Journal of Information Systems Engineering & Management, 1:2 (2016), 83-90

ISSN: 2468-2071

A knowledge federation architecture for rare disease patient registries and biobanks

Pedro Sernadela[1*], Pedro Lopes[1] & José Luís Oliveira[1]
[1]*University of Aveiro, PORTUGAL*

ABSTRACT

Patient registries are a source of standardized electronic patient information. These records are vital to identify and coordinate a proper cohort, especially for the rare disease domain. Likeness, biobanks are currently an essential instrument for biomedical research, since they provide the very first piece of the biomedical research cycle, i.e. the biological samples. However, connection between rare diseases, patient registries and biobanks has been very limited, due to the lack of common data models and procedures. As they were built with security and privacy in mind, available tools lack comprehensive data access mechanisms, thus making data sharing a complex process. To tackle these challenges, we introduce a semantic web-based architecture to connect distributed and heterogeneous registries and samples. The outcome is a unique knowledge layer, connecting miscellaneous datasets and enabling state-of-the-art semantic data sharing mechanisms.

Keywords

rare diseases,
data integration,
interoperability,
semantic web,
patient registries,
biobanks

Received: 24 Jan 2016
Revised: 13 Feb 2016
Accepted: 1 Mar 2016

DOI: 10.20897/lectito.201615

INTRODUCTION

Current research asserts that a rare disease is a particular condition affecting at most 1 in 2000 patients (Nabarette, Oziel, Urbero, Maxime, & Aymé, 2006). The European Organization for Rare Diseases (EURORDIS) estimates that there are approximately 6000 to 8000 rare diseases, affecting about 6% to 8% of the population (Aymé & Schmidtke, 2007). From these, about 80% have a genetic origin. Still, personal health implications behind rare diseases are seldom considered in medical care. Due to these diseases' low incidence rate and their complex treatment process, their research is still deemed an underrated field in the life sciences. At the patient level, it is difficult to find clinical and psychological support (Seoane-Vazquez, Rodriguez-Monguio, Szeinbach, & Visaria, 2008), due to the reduced incidence of each individual disease. The existence of a small number of cases for each disease creates additional barriers in the translational research pathway, as it is difficult to identify and coordinate a substantial cohort (Schieppati, Henter, Daina, & Aperia, 2008)(Cooper et al., 2010). Nevertheless, the value behind studying rare diseases cannot be ignored, as the combined amount of patients suffering from similar diseases is considerably high, despite the low statistic impact.

During the last decade, several small disease-specific databases related, for instance, to neurological disorders (Gowthaman, Gowthaman, Rajangam, & Srinivasan, 2007) or muscular problems (Aartsma-Rus, Van Deutekom, Fokkema, Van Ommen, & Den Dunnen, 2006) were developed. Despite providing high quality information and resources, their disease coverage is small and their scope is typically regional or national. To achieve higher statistical evidences, the creation of virtual cohorts of patients with similar features spread worldwide is required. Moreover, it is in these particular conditions that the strongest relations between genotypes and phenotypes are identified.

In addition to long-term patient care improvements, understanding gene-disease associations is a

*Correspondence to: sernadela@ua.pt

fundamental goal for bioinformatics research, especially in rare disease where genotype-phenotype connections are typically limited to one or a few more genes (Aronson, 2006)(Wastfelt, Fadeel, & Henter, 2006). Hence, connecting knowledge that is widespread throughout miscellaneous registries is essential to fully understand the underlying causes of diseases. Usually, these are closed data silos with independent data models and relying on primitive formats. Moreover, there is a clear difficulty in finding the adequate ontologies to map internal data from patient registries to an external shared common language, which further compounds this scenario. This results in a lack of interest in sharing data, locking even further the potential behind collected data. Therefore, not only the proper tools to extract information from these databases are needed, but also a common shared model to where available knowledge can be mapped.

In this work, we introduce semantic web-based strategies to provide a seamless working environment for everyone involved in rare disease research. Our goal is to deploy a semantic web layer on top of existing and miscellaneous datasets. With this add on, we will extract anonymised data, translate them to a common shared exchange model and make them available to the research community.

