LORD: a phenotype-genotype semantically integrated biomedical data tool to support rare disease diagnosis coding in health information systems

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

[1]Banque Nationale de Données Maladies Rares, Hôpital Necker Enfants Malades, APHP, F-75015, Paris, France; [2]INSERM, U1142, LIMICS, F-75006, Paris, France; [3]Service de Médecine Fœtale, Hôpital Armand Trousseau, APHP, UPMC, Paris, France; [4]Montpellier1 University, EA2415 & BESPIM, University Hospital, Nîmes, France; *both authors equally contributed to the work

Abstract

Characterizing a rare disease diagnosis for a given patient is often made through expert's networks. It is a complex task that could evolve over time depending on the natural history of the disease and the evolution of the scientific knowledge. Most rare diseases have genetic causes and recent improvements of sequencing techniques contribute to the discovery of many new diseases every year. Diagnosis coding in the rare disease field requires data from multiple knowledge bases to be aggregated in order to offer the clinician a global information space from possible diagnosis to clinical signs (phenotypes) and known genetic mutations (genotype). Nowadays, the major barrier to the coding activity is the lack of consolidation of such information scattered in different thesaurus such as Orphanet, OMIM or HPO. The Linking Open data for Rare Diseases (LORD) web portal we developed stands as the first attempt to fill this gap by offering an integrated view of 8,400 rare diseases linked to more than 14,500 signs and 3,270 genes. The application provides a browsing feature to navigate through the relationships between diseases, signs and genes, and some Application Programming Interfaces to help its integration in health information systems in routine.

Introduction

Eighty percent of Rare Diseases (RD) have a genetic origin. Presently, 3,000 among more than 7,000 rare diseases have an identified gene.[1] Moreover, clinical signs are often used to characterize RD.[2] Even when a consensual disease classification is used in a registry for RD centers,[3] the "not confirmed" category represents 1 out of 4 cases (17% undetermined, 9% could not be classified yet). The identification of patients with rare diseases at international level is a key requirement to accelerate patient recruitment for clinical trials and cohorts. This work is proposed within the framework of the second French National Plan for Rare Diseases (2011-2016) where the French National Database for Rare Diseases infrastructure (BNDMR)[4] is defined. The BNDMR is a priority project of the National Plan and aims at gathering a common minimal data set for all rare diseases patients at nationwide scale[5] to facilitate patient recruitment in clinical trials and in cohorts as well as public health studies. A common terminology is required to register patient diagnosis and enable statistical use of the data. The French Ministry of Health encourages the use of Orphanet[6] for national coding of RD while the European Commission recently decided to recommend Orphanet as an European terminology for RD diagnosis coding.[7]

In order to properly code rare disease patients into health information systems (medical records, registries or cohorts), several issues have been discussed by health information specialists and clinicians: (i) poor information on rare disease is available in standard morbidity classifications (ICD-9 to ICD-10); the ICD classification is often used for reimbursement purposes in pay-for-performance systems, and is not well adapted to describe patient's clinical status, (ii) to code a patient diagnosis, there is a possible need to gather information from other knowledge bases, such as genetic and phenotype information, (iii) navigating within more than 7,000 RD cannot be done through a tree navigation or a list navigation. Search algorithms are nice but context dependent navigation is a key feature (a context can be a medical field), (iv) RD are often multi-systemic, thus mono-parental classifications such as ICD might not be adapted.

Coding systems (thesaurus, classifications, terminologies, controlled vocabularies) are used to enable controlled data entry in health information systems in order to facilitate data analysis for quality of care,[8] epidemiology[9] or research.[10] Historically, the International Classification of Diseases (ICD) was used to measure mortality in

populations. Its usage was progressively redefined to measure morbidity and was then integrated within global hospital information systems with some difficulties.[11] In the rare disease field this classification does not provide us with the necessary RD completeness (only 300 rare diseases are described) and when described, rare diseases are not necessarily used as the primary or secondary diagnosis since the visit of the patient might be related to a comorbidity (e.g. respiratory distress for a Cystic fibrosis patient). There are few resources available that can help in classifying RD patients: Orphanet, OMIM,[12] HPO,[13], the GARD disease nomenclature (**https://rarediseases.info.nih.gov/gard**) or SNOMED CT. These resources do not present the same information type or completeness but they could be used together to specify patient rare conditions with a relatively good phenotypic or genotypic precision.

