Un outil de visualisation de classifications et d'intégration de données phénotypiques et génétiques pour faciliter le codage des maladies rares

Coding rare diseases in health information systems: a tool for visualizing classifications and integrating phenotypic and genetic data

Rémy Choquet[1,2], Yannick Fonjallaz[1], Albane de Carrara[1], Meriem Maaroufi[1,2], Pierre-Yves Vandenbussche[2], Ferdinand Dhombres[2,3*], Paul Landais[1,4*]

[1]*Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, Assistance Publique Hôpitaux de Paris, Paris, France.* [2]*INSERM, U1142, LIMICS, Sorbonne Universités, UPMC Paris 6, et Université Paris 13, Villetaneuse, France.* [3]*Unité de Diagnostic Prénatal et d'Échographie et Centre Pluridisciplinaire de l'Est Parisien, hôpital Armand Trousseau, UPMC et AP-HP, Paris, France.* [4]*Montpellier1University, EA2415 & BESPIM, Carémeau University Hospital, Nîmes, France. *Ces deux auteurs ont contribué de manière analogue à ce travail.*

Résumé

Pour diagnostiquer et coder une maladie rare (MR), il est nécessaire de caractériser des informations multiples, y compris génotypique et phénotypique. Aujourd'hui, un obstacle à l'activité de codage des diagnostics MR est le manque de consolidation de ces informations dispersées dans différentes bases de connaissances telles qu'Orphanet, OMIM ou HPO. Nous avons développé l'application web LORD (Linking Opendata for Rare Diseases), en offrant une vue intégrée de 8.336 maladies rares et groupes de maladies liés à plus de 12.500 signes et 3.000 gènes. LORD offre une fonctionnalité de navigation contextuelle (médicale) dans les relations entre les groupes de maladies, les maladies, les signes et les gènes. Il est développé sur un ensemble d'interfaces de programmation d'application (APIs) permettant son intégration au sein de systèmes d'informations propriétaires. Il est dédié aux 131 centres français de référence MR et aux 501 centres de compétences, mais aussi aux départements d'information médicale pour coder les diagnostics MR dans les systèmes d'information de santé.

Abstract

Establishing the diagnosis and coding for a rare disease (RD) needs to characterize multiple information including patient's phenotype and genotype. A major barrier to coding is a lack of consolidation of such information, scattered in several resources such as Orphanet, OMIM or HPO. We developed a web portal, Linking Open data for RD (LORD), offering an integrated view of 8,336 RDs linked to more than 12,500 signs and 3,000 genes. It allows navigating through the relationships between diseases, signs and genes, and provides Application Programming Interfaces for its integration in information systems (IS). LORD is dedicated to the 131 French RD reference centers and 501 competence centers, for coding RD diagnoses in the health IS.

Articles courts des 15es Journées francophones d'informatique médicale, JFIM 2014, pages 198–203
Fès, Maroc, 12–13 juin 2014

Mots-clés : Ontologies des maladies rares ; Big Data ; Codage diagnostique ; Web sémantique ; Informatique médicale ; Services web

Keywords: *Rare Diseases Ontologies ; Big data ; Coding support ; Semantic web ; Medical Informatics ; Web services*

1 Introduction et état de l'art

Les systèmes de codage (thésaurus, classifications, terminologies, vocabulaires contrôlés) sont utilisés pour permettre l'entrée de données contrôlées dans les systèmes d'information de santé afin d'en faciliter l'analyse pour l'amélioration de la qualité des soins [1], l'élaboration d'études épidémiologiques [2] ou la comparabilité des protocoles de recherche [3]. Historiquement, la classification internationale des maladies a été utilisée pour mesurer la mortalité dans les populations. Son utilisation a été progressivement adaptée pour mesurer la morbidité puis généralisée aux systèmes d'information hospitaliers non sans quelques difficultés [4]. Les causes en sont nombreuses comme notamment l'utilisation de structures de classification à base d'arbres simples qui ne permettent pas de modéliser les différents usages possibles de ces classifications. L'utilisation d'arbres simples peut en effet nuire à la capacité expressive des éléments d'un domaine et n'est pas adaptée à la formalisation du sens médical [5] de l'information codée.