This architecture addresses three key requirements from the rare disease research community, as it is: 1) model agnostic; 2) distributed and independent; and 3) knowledge-oriented.

First, as we are dealing with systems featuring assorted characteristics, the created strategies must be model agnostic and work regardless of registries' data format and internal structure. Although there are modern registries with relational databases and service endpoints, we also come across registries stored in single Excel spreadsheets. Nevertheless, this should not be an obstruction to integrating registries into the semantic knowledge layer. Next, the architecture must be distributed and independent since data anonymity and privacy are key issues when dealing with rare disease patients.

Hence, we must develop tools able to extract meaningful data, while maintaining the original patients' metadata hidden. Likewise, we must also ensure that the new system works without changing the original structure. The entropy of adding this new component to existing systems must be as minimal as possible.

At last, the new system must take advantage of semantic web technologies to extract the true added value of connected knowledge. The semantic web paradigm brings unique standards to improve how we access, express and share knowledge. From a technological perspective, the system was built on top of COEUS (Pedro Lopes & Oliveira, 2012), an application framework that streamlines data integration with semantic representation. COEUS is an application framework designed to streamline the creation of semantic web-oriented systems. By using these technologies, researchers will be allowed to explore the true meaning of their data, since all integrated systems will be seen as a unique virtual component. Researchers and developers will be able to perform distributed queries, covering miscellaneous databases just as if they would query a single local dataset, as patient registries and samples will share their data with a common model.

In summary, we explored a semantic web approach and a non-intrusive strategy to interconnect, enrich and federate data from multiple rare disease patient registries and biobanks, allowing extending the knowledge behind these distributed repositories.

BACKGROUND

Patient registries

Personal genetic records are increasingly important for the diagnosis and therapeutic treatment of rare diseases. This took medicine to a level where wet-lab research is crucial to unravel disease causes and consequences. Hence, databases with information about human genome, such as the Human Gene Mutation Database (HGMD) (Stenson et al., 2003) or the 1000 Genomes Project (Via, Gignoux, & Burchard, 2010), have currently a growing relevance. Moreover, it is important to reuse these data in novel biomedical software to enable its usage on daily medical workflows. The value of individual data increases when it is aggregated and presented in a unified way, both for humans and computers (Mons et al., 2011).

The *de facto* standard in rare diseases software is Orphanet (Rath et al., 2012), a web platform directed to the general public, health professionals and patients, to inform about orphan drugs and rare diseases. It also displays information on specialized consultations, diagnostics, research projects, clinical trials and support groups. Another platform that aggregates genotype-to-phenotype information regarding rare diseases, pointing to key elements for both the education and the biomedical research field is Diseasecard (Pedro Lopes & Oliveira, 2013). Although these systems do not provide repositories for patient level data, they are useful resources for disseminating and sharing existing knowledge.

Another major challenge to support personalized medicine, besides the important role of these specialized

repositories, is the integration of knowledge that can be extracted from distinct electronic health records (EHR. Data from gene sequences, mutations, proteomics, and drug interactions (the genotype) can now be combined with data from EHRs, medical imaging, and disease-specific information stored in patient registries (the clinical phenotype). Hence, it is crucial to start exploring patient-level data from rare diseases registries, which often include personal data, diagnosis, clinical features, phenotypes, genotypes, treatments, and clinical follow up.

These patient-centric databases offer unique specialized views over their internal datasets. However, while there are huge amounts of data scattered throughout multiple stakeholders, they are extremely difficult to obtain. In the end, this results in not enough data to generate statically meaningful conclusions. As such, without having access to a minimal amount of patient data, we cannot discover or infer new knowledge.

To cope with these challenges, we need a system that offers a unique holistic view promoting the collaboration of multiple entities towards the study of rare diseases and assessment of patients' evolution (Thompson et al., 2014).