In this paper, we will not discuss the adequacy of using Orphanet as a coding system to record morbidity in the rare diseases field nor the curation made by Orphanet, OMIM and HPO to interlink their resources. We will focus on implementing a tool that could use the Orphanet proposed views (multi-parental and multi-classified diseases) to help guiding professionals to the right disease (medical coding assistance). Besides, we will propose to enrich each of the Orphanet disease with genetic and phenotypic information available from OMIM and HPO to provide a richer presentation of the disease. Many terminology sources are available, and it is now crucial to be able to bring data from different sources together into one application to help clinicians in their coding task. Available web applications do provide parts of the information for a given disease (http://www.orpha.net, http://omim.org). Few projects offer to link the diseases all together (Diseasecard,[14] HeTOP,[15] or Orpha.net) but they do not provide the actual information from all sources into the same web interface. We propose in this paper a new generation application for rare diseases coding assistance based on semantic web/RDF technologies for the rich semantic linked data DB on one hand and NoSQL/JSON for the web services technologies access on the other hand.

Methods

The semantic integration of biomedical knowledge sources and datasets is a complex data engineering issue[16]. In order to enable the integration of different datasets, our architecture is based on a first semantic integration layer that deals with the heterogeneity of source formats as well as the semantic heterogeneity of proposed representations of diseases. We then defined a presentation layer composed of a NoSQL database well fit to represent JSON complex biomedical objects to be displayed through a Ruby on Rails web application. That architecture combines the use of semantic web technologies that have great expressivity capabilities (RDF, OWL) but have a footprint on performance for real-time applications, with the flexibility and performance of NoSQL technologies to provide the best user experience when navigating through complex biomedical data.

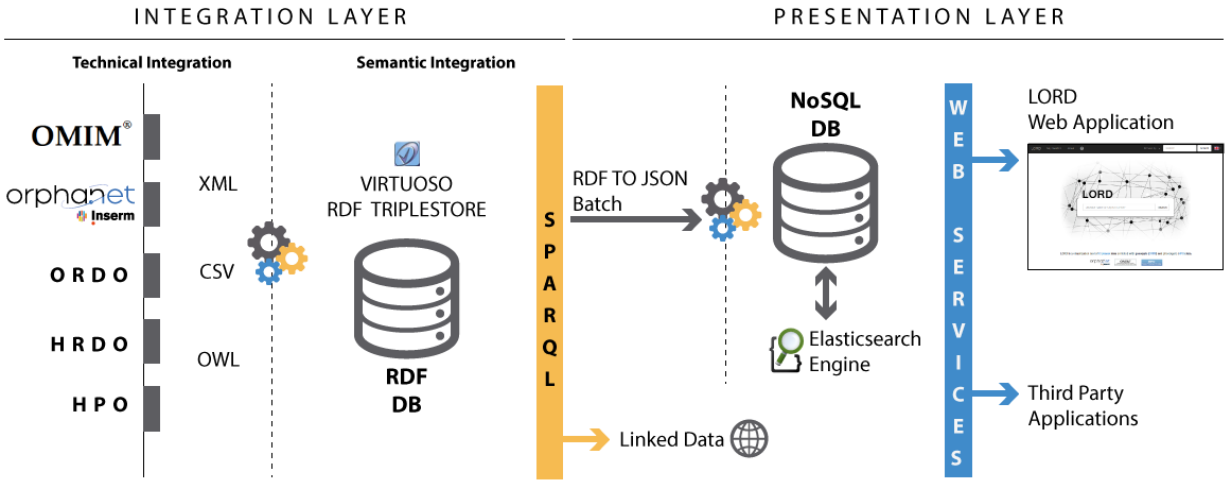

Figure 1. Genetic-phenotypic semantic database integration architecture for rare diseases. The first layer function deals with the technical and the semantic integration of various data sources. The second layer deals with presenting the data and navigating through views.