Près de 80% des maladies rares (MR) ont une origine génétique. Actuellement, 3.000 MR parmi plus de 8.000 ont un gène identifié. Coder un cas de maladie rare est une tâche qui peut être complexe, et qui nécessite un niveau de connaissances médicales spécifiques provenant de diverses sources d'information. Par exemple, la Cystinose est une maladie métabolique accompagnée d'une atteinte cornéenne et rétinienne, mais aussi d'une néphropathie. Si la première information est informative pour un ophtalmologiste, la deuxième l'est pour un néphrologue. La cystinose peut atteindre d'autres organes et présenter des signes cliniques complexes : photophobie, hypothyroïdie, troubles de la croissance, atteinte neurologique,.. Actuellement, une ressource de représentation des maladies rares est désignée pour coder les diagnostics de MR : Orphanet. Elle est actuellement utilisée par 44% des sites nationaux de prise en charge MR (réseau des centres de référence MR labellisés par le ministère de la santé) [6]. Cependant, les diagnostics « non confirmés » représentent 1 cas sur 4 (17% indéterminés, 9% non encore classables). Pour coder les patients atteints de MR dans les systèmes d'information de santé (dossiers médicaux, registres ou cohortes), plusieurs questions ont été examinées par des spécialistes et des cliniciens de l'information en santé : i) le manque d'information sur les maladies rares dans la CIM-10 (moins de 400 maladies rares représentées), ii) le recours souvent nécessaire à plusieurs ressources pour coder un diagnostic MR, comme par exemple de l'information phénotypique et/ou génotypique, iii) la navigation dans 8.336 MR et groupes de MR qui ne peut se faire à travers une liste à plat ou à travers un simple algorithme de recherche, iv) enfin, pour des maladies souvent multi-systémiques, la limite des représentations sous forme d'arbres monoparentaux.

En termes de ressources d'aide au codage pertinentes pour les MR, Diseasecard [7] vise à relier chaque MR aux bases de connaissances connues. Cependant, elle recueille des données à partir de toutes ses sources au sein d'une même interface utilisateur, et ne propose pas de navigation contextuelle dans sa base de connaissances. HeTOP [8] relie de nombreuses bases de connaissances médicales, y compris les principales pour les MR. Cependant, les données d'une seule source peuvent être affichées à la fois et l'application ne propose pas de navigation contextualisée dans l'arborescence des MR. Nous avons donc défini et construit une nouvelle application web pour aider les cliniciens et les spécialistes de l'information de santé à naviguer dans des graphes représentant les diagnostics MR proposés par Orphanet.

2 Matériel et méthodes

2.1 Les sources de données

Au cours de la dernière décennie, de nombreuses ressources terminologiques ou ontologiques (RTO) ont été développées. Peu d'outils sont proposés pour utiliser leur complète représentativité de l'information médicale dans un contexte de codage. Dans le contexte des MR, nous avons identifié quatre ressources d'intérêt avec l'aide d'experts. (i) Orphanet [9] a été notre principale ressource pour naviguer dans les maladies. Cette unité de service INSERM diffuse, via son site OrphaData (www.orphadata.org), 30 classifications MR, ainsi que les synonymes, les signes cliniques, les gènes et les liens à d'autres ressources externes. (ii) OMIM [10] donne accès à une information textuelle importante sur les maladies, les gènes ou les mutations fournis dans des fichiers plats. (iii) HPO [11] propose un grand nombre de phénotypes qui peuvent être utilisés pour le diagnostic clinique des maladies génétiques. Son ontologie est disponible à partir de leur site web dans différents formats. (iv) HRDO [10] enfin, une ontologie dédiée aux MR, propose un méta-modèle pour les données Orphanet. HPO, Orphanet et OMIM mettent à jour leurs données en continu et de manière asynchrone. Leur utilisation comme source primaire pour le codage est un enjeu majeur pour établir des registres de données de patients homogènes et de qualité.

2.2 Architecture

Afin de réunir les informations sur les maladies, les classifications, les gènes et les signes cliniques dans un même espace d'information pour les cliniciens, nous avons tout d'abord effectué une intégration des formats des sources (XML, CSV, OWL) en RDF dans un triplestore Virtuoso. Habituellement, les applications basées sur des données RDF sémantiques sont assez lentes, en raison de la quantité de requêtes SPARQL nécessaires. Donc nous avons choisi de préparer les données à l'affichage en générant chaque maladie et ses données associées en objet JSON que nous stockons dans une collection MongoDB (NoSQL). Enfin, nous avons construit une application web pour aider l'utilisateur à naviguer dans la base de connaissances (figure 1).

2.3 Intégration de données médicales et génétiques dans un entrepôt RDF et définition des objets JSON

Pour construire une base de connaissances cohérente intégrant diverses terminologies et classifications, l'intégration des données dans un espace commun d'information RDF (Virtuoso triplestore) était nécessaire. La source d'information sur laquelle nous avons intégré d'autres données externes (HPO et OMIM) était Orphanet. L'utilisation de SPARQL pour obtenir des triplets de maladies n'a pas été efficace dans un cadre de navigation web. Nous avons donc utilisé un modèle d'application web moderne qui échange des données avec le serveur via des services web, par opposition au rechargement de page classique. Le format de données commun utilisé pour ces échanges était JSON.

Pour la conception de l'interface utilisateur, nous avons d'abord conçu une interface web maquette en collaboration avec des experts MR afin d'identifier les données appropriées à des fins de codage de diagnostic. Les informations et les caractéristiques suivantes ont été retenues : (i) le nom et OrphaCode de la maladie, (ii) les liens externes par source (1-n relations possibles), (iii) les classifications auxquelles la maladie est liée, (iv) un graphique pour naviguer dans la classification actuelle (en raison de la nature multi-parentale du graphique, seuls 3 niveaux de navigation - les parents, la maladie, les enfants - sont conservés) ; (v) les signes ou groupes de signes liés à la

maladie actuelle ; (vi) des informations générales sur la maladie (synonymes, la prévalence, l'héritage, l'âge de début et de décès), (vii) la liste des gènes liés à la maladie à partir de différentes sources le cas échéant (OMIM et Orphanet), (viii) le synopsis clinique ; (ix) les autres informations textuelles connexes lorsqu'elles sont disponibles avec la possibilité de choisir l'affichage de chaque section.

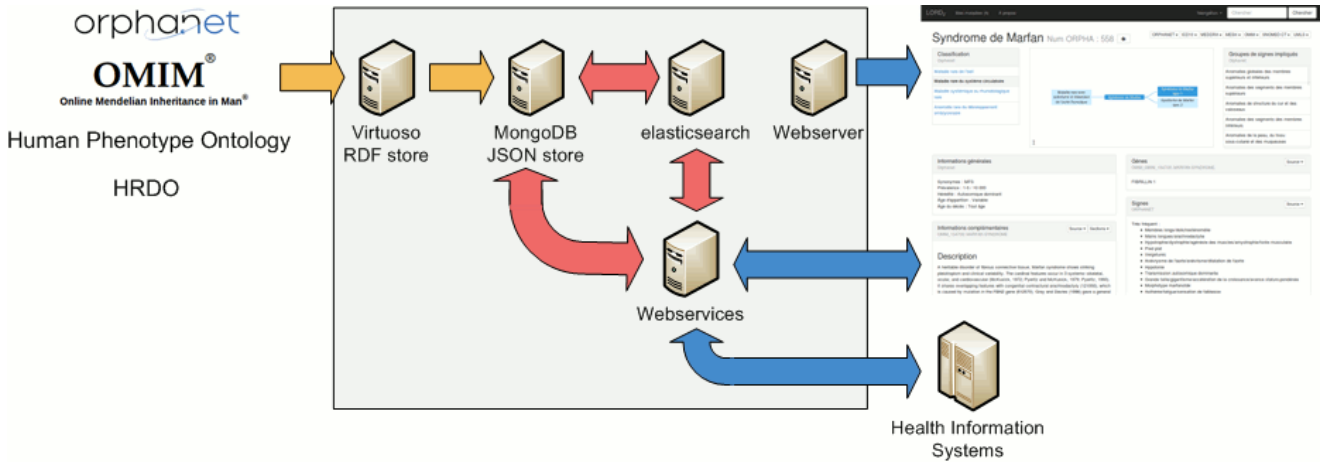

Figure 1 : Architecture de l'application LORD

Cette maquette nous a également permis de définir la structure de l'objet JSON que nous avons utilisé pour transmettre les données des services web. Nous avons créé un script Ruby qui a permis d'intégrer des objets JSON dans une base de données MongoDB.

Catégories	**Information**	**Orphanet**	**OMIM**	**HPO**
Disorders	*PrefLabels*	8336	7666	0
	Synonyms	9750	8650	0
	Description	472	2604	0
	Diagnosis	0	412	0
	Clinical features	0	4783	0
Signs	*PrefLabels*	1360	0	10152
	Synonyms	0	0	6424
	Descriptions	0	0	6844
	Broaders	1310	0	13371
Genes	*PrefLabels*	3045	14419	0
	Synonyms	7168	21275	0
	Symbols	3045	23429	0
	Methods	0	20639	0

Tableau 1 : Nombre de triplets intégrés dans l'entrepôt RDF Virtuoso ; PrefLabels *: nom le plus répandu d'une entité ;* Synonyms *: noms alternatifs d'une entité ;* Description *: définition générale ;* Diagnosis *: description textuelle détaillée du diagnostic ;* Clinical features *: description textuelle des signes rencontrés ;* Broaders *: lien vers un élément parent ;* Symbols *: nom abrégé d'un gène ;* Methods *: caractérisation de la mutation génétique.*

3 Résultats

L'application web est accessible à l'adresse suivante : http://lord.bndmr.fr. Nous avons intégré des ensembles de données d'OMIM, HPO, HRDO et d'Orphanet au 1er Janvier 2014. Nous intégrons

des données annuelles dans notre base de connaissances. Le volume et la nature des données intégrées sont décrits dans le tableau suivant.

Un serveur elasticsearch, relié à la collection MongoDB, fournit le moteur de recherche rapide et paramétrable adapté à une recherche efficace. La taille de l'ensemble de données RDF est de 126 Mo, alors que la taille de l'ensemble de données JSON est 63,55 Mo. Nous avons effectué des tests de performance sur le même matériel (machine virtuelle avec 2 cores Opteron 6132 et 8Go de RAM). Obtenir les données nécessaires à l'affichage depuis l'entrepôt RDF prend 2 secondes. Ce délai a été réduit à 1 seconde en utilisant des technologies de mise en cache. Obtenir les mêmes informations depuis MongoDB prend 0,03 seconde. Nous avons également défini un ensemble de services web qui sont utilisés par notre application web. Ces services sont accessibles au public. Ils permettent d'obtenir : la description d'une maladie, la liste des classifications et la recherche de maladies.

4 Conclusion

L'identification des patients atteints de MR est un objectif du plan national maladies rares 2. L'utilisation d'une classification telle qu'Orphanet, qui vise à représenter les maladies rares pour diverses utilisations (codage, diagnostic clinique, registres, cohortes) peut aider à répondre à cet objectif mais elle pose des problèmes de représentativité pour les codeurs. Les cliniciens ont exprimé un besoin spécifique pour avoir à la fois des données phénotypiques et génotypiques, au même niveau, afin d'attribuer le code plus approprié au diagnostic retenu. A titre d'exemple, la *néphronophtise*, une maladie rénale kystique de la médullaire, dispose de 11 formes génétiques distinctes (http://lord.bndmr.fr/#disorders/655). La *Cystinose*, une maladie de surcharge lysosomale, peut être considérée comme une maladie rénale rare par le néphrologue ou une MR de l'œil par l'ophtalmologiste (http://lord.bndmr.fr/#disorders/213). Il pourrait toutefois être intéressant de pouvoir sélectionner plus d'une spécialité médicale. Les bases de connaissances sur les MR étant en constante évolution, l'arborescence des maladies change chaque jour, ce qui peut compliquer la navigation et les habitudes des médecins. C'est pourquoi une version est créée chaque année, qui peut dès lors être utilisée comme référence. Un processus mensuel d'intégration des nouvelles maladies est en cours d'étude. L'application est étudiée pour permettre l'ajout de nouvelles sources de données au cours de l'année. Permettre la navigation sur des grands graphes sémantiques pose des problèmes de performances sur triplestores RDF classiques. Nous avons utilisé des technologies Big Data pour la persistance des données et permettre une navigation temps réel sur notre ressource intégrant sémantiquement des données génétiques et phénotypiques. Cette application a été requise dans le cadre du 2e Plan National Maladies Rares afin d'aider les centres de référence et de compétences MR à coder avec un code Orphanet leurs patients avec le même système national de codage dans les différents systèmes d'information de santé. L'application a d'ores et déjà été traduite en anglais pour satisfaire à des opportunités de collaboration européennes sur le sujet.

Remerciements

Nous remercions chaleureusement C. Messiaen, J.-P. Necker et C. Angin pour leur aide. Ce travail a été financé par le ministère de la Santé.

Références

[1] JM. Januel et al., “ICD-10 adaptation of 15 Agency for Healthcare Research and Quality patient safety indicators.” *Rev Epidemiol Sante Publique*, vol. 59, no. 5, pp. 341–50, 2011.

[2] E. R. Dubberke, A. M. Butler, H. A. Nyazee, K. A. Reske, D. S. Yokoe, J. Mayer, J. E. Mangino, Y. M. Khan, and V. J. Fraser, “The impact of ICD-9-CM code rank order on the estimated prevalence of Clostridium difficile infections.,” *Clin. Infect. Dis.*, vol. 53, no. 1, pp. 20–5, 2011.

[3] P. Bruland, B. Breil, F. Fritz, and M. Dugas, “Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM.,” *Stud Health Technol. Inform.*, vol. 180, pp. 564–8, 2012.

[4] E. J. Picardi and J. B. Peoples, “Mesenteric venous thrombosis: ten year record review and evaluation of difficulties with the ICD coding system.,” *S D J Med*, vol. 44, no. 2, pp. 33–7, 1991.

[5] A. Sollie et al., “A new coding system for metabolic disorders demonstrates gaps in the international disease classifications ICD-10 and SNOMED-CT, which can be barriers to genotype-phenotype data sharing.,” *Hum. Mutat.*, vol. 34, no. 7, pp. 967–73, 2013.

[6] P. Landais et al., “CEMARA an information system for rare diseases.,” *Stud Health Technol Inform*, vol. 160, no. Pt 1, pp. 481–5, 2010.

[7] P. Lopes and J. L. Oliveira, “An innovative portal for rare genetic diseases research: The semantic Diseasecard,” *J Biomed Inform*, vol. 46, no 6, pp.1108-15, 2013.

[8] J. Grosjean, T. Merabti, N. Griffon, B. Dahamna, and S. J. Darmoni, “Teaching medicine with a terminology/ontology portal.,” *Stud Health Technol Inform*, vol. 180, pp. 949–53, 2012.

[9] H. Nabarette, D. Oziel, B. Urbero, N. Maxime, and S. Aymé, “Utilisation d’un annuaire des services spécialisés et orientation dans le système de soins: l’exemple d’Orphanet dans les maladies rares,” *Rev Epidemiol Sante Publique*, vol. 54, no. 1, pp. 41–53, 2006.

[10] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.,” *Nucleic Acids Res*, vol. 33, pp. D514–7, 2005.

[11] S. Köhler et al., “The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data.,” *Nucleic Acids Res.*, vol. 42, no. 1, pp. D966–74, 2014.

Adresse de correspondance

Rémy Choquet, PhD, Hôpital Necker Enfants Malades, Banque Nationale de Données Maladies Rares, Bâtiment Imagine, 149 rue de Sèvres, 75743 Paris. http://www.bndmr.fr