Biobanks

Biobanks provide the very first piece of the biomedical research cycle: the biological samples. They store samples and related data that can be used to produce results and generate data and knowledge to be reused by other research studies. In Europe, there are two major relevant biobaking infrastructures: BBMRI (Yuille et al., 2008), primarily focused on population biobanks, and EuroBioBank (Lochmüller & Schneiderat, 2010), focused on neuromuscular diseases. Most biobanks use LIMS (Lab Information Management System) to manage samples and bio-resources. The informatics management systems differ from one biobank to another, not only regarding the software provider, but most importantly, regarding data models, data annotation and data representation. Even though there is an increasing effort towards biobank harmonization, standardization and integration, there is still a long way to make possible the finding of samples according to specific requirements in a distributed network of biobanks. Across Europe, millions of samples with related data are held in different types of collections. Nevertheless, one of the most challenging tasks is to build the "provenance" of the sample from the sample donor to the data generated when used in biomedical research studies or in clinical analyses. Ideally, samples should be formally linked not only to all the processes carried out in the biobank, but also to information about the donor and to the data and knowledge generated in the research process or in the clinic.

Historically, the connection inter- and intra- registries and biobanks has been very limited, due to the lack of common standards for data collection, the use of free text non-standardized descriptions, and the variability in data modelization that convert patient registries and biobanks in data silos. In addition, the most common situation is when the same patient is associated with multiple entries in these different registry systems, making data-linkage a more complex task. Hence, there are other challenges to overcome for data sharing and data management such as the high heterogeneity and complexity of the data types, the variability among patient registries and their distributed nature, patient data fragmentation, and the requirement to protect data.

Semantic Web

The Semantic Web arises as a ground breaking paradigm to foster the intelligent integration of structured information. Sustained by state-of-the-art standards such as RDF, OWL, SPARQL and LinkedData, the Semantic Web promote better strategies to express, infer and make knowledge interoperable.

Latest advances in the area cover the research and development of new algorithms to further improve how we collect data, transform data into meaningful knowledge assertions, and publish connected knowledge. State-of-the-art solutions, including the EBI RDF Platform (Jupp et al., 2014), COEUS (Pedro Lopes & Oliveira, 2012) or SADI (Wilkinson, Vandervalk, & McCarthy, 2009), pave the way towards interoperable scientific knowledge. From a large-scale perspective, we can now see the Semantic Web as a single knowledge network. Available technologies foster data integration and publishing, enabling an effortless connection between heterogeneous distributed knowledge.

The true value behind Semantic Web technologies lies in on how easy it is to access and exchange knowledge between independent systems. The Linked Data guidelines, from the W3C working group, promote accessing data via unique URIs that, besides identifying knowledge, must resolve to real data. SPARQL, the Semantic Web query language, complements Linked Data.

Knowledge bases with an open SPARQL endpoint enable direct queries to their content. This empowers researchers and developers alike with an open knowledge highway. In this area, COEUS can

play a fundamental role by delivering a "Semantic Web in a box" approach, enabling the rapid development of new knowledge management systems with semantic web technologies (Pedro Lopes & Oliveira, 2011). COEUS allows gathering data from heterogeneous repositories and publish them via SPARQL endpoint and Linked Data interfaces.

METHODS

Semantic data integration is a complex data engineering issue (Gardner, 2005)(Pasquier, 2008), and the personalized medicine field further increases this complexity. Leveraging on previous results (Pedro Lopes, Sernadela, & Oliveira, 2015), we use COEUS as the baseline framework of our architecture. Exploring its flexible integration engine enables simplifying the overall system architecture through the creation of a comprehensive dependency-based resource integration network.

COEUS framework is focused on helping researchers in the construction and publishing process of new semantically enhanced systems. It offers a good starting point to integrate disparate data due to the advanced ETL (Extract-Transform-Load) processes in its engine. These algorithms facilitate the "triplification" process, in which all data are converted to a simple *subject-predicate-object* model. Moreover, it makes the integrated information available through a hierarchical model establishing relationships between data in an "Entity-Concept-Item" structure (e.g. Protein-Uniprot-P51587). To create each knowledge base according to this organized model, we must follow a comprehensive workflow.

Figure 1 describes the key steps in this semantic integration and translation pipeline: 1) ontology mapping; 2) COEUS setup; 3) semantic translation and 4) data publishing.

The first step consists in defining the best ontologies to map common patient's data. HPO (Robinson & Mundlos, 2010), UMLS (Miličić Brandt, Rath, Devereau, & Aymé, 2011), ICD (Hougland et al., 2008) or ORDO (Rath et al., 2012) are the most widely ontologies used in the rare diseases field. One of the great advantages of using semantic web technologies is that any external ontology can be used to complement or extend COEUS internal model. As long as clinicians understand the new predicates, any number of properties can be included, semantically mapping concepts or entities to existing ontologies, or adding further properties to describe entities or concepts. Moreover, we may combine multiple ontologies, i.e., the same data element can be mapped to terms from more than one ontology, optimising its expressiveness and enriching the way it can be used in future research environments.

The second step of the pipeline consists on the configuration and deployment of a new COEUS instance. The setup involves defining how data will be extracted and mapped into the selected ontology terms. Using COEUS connectors, we have to specify where the data comes from (Excel, CSV or XML files; SQL databases; or SPARQL/LinkedData endpoints), and how we will map them to the ontologies. For instance, for a patient registry available as a CSV file, we need to specify the file location and, for each mapped ontology term, the column containing the actual data elements.

COEUS' configuration enables the semantic translation process. At this stage, new individuals are created for the miscellaneous knowledge base elements and their data and object properties are created in real-time from the integrated data. Along with data format and location diversity, the heterogeneity of each patient registry data model increases the complexity associated with COEUS data integration process. To overcome the fact that data are in all sorts of formats and models, COEUS adds an intermediate abstraction layer between the external resources and the internal knowledge base. The goal is to convert data into a general model-independent format. This process elevates data in primitive formats to a new semantic abstraction level. This step is complete when all data are imported into a new COEUS triplestore, making it available for external use through the various data publishing endpoints.

RESULTS

Migrating systems to a Semantic Web environment is no different from the transition to previous paradigms. New technologies, algorithms and development strategies are introduced, making this transition a cumbersome task. The COEUS framework was built to overcome these challenges. COEUS' flexible integration engine improves traditional data warehousing Extract-Transform-Load tasks, enabling the acquisition of data from heterogeneous resources (in CSV, JSON, XML, SQL, SPARQL, RDF and LinkedData) and its translation to a semantic data abstraction. The latter organizes knowledge in a cohesive structure, ready to be explored by a common and shared model.

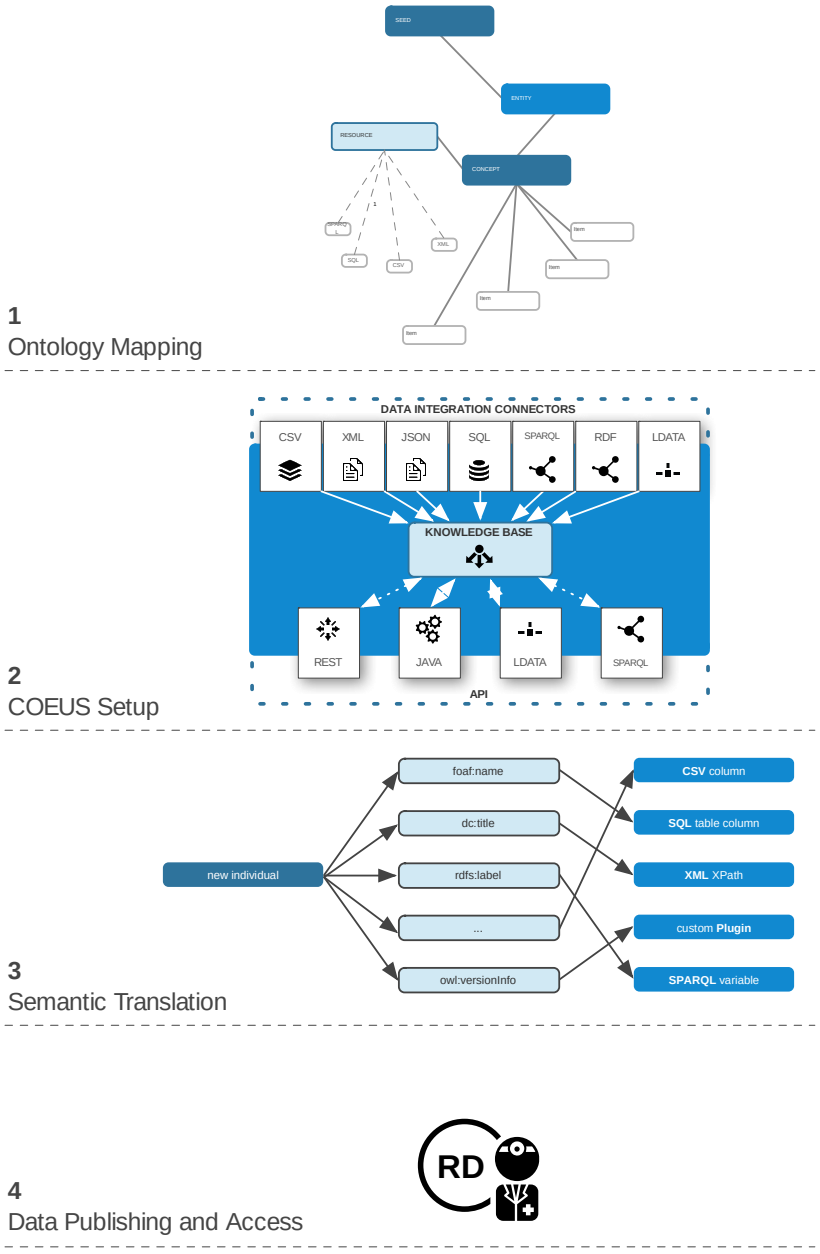

Figure 1. Semantic integration and translation pipeline via COEUS

Figure 2 presents our results, a federated architecture organised in four levels: 1) Patient, 2) Semantic, 3) Federation, and 4) Research.

At the patient level, we gather information from the distributed and heterogeneous patient registries and biobanks, which can be stored in multiple formats and using various technologies (e.g., relational databases, text files, spreadsheets, …). Although Figure 2 only features four components, this solution envisages the inclusion of any number of instances. Patient registries and biobanks can be integrated regardless of their location, as long as an Internet connection is available.

At the second level, we include additional semantics to datasets using COEUS, which acts as the main abstraction, storage and publishing engine. Here, we manage the anonymised data, translating them from their primitive format to common biomedical ontologies.

The third level provides the knowledge federation and data exploration capabilities, i.e., SPARQL queries

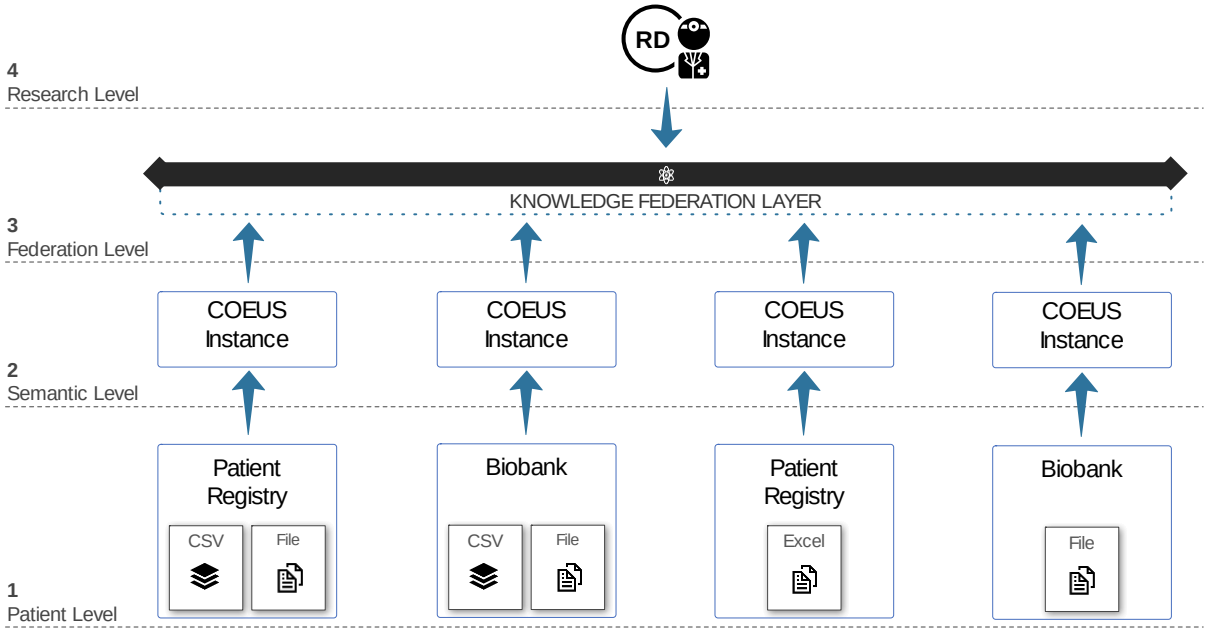

Figure 2. Knowledge federation architecture, integrating distributed patient registries and biobanks

can be forwarded to several distributed endpoints. COEUS acts here as a middleware component between each registry triplestore and the public knowledge federation layer.

Finally, at the upper level, researchers can perform general queries that combine data from several patient's registries and biobanks. In a sense, query federation enables performing SQL-like UNIONs or JOINs across multiple knowledge bases.

SPARQL Endpoint becomes the main preference to access data, since it is a flexible way to interact with Web of Data, by formulating queries like SQL in traditional databases. Knowledge bases with an open SPARQL endpoint enable direct queries to their content. This empowers researchers and developers alike with an open knowledge highway. With these federation systems, the data is discovered by following HTTP URIs of distributed endpoints, each distinct repositories providing a wide and heterogeneous query engine that supports the principles of Linked Data (Bizer, Heath, & Berners-Lee, 2009). This type of federation strategies has been topic of recent research in the Semantic Web research community (Freitas, Curry, Oliveira, & O'Riain, 2012).

DISCUSSION

In our research work, we identified how semantic web technologies can be tailored to the patient registries and biobanks integration scenario. Although our results are successful, they highlight two major issues.

First, identifying the proper common ontology to be used across patient registries and biobanks is a cumbersome challenge. While COEUS empowers this process at the technical level, there still has to be an agreement between stakeholders on what ontologies will be used and how will their data be properly mapped to them. This introduces a new challenge, as distinct ontologies need to be adequately mapped (Kumar & Harding, 2013).

Second, convincing data owners of the true value in sharing their data is a difficult task. In addition to the privacy and security issues, data owners fail to realize the incentives underlying the sharing of their data. To overcome this in the future, financing projects should include clear guidelines to mandate the anonymous sharing of data for research purposes. Including these political policies would shed a new light on the benefits of sharing rare diseases data to a broader community, truly unlocking its potential.

CONCLUSIONS

This work introduces a unique semantic web-based architecture that moves us towards knowledge federation in rare diseases patient records ecosystems. This delivers a lightweight holistic perspective over the

wealth of knowledge stemming from connected patient registries and biobanks supported by the growing number of research projects.

Our results are significant in at least three major respects: 1) The use of a model agnostic system, which enables the mapping of patient data from any format to a common shared ontology. 2) The creation of an independent system that can be plugged into any existing infrastructure without changing it. This enables the extraction of relevant data elements, while maintaining patients' data privacy and security. 3) The adoption of Semantic Web technologies to promote a better translation, interpretation, and federation of knowledge.

Finally, this architecture enables researchers to easily access a broad set of patients' records by using SPARQL federated queries. As a result, distributed repositories can be accessed towards semantic interoperability on rare disease research.

COMPETING INTERESTS

The authors declare no competing interests on the presented work.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 305444 – the RD-Connect project. Pedro Sernadela is funded by Fundação para a Ciência e Tecnologia (FCT) under the grant agreement SFRH/BD/52484/2014.

REFERENCES

Aartsma-Rus, A., Van Deutekom, J.C.T., Fokkema, I.F., Van Ommen, G.B., and Den Dunnen, J.T., 2006. Entries in the Leiden Duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle & Nerve*, *34*(2), pp.135–144. http://doi.org/http://dx.doi.org/10.1002/mus.20586

Aronson, J.K., 2006. Rare diseases and orphan drugs. *British Journal of Clinical Pharmacology*, *61*(3), pp. 243–245. http://doi.org/http://dx.doi.org/10.1111/j.1365-2125.2006.02617.x

Aymé, S., and Schmidtke, J., 2007. Networking for rare diseases: A necessity for Europe. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *50*(12), pp. 1477–1483. http://doi.org/10.1007/s00103-007-0381-9

Bizer, C., Heath, T., and Berners-Lee, T., 2009. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*. Retrieved from http://www.igi-global.com/article/linked-data-story-far/37496

Cooper, D.N., Chen, J.M., Ball, E.V., Howells, K., Mort, M., Phillips, A.D., Chuzhanova, N., Krawczak, M., Kehrer-Sawatzki, H. and Stenson, P.D., 2010. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Human mutation*, *31*(6), pp.631-655. http://doi.org/http://dx.doi.org/10.1002/humu.21260

Freitas, A., Curry, E., Oliveira, J.G., and O'Riain, S.,2012. Querying heterogeneous datasets on the linked data web: Challenges, approaches, and trends. *IEEE Internet Computing*, *16*(1), pp. 24–33. http://doi.org/10.1109/MIC.2011.141

Gardner, S.P.,2005. Ontologies and semantic data integration. *Drug Discovery Today*, *10*(14), pp. 1001–1007.

Gowthaman, R., Gowthaman, N., Rajangam, M.K. and Srinivasan, K., 2007. Database of neurodegenerative disorders. *Bioinformation*, *2*(4), p.153.

Henriksen, K., Battles, J.B., Keyes, M.A., Grady, M.L., Hougland, P., Nebeker, J., Pickard, S., Van Tuinen, M., Masheter, C., Elder, S. and Williams, S., 2008. Using ICD-9-CM codes in hospital claims data to detect adverse events in patient safety surveillance. *Advances in Patient Safety: New Directions and Alternative Approaches (Vol 1: Assessment)*.

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N. and Wimalaratne, S.M., 2014. The EBI RDF platform: Linked open data for the life sciences. *Bioinformatics*, *30*(9), pp.1338-1339.

Kumar, S.K., and Harding, J.A., 2013. Ontology mapping using description logic and bridging axioms. *Computers in Industry*, *64*(1), pp. 19–28.

Lochmüller, H. and Schneiderat, P., 2010. Biobanking in rare disorders. In *Rare Diseases Epidemiology* (pp. 105-113). Springer Netherlands.

Lopes, P., & Oliveira, J.L., 2011. COEUS: A semantic web application framework. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences* (pp. 66–73). ACM. http://doi.org/http://dx.doi.org/10.1145/2166896.2166915

Lopes, P., and Oliveira, J.L., 2012. COEUS: "Semantic web in a box" for biomedical applications. *Journal of Biomedical Semantics*, *3*(1), p. 11. http://doi.org/10.1186/2041-1480-3-11

Lopes, P. and Oliveira, J.L., 2013. An innovative portal for rare genetic diseases research: The semantic Diseasecard. *Journal of biomedical informatics*, *46*(6), pp.1108-1115. http://doi.org/10.1016/j.jbi.2013.08.006

Lopes, P., Sernadela, P., and Oliveira, J.L., 2015. Towards a knowledge federation of linked patient registries. In *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–5). IEEE. http://doi.org/10.1109/CISTI.2015.7170546

Miličić Brandt, M., Rath, A., Devereau, A., and Aymé, S., 2011. Mapping orphanet terminology to UMLS. In M. Peleg, N. Lavrač, & C. Combi (Eds.), *Artificial intelligence in medicine* (Vol. 6747, pp. 194–203). Springer Berlin Heidelberg. http://doi.org/http://dx.doi.org/10.1007/978-3-642-22218-4_24

Mons, B., van Haagen, H., Chichester, C., den Dunnen, J.T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J. and Kiesel, B., 2011. The value of data. *Nature genetics*, *43*(4), pp.281-283. http://doi.org/10.1038/ng0411-281

Nabarette, H., Oziel, D., Urbero, B., Maxime, N., and Aