps nous intégrons notre représentation sémantique au processus déjà existant et fonctionnel. A terme, nous envisageons l'ensemble du processus au travers de représentations sémantiques.

Nous distinguons dans nos premiers résultats 1) l'ontologie produite et sa validation 2) les apports de cette ontologie dans notre architecture cible. Ces apports sont de deux types : d'une part au niveau des procédures éditoriales et d'autre part au niveau formel de représentation de la base de connaissance d'ORPHANET.

5.1. Une ontologie de domaine des maladies rares

Notre méthode a permis d'établir une procédure de construction d'une première version de l'ontologie. Cette procédure entièrement automatique et exécutable à la demande est encourageante pour envisager une mise en production : 7 minutes pour télécharger les fichiers extraits de la base Sybase, produire l'ontologie et la télécharger dans l'entrepôt RDF (l'étape de construction de l'ontologie elle-même durant moins d'une minute). Les caractéristiques de l'ontologie sont déterminées par des requêtes SPARQL sur l'entrepôt RDF (*cf.* tableau 2). Le détail des annotations de la classe *maladie* (*prefLabel*, *altLabel*, *definition* et les annotations de références externes) est résumé tableau 3.

La vérification des procédures d'exportation ETL confirme que les entrées de la bases sont bien toutes converties en classes (tableau 4) ; il existe plus

owl :Class	11 077	owl :Restriction	
owl :ObjectProperty	10	orpha :occasionalSignOf	10 530
owl :AnnotationProperty	29	orpha :frequentSignOf	12 384
AnnotationAssertion	179 567	orpha :veryFrequentSignOf	21 281
Classes polyparentales	2 843	orpha :geneOf	3 819

TABLEAU 2: Caractéristiques d'ONTOORPHA (412 138 triplets RDF).

	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

TABLEAU 3: Les annotations de la classe *maladie* dans ONTOORPHA.

de classes dans l'ontologie que d'entrées dans la base, cette différence s'explique par l'enrichissement du haut de l'ontologie (concepts de maladie rare, de maladie non rare,...) défini dans l'en-tête du fichier.

	maladies	signes	gène	hérarchies
En base (nombre d'identifiants)	7 161	1 360	2 415	106
ONTOORPHA (nombre de classes)	7 164	1 361	2 416	111

TABLEAU 4: Bilan de la construction d'ONTOORPHA.

5.2. Une évolution des méthodes éditoriales

Grâce à la séparation des connaissances du domaine des maladies rares d'une part et des règles métier d'autre part, les méthodes éditoriales sont plus simples. Toute connaissance qui peut être générée automatiquement par des règles n'est plus éditée manuellement.

L'éditeur d'ontologies PROTÉGÉ [11] est utilisé avec un *plugin* spécifique : ARCHONTE [12]. Nous avons comparé les procédures courantes et les procédures expérimentales [13] au sein de notre architecture cible pour deux cas d'uti-

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

12. Le *plugin* ARCHONTE a été développé dans notre unité de recherche par L. Mazuel, il correspond à l'intégration d'une partie des fonctionnalités du logiciel DOE dans PROTÉGÉ associée à une interface d'annotation gérant le multilinguisme et les labels SKOS (*definition*, *prefLabel* et *altLabel*).

13. L'accès sécurisé aux données et la gestion des droits ne sont pas détaillés ici car non implémentés dans nos procédures expérimentales.

lisation de routine de mise à jour des connaissances par les experts du domaine (tableau 5).

Cas 1 : mise à jour du libellé d'une maladie – PROCÉDURE COURANTE
– ouverture de l'interface proriétaire de mise à jour MAJOR – choix des données éditables (sélection du bouton "disease") – recherche par identifiant numérique de la maladie à éditer – édition du champ "libellé principal" FR/EN (édition des synonymes 2 écrans plus loin) – validation (et mise à jour de la base par un bouton "update")
Cas 1 : mise à jour du libellé d'une maladie – PROCÉDURE EXPÉRIMENTALE
– ouverture de PROTÉGÉ et du fichier RDF/XML généré : ontoOrphanet.owl – recherche de la maladie par label (champs recherche avec autocomplétion) – édition et visualisation de tous les libellés et synonymes par langue (*plugin* ARCHONTE) – sauvegarde du fichier
Cas 2 : déplacement d'une maladie dans une hiérarchie – PROCÉDURE COURANTE
– ouverture de l'outil d'extraction de vue de la base (PLATOR) – sélection (4 écrans) et téléchargement de la vue correspondant aux hiérarchies des maladies (fichier tabulé : PatPrt.Txt) – démarrage de l'application de visualisation des hiérarchies (ARBOR, produit une arborescence statique des hiérarchies à partir de la vue PatPrt.) – édition ligne par ligne du fichier PatPrt.Txt (plus de 40 000 lignes) dans un tableur : (1) invalidation des lignes à supprimer, (2) création de lignes [hiérarchie-maladie-maladie mère] pour toutes les hiérarchies concernées par les modifications – sauvegarde des lignes modifiées du fichier – connexion à l'outil de mise à jour de la base à partir de fichier (INJECTOR), sélection et envoi du fichier de modifications, puis réception d'un mail de rapport de mise à jour – nouvelle extraction du fichier PatPrt.Txt (PLATOR) et réouverture d'ARBOR pour une visualisation des modifications dans la base de données.
Cas 2 : déplacement d'une maladie dans une hiérarchie – PROCÉDURE EXPÉRIMENTALE
– ouverture de PROTÉGÉ et du fichier RDF/XML généré : ontoOrphanet.owl – recherche de la maladie à déplacer par son label – déplacement par glisser-déposer (*drag'n drop* de la maladie dans l'arborescence, avec visualisation continue de l'évolution de l'ensemble des hierarchies – sauvegarde du fichier

TABLEAU 5: Procédures courantes et expérimentales de mise à jour.

5.3. Procédures de validation d'intégrité

Grâce aux définitions formelles du modèle des maladies rares contenues dans l'ontologie, il devient aisé de construire des règles de validation d'intégrité. Les contraintes exprimées dans l'ontologie sont des premières vérifications d'intégrité, par exemple la relation *signOf* s'applique par définition à un

0	maladie(s) sans libellé
1	mode de transmission non lié à une maladie
4	signe(s) non liés à une maladie
20	gène(s) non liés à une maladie
2 494	maladie(s) génétique(s) sans mode de transmission (sur un total de 5272 maladies génétiques rares)

TABLEAU 6: Exemple d'audit de la base de connaissance (SPARQL).

signe et a pour valeur un objet de type *maladie* ; si des faits dans la base de connaissances ne respectent pas cette définition, une erreur est générée. Les règles de validation peuvent donc s'appuyer sur l'ontologie, par exemple ses relations transitives pour prendre en compte l'ensemble des maladies avec un signe donné. Ainsi s'il apparait une nouvelle maladie avec ce signe, la règle s'appliquera dynamiquement à elle.

Nous utilisons le langage de requêtes SPARQL pour décrire nos règles de validation. Le choix de ce langage est motivé par son adoption par la communauté web sémantique, la disponibilité des implémentations et la richesse des fonctionnalités qu'il propose (Polleres *et al.*, 2007). Un exemple d'audit par requêtes SPARQL (*cf.* tableau 6) fournit une vue de l'état de la base de connaissance : toutes les maladies ont un libellé, il existe des signes, des gènes et un mode de transmission non liés à une maladie, et le mode de transmission de la moitié des maladies génétiques rares n'est pas renseigné. Le tableau 3 fournit une vue intéressante de l'avancement des traductions des annotations des maladies.

6. Discussion et perspectives

6.1. Évolution des méthodes éditoriales et outils d'édition

Les procédures pour les deux cas étudiés ont été simplifiées par les outils expérimentaux. L'amélioration principale est la visibilité des modifications en cours d'édition (en particulier pour les hiérarchies). La diminution des erreurs lors de l'édition manuelle des lignes du tableur n'est pas quantifiée dans ce travail, mais elle semble acquise. Le *plugin* de gestion de la terminologie est également plus simple d'utilisation. Le choix initial de PROTÉGÉ (éditeur d'ontologie) permet donc des avancées notables. D'autres approches, en particulier l'utilisation d'éditeurs d'entrepôt couplé à des outils de visualisation

(*e.g.* ALLEGROGRAPH-GRUFF [14]) sont à explorer par la suite car ils permettraient une meilleure intégration au sein de notre architecture cible.

6.2. OWL-DL et annotations : atteinte des limites d'expressivité ?

OWL-DL nous semble adapté pour supporter les procédures d'édition de l'ontologie dans un outil comme PROTÉGÉ. Cependant le choix des annotations pour certaines caractéristiques comme la prévalence ou l'âge d'apparition n'est pas satifaisant car il limite les possibilités de raisonnement sur ces carcatéristiques (par exemple, une maladie fille ne peut avoir une prévalence supérieure à celle de sa maladie mère). Dans ce contexte, les deux approches possibles sont 1) l'utilisation de règles en dehors de l'ontologie pour contrôler ces annotations ou 2) l'utilisation d'un formalisme OWL-*full* en créant des classes et des instances correspondant à ces annotations.

Dans une seconde version de l'ontologie, nous envisageons de créer des classes pour les différents types de maladies [15] comme les "syndromes malformatifs" ou les "groupements de maladies". La notion de type de maladie, nécessaire à l'expression d'un grand nombre de règles métier, permettra également de redéfinir certaines annotations (*e.g.* le "mode de transmission" (orpha :inheritance)) comme des classes. La règle *une maladie du groupe maladie génétique doit avoir un mode de transmission* pourra par exemple s'appuyer sur ce typage.

6.3. Enjeux opérationnels

Nous espérons, grâce à l'apport de langages et de formalismes plus expressifs, améliorer la qualité des données (connaissances) que propose ORPHANET en France et en Europe, mais aussi la flexibilité de l'outil de production actuel (tableurs + interfaces + base de données relationnelle). Nous pensons qu'il est nécessaire de réorganiser l'architecture de l'information au sein même d'ORPHANET en utilisant par exemple des langages de représentation appropriés en fonction de la nature de la donnée traitée ; la séparation de l'information liée aux concepts et aux classifications (OWL), la gestion des termes (SKOS) et la gestion des règles métier (N3) permettront de gagner en flexibilité, en interprétation, mais faciliteront aussi l'extraction et la

14. `http://www.franz.com/agraph/`

15. Les types sont actuellement en cours d'édition par les experts du domaine sous forme d'annotations (orpha :patType, *cf.* figure 2)

maintenance des données.

Nous devons proposer une architecture de l'information qui soit compatible avec un service de production où la plage de fonctionnement du service est continue et où plusieurs utilisateurs à travers l'Europe peuvent agir sur le contenu de la base d'Orphanet. Nous observons dès aujourd'hui les limites d'outils tels que Sesame et Protégé pour la gestion des droits utilisateur ou des accès multi-poste. Aussi, dans un cadre hypothétique de remplacement de la base de données ORPHANET (Sybase) par un entrepôt RDF, une gestion de la sécurité d'accès et des droits doit être assurée. De plus, des mesures de montée en charge cohérentes avec la fréquentation du site web devront être effectuées.

6.4. Partage des connaissances formalisées

L'automatisation possible de procédures d'alignement à partir des annotations des concepts de l'ontologie par des références externes constitue un apport pour une intégration à des portails de partage d'ontologies biomédicales (type BioPortal [16]). Les connaissances des maladies rares d'ORPHANET pourront être utilisées dans le portail web mais aussi disposées sur un nouveau serveur de triplets RDF ouvrant des possibilités de nouveaux web services. Par des règles de transformation et d'export SPARQL, une génération de fichiers d'export à la demande dans les formats du web sémantique est envisageable. Ces règles de transformation utilisant la définition formelle de l'ontologie présentent des avantages certains : facilité de maintenance et possibilités d'évolutivité. En effet ces règles reposant sur la sémantique vont évoluer dynamiquement à mesure que le contenu change.

7. Conclusion

La première étape du projet ORPHAONTO a produit des résultats positifs ; une première mouture valide de l'ontologie des maladies rares formalisée en OWL-DL permet une évolution des méthodes éditoriales et une approche nouvelle de l'audit des bases de connaissances d'ORPHANET. La limite de l'expressivité du formalisme OWL-DL nous semble cependant atteinte pour représenter l'ensemble du domaine, et le choix d'OWL-*full* paraît envisageable. La formalisation des règles de validation et des règles métiers, le développement de prototypes de mise à jour de la base de donnée relationnelle et la

16. `http://bioportal.bioontology.org/`

prise en compte des enjeux opérationnels constituent les prochaines étapes de ce projet de recherche.

Références

AYMÉ S. (2002). Orphanet : serveur d'information sur les maladies rares et les médicaments orphelins, INSERM SC11. http ://www.orpha.net/.

BERNERS-LEE T., FIELDING R. & MASINTER L. (2005). RFC 3986/STD 0066 - uniform resource identifier (URI) : generic syntax.

DHOMBRES F., CHARLET J., JOUANNIC J., MAZUEL L. & 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.*, p. 1–12, Lisbonne, Portugal : Springer-Verlag.

DHOMBRES F., JOUANNIC J., JAULENT M. & 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*, p. 1–12, Nîmes, France. [http ://hal.archives-ouvertes.fr/hal-00487736/fr/].

KRIVINE S., NOBÉCOURT J., SOUALMIA L., CERBAH F. & DUCLOS C. (2009). Construction automatique d'ontologie à partir de bases de données relationnelles : application au médicament dans le domaine de la pharmacovigilance. In F. GANDON, Ed., *20 th French Knowledge Engineering Workshop*, p. 73–84, Hammamet, Tunisie.

MILES A. & BECHHOFER S. (2009). SKOS simple knowledge organization system namespace document - HTML variant, 18 august 2009 recommendation edition.

O'CONNOR M. J. & DAS A. (2010). Semantic reasoning with xml-based biomedical information models. In C. SAFRAN, H. F. MARIN & S. R. RETI, Eds., *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*, p. 986–90, Cape Town, South Africa : IOS Press.

POLLERES A., SCHARFFE F. & SCHINDLAUER R. (2007). SPARQL++ for mapping between RDF vocabularies. In *6th International Conference on Ontologies, Databases, and Applications of Semantics (ODBASE 2007)*, p. 4803, Vilamoura, Portugal : Springer-Verlag.