1. Integration layer

The integration of various externally curated multi-lingual biomedical datasets requires expressive models and semantic robustness to express exact, broader-than and narrower-than alignments as well as reasoning. We have

chosen to integrate the datasets into a RDF triple store (Openlink Virtuoso©) through a homemade integration layer (xml and java data integrators). Each data integrator was developed specifically for each data source as data structures and format of sources were heterogeneous. Orphanet data is released using xml based on a local model extraction without xsd, OMIM is available as a set of flat text files and HPO as an OBO or OWL file. Whilst the HPO file has a rich semantic web compatible serialization format, Orphanet and OMIM are not semantic web compatible, semantic links between concepts were thus rebuilt to be compliant with RDF. Table 1 reflects the volume of integrated triples in the RDF triple store.

Table 1. Dataset integrated triple volumes (at 1st January 2015)

Categories	Information	Orphanet	OMIM	HPO
Disorder	PrefLabels	8,489	7,666	4,170
	Synonyms	10,412	8,650	0
	Description	472	2,604	0
	Diagnosis	0	412	0
	Clinical features	0	4,783	0
Signs	PrefLabels	1,360	0	12,626
	Synonyms	0	0	6,424
	Description	0	0	6,844
	Broaders	1,310	0	13,371
Genes	PrefLabels	3,270	14,419	0
	Synonyms	7,168	21,275	0
	Symbol	3,270	23,429	0
	Method	0	20,639	0

Overlaps between these three datasets to represent disorders (diseases) were enumerated in the disease ontology.[17] The Orphanet source also offers cross referencing to HPO and OMIM and other datasets. They also produce more than 30 views (named diseases classifications by Orphanet) organized per rare disease groups (including 1 non-rare diseases group). Each disease group is organized into several entities such as categories of diseases, diseases, clinical subtypes, etiological subtypes, etc. These views are curated by Orphanet and will be used to implement the navigation functions of the LORD application. Rare diseases are numerous and can often have several medical expressions. For instance, patients diagnosed with *Fabry disease* (http://enlord.bndmr.fr/#disorders/324) can present several anomalies: renal, cardiac, neurologic, dermatologic, ophthalmic etc. The classification can also be used to make differential diagnosis, if the patient present a *Cataract associated with a metabolic disease* (http://enlord.bndmr.fr/#disorders/98644), then the patient could be suffering from *Fabry disease*, but also *Refsum disease*, *De Barsy syndrome*, *Galactosemia* or *Alpha-mannosidosis diseases* (given by Orphanet views). Searching from clinical signs is not supported in this version of the LORD application.

2. Presentation layer

Semantically enriched biomedical data stored in native RDF triplestore generally suffers from a performance footprint when queried on a large scale application with real-time navigation features on the web.[18,19] On the other hand, new generation web applications (Ruby on Rails, Backbone.js) can be easily built up using a full web service architecture managing JSON objects. JSON objects are similar to rich XML objects and can be stored in NoSQL databases such as MongoDB, allowing flexibility for model changes as well as best in class performance and scalability using sharding techniques if needed (data atomization over MongoDB servers and query balancing).

To generate JSON objects, SPARQL queries are processed on the RDF triplestore, and resulted JSON objects are stored into the NoSQL database in batch mode, every day. Such a process can be applied from the same RDF triplestore to numerous presentation databases for various projects if needed. Once integrated and formatted for presentation as JSON objects, the NoSQL database is indexed and searched through an Elasticsearch engine (http://www.elasticsearch.org) enabling auto-completion, multi-field, ranked searches as well as various possible optimizations. The search engine enables search over disorder names and synonyms, Orphanet numbers and genes. Given the nature of the Orphanet resource (multi-view and multi-parental), we pre-calculated all possible concept parents and sons to help with the navigation feature of the interactive graphical interface. Hence, when a user selects a disorder, he can view direct neighbors of the given disease.

The result is returned and displayed to the user in a graphical and textual interface. We defined a navigational area and a content area. Users can navigate through classifications, filter classifications and diseases, and add current disease to a disease 'basket' for exportation if the application is integrated with an EHR or a patient registration system. They have then access to all information available from Orphanet, HPO and OMIM for the selected disease on the same screen. The Human Rare Disease core Ontology (HRDO) provides a meta-model for this presentation layer implementation.[20]

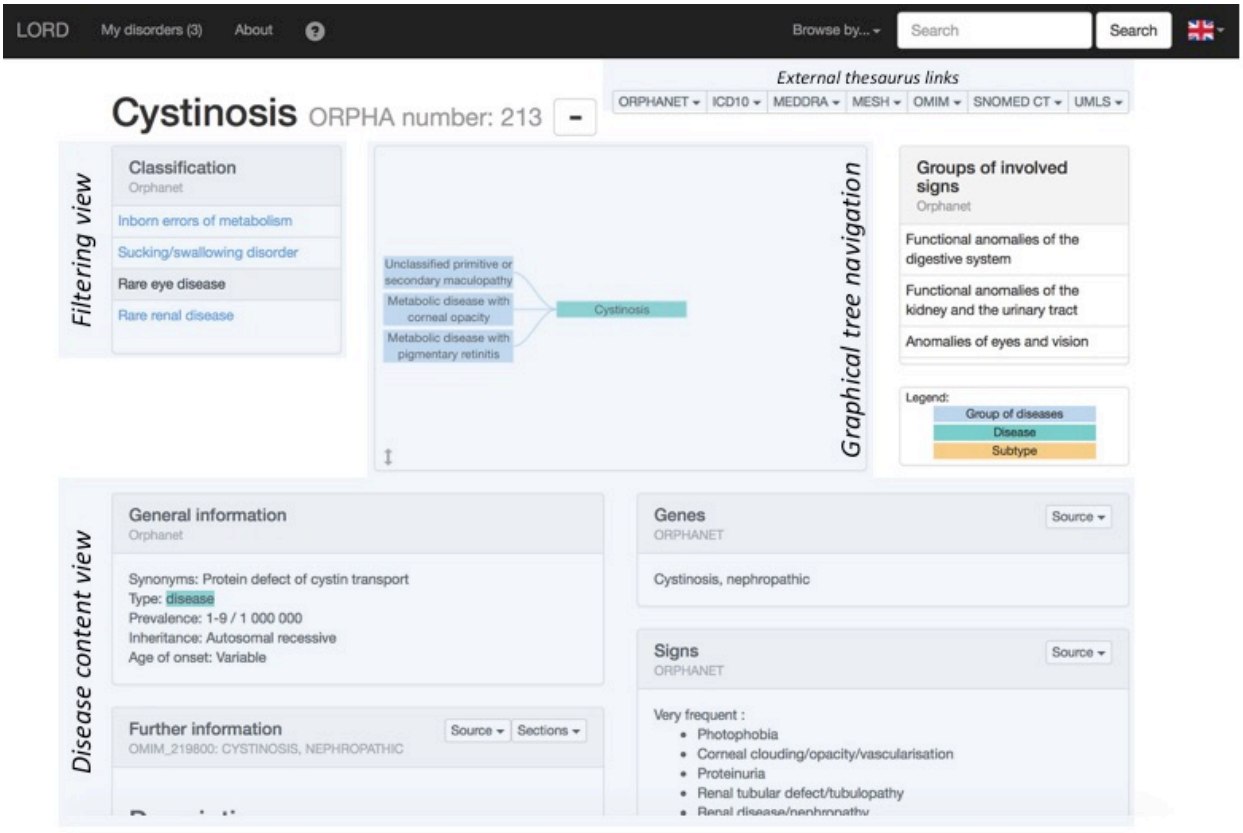

Figure 2. The disease interface of the LORD application. *The filtering view* gives access to filters based on medical specialities. *The graphical tree navigation* enables navigation through diseases, symptoms and group of diseases. *The external thesaurus links* allow the user to be redirected to source sites pages of diagnosis. *The disease content view* represents disease definition data from Orphanet, signs from HPO and clinical synopsis and genetic data from OMIM.

Results

One of the objectives of the project was to be able to create a common, curated and semantically rich knowledge repository to help managing the complexity and heterogeneity of biomedical databases. Data are openly available but with poor usability, as they are not integrated into a single application. Update frequencies are heterogeneous and as the sources are interlinking their datasets, they present some data in common. It is therefore necessary to build an integrated semantic database that can also serve as a national terminology server for several purposes: search terms, search based on clinical signs, disease complete data retrieval, integration within another application of the curated database. The RDF datastore can be queried over the semantic web SPARQL query language which is the most advanced query language enabling reasoning, inference and federation. It is compatible with Linked Data principles. Data can also be queried directly through a secured web service directly on the NoSQL database, depending on the use case of the user.

When connecting to the application (http://enlord.bndmr.fr) the user can browse the data through 2 modes. Either using the search function, or by browsing using proposed Orphanet views. An Elasticsearch engine that provides multi-criteria searches over several data sources powers the search function. Users can search through diseases preferred label, alternative label, gene or Orphanet number. The search engine enables word search regardless of the position of the word within the disease label. Given the length of some disease names, this feature is important. For instance, the *Anhidrotic ectodermal dysplasia - immunodeficiency - osteopetrosis – lymphedema* Orphanet disease (http://enlord.bndmr.fr/#disorders/69088) which is composed as a set of symptoms. By navigating backwards in the hierarchy to *Osteopetrosis* (http://enlord.bndmr.fr/#disorders/2781) the user can search over a limited set of diseases that share Osteopetrosis, if it is the observed symptom.

When users are navigating through diseases graphs, they can filter the displayed concepts. For example, while looking at *Cystinosis* information (http://enlord.bndmr.fr/#disorders/213/97966), the user can filter on Rare eye diseases (if he is an Ophthalmologist) and then navigate through symptoms that are related to his medical expertise: *Unclassified maculopathy*, *metabolic disease with corneal opacity* or *metabolic disease with pigmentary retinisis*. We should emphasis that the nature of the disease "groups" here differs from the previous example. While the first "groups" are related to symptoms of a disease, the seconds are diseases classifiers.

The user then can have access to textual information from OMIM and HPO for a given disease and epidemiological information from Orphanet. As relations between data sources can be one to many (1 Orphanet disease for n OMIM entries) then we have developed an OMIM entry selector to help user navigation and data retrieval. From the web service standpoint, all data of all OMIM entries are gathered in the same JSON object for an individual Orphanet disease.

Discussion

We defined and built a new generation web application to help clinicians and health information specialists to navigate in the Orphanet rare diseases for diagnosis coding. We first integrated and curated various terminologies and classifications (Orphanet, HPO, OMIM, HRDO) related to our domain within a semantic database adapted to reasoning and compatible with linked data principles. Once curated and qualified, the semantic biomedical data is transformed into JSON complex objects and stored into a NoSQL database for presentation and navigation through a new generation web portal. This 2-layer application architecture has proven to combine both generally distinct paradigms: expressivity and performance.

The biomedical data we integrated can be used in other applications, either using directly SPARQL (SPARQL end point not publicly available yet), or through web services. For instance, through the search engine, the "cystin" search query could be executed following a command line or web request (GET) *http://enlord.bndmr.fr/disorders/search/cystin.json[?page=N]*. The result will be a JSON object:

{"term":"cystin","page":1,"total":1,"lastPage":1,"results":[Full collection of diseases found]} that can be directly processed within a third party application. The access to the data through APIs and the application browser are freely available in French and English.

The application was evaluated by clinicians and information managers and used in real setting. The quantity of information on screen was judged good as well as the navigation functions. Clinical geneticists requested more OMIM data and clinicians less Orphanet concepts and more precision upon signs. Some indicated that not all information is useful to all professionals. Medical experts also argued that Orphanet classifications cannot be seen as a diagnosis assistance tool (guiding the clinician from presentations of the disease to the precise disease). Others considered that Orphanet as a master terminology might not be sufficient, as rare manifestations do not have necessary a known diagnosis and diagnosis certainty is a key requirement in the rare disease field. Geneticists also criticized data update latency of OMIM and Orphanet as new genes or mutations are found regularly. Last and most important, not all patients can be given a diagnosis at a certain point of time. The disease might naturally evolve (in case of a genetic disease) over time or the patient might or might not express it. Hence the application should update the thesaurus rapidly but, even with fast updates, the application solely relies on pre-curated terminologies by external parties. The presence of a correct entity to represent at the right granularity (clinical sub-type, disease, group of diseases, disease category) is, for the moment, entirely relying on our data providers. Granularity is also seen by health professionals as a barrier to a homogeneous national data collection for later statistical exploitation of data. A better updating process of these databases by users should be set in order to help Orphanet/HPO/OMIM in their curation tasks which are time consuming. Finally, for the coding task, instructions for coding RD should be elaborated in collaboration with RD expert networks. We foresee in the next months a formal evaluation of

the application whilst the new generation application for data collection for RD centres is rolled out in France (BaMaRa). During the rollout, data accuracy and quality, as well as navigation within the LORD application will be evaluated. Also, coding cases will be set to assess the views proposed by Orphanet as well as the navigational behaviours of the application for that particular purpose.

The application is not yet used in routine in France in all hospital information systems as the coding of rare diseases has not yet started in hospital information systems. As at today, the application is used by 200 users monthly. Users are located in France, Spain, Belgium, Switzerland, Germany and the United States. It is also used by some French university hospitals in clinical routine and by Orphanet France and Switzerland for navigating through their data and for curation.

This first version of the application could be enhanced in many ways. First, we could integrate other existing initiatives like DiseaseCard, which offers more links to external sources than Orphanet. Second, we could help the coding specialist or the expert physician to code the diagnosis from the patient discharge letters. Third, we could develop a search by sign engine that could provide a list of candidate diseases. Fourth, we could propose on the same infrastructure a signs navigation engine for HPO. Fifth, we could enrich the coding possibilities to other nomenclatures such as HPO, OMIM, CIM10, SNOMED since not all rare diseases are represented in Orphanet only. The application is also seen as a potential coding application for the RD-ACTION European joint action (2015-2018) that could also serve as a European terminology service integrated hub for coding rare diseases across European countries.

Conclusion

We introduced a new generation application to lighten the burden of rare diseases codification of cases based on controlled vocabularies. LORD results are significant as (1) the proposed architecture is set on the latest standards (Semantic Web, RDF, JSON, NoSQL, REST) and is flexible enough to curate complex biomedical objects and navigate through them easily at large scale, (2) it brings to users many information about a rare disease into the same UI (genotypic, phenotypic, epidemiological) (3) it could help in some rare cases, in defining the right disease by disease similarity analysis for differential diagnosis, (4) it can help in being more precise in the classification of a disease, (5) it facilitates the specialist navigation through complex graphs of concepts by applying filters on the data (rare eye diseases, rare kidney diseases, etc.).

Acknowledgements

The French Ministry of Health funded this work under the second national plan for rare diseases. We thank our data sources (Orphanet, OMIM, HPO) for their great work in curating and interlinking constantly their respective databases. We warmly thank Céline Angin for her editing and design work.

References

1. OMIM stats. Available at: http://www.ncbi.nlm.nih.gov/Omim/mimstats.html. Accessed March 11, 2015.
2. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 2014;42(1):D966–74. doi:10.1093/nar/gkt1026.
3. Landais P, Messiaen C, Rath A, et al. CEMARA an information system for rare diseases. *Stud Health Technol Inform*. 2010;160(Pt 1):481–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20841733. Accessed February 25, 2013.
4. R. Choquet and P. Landais, « The French national registry for rare diseases: an integrated model from care to epidemiology and research, » *Orphanet journal of rare diseases*, vol. 9, iss. Suppl 1, p. O7, 2014.
5. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. *J Am Med Inform Assoc*. 2014:1–7. doi:10.1136/amiajnl-2014-002794.
6. Orphanet knowledge base for rare diseases. Available at: http://www.orpha.net. Accessed January 1, 2014.
7. *Recommendation on Ways to Improve Codification for Rare Diseases in Health Information Systems*.; 2014. Available at: http://ec.europa.eu/health/rare_diseases/docs/recommendation_coding_cegrd_en.pdf.
8. Januel J-M, Couris C-M, Luthi J-C, et al. [ICD-10 adaptation of 15 Agency for Healthcare Research and Quality patient safety indicators]. *Rev Epidemiol Sante Publique*. 2011;59(5):341–50. doi:10.1016/j.respe.2011.04.004.

9. Dubberke ER, Butler AM, Nyazee HA, et al. The impact of ICD-9-CM code rank order on the estimated prevalence of Clostridium difficile infections. *Clin Infect Dis*. 2011;53(1):20–5. doi:10.1093/cid/cir246.
10. Bruland P, Breil B, Fritz F, Dugas M. Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM. *Stud Health Technol Inform*. 2012;180:564–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22874254. Accessed January 27, 2014.
11. Picardi EJ, Peoples JB. Mesenteric venous thrombosis: ten year record review and evaluation of difficulties with the ICD coding system. *S D J Med*. 1991;44(2):33–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2008604. Accessed January 27, 2014.
12. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514–7.
13. Robinson PN, Mundlos S. The human phenotype ontology. *Clin Genet*. 2010;77(6):525–34. doi:10.1111/j.1399-0004.2010.01436.x.
14. Lopes P, Oliveira JL. An innovative portal for rare genetic diseases research: the semantic Diseasecard. *J Biomed Inform*. 2013;46(6):1108–1115.
15. Grosjean J, Merabti T, Dahamna B, et al. Health multi-terminology portal: a semantic added-value for patient safety. *Stud Health Technol Inform*. 2011;166:129–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21685618. Accessed July 26, 2013.
16. Gardner S. Ontologies and semantic data integration. *Drug Discov Today*. 2005;10(14):1001–1007.
17. Kibbe WA, Arze C, Felix V, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res*. 2014:gku1011.
18. Kilintzis V, Beredimas N, Chouvarda I. Evaluation of the performance of open-source RDBMS and triplestores for storing medical data over a web service. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.Vol 2014. IEEE; 2014:4499–4502. doi:10.1109/EMBC.2014.6944623.
19. Salvadores M, Alexander PR, Musen MA, Noy NF. BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. *Semant Web*. 2013;4(3):277–284. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4159173&tool=pmcentrez&rendertype=abstract. Accessed March 11, 2015.
20. Aimé X, Charlet J, Furst F, Kuntz P, Trichet F, Dhombres F. Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM. *Stud Health Technol Inform*. 2012;180:88–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22874158. Accessed February 25, 2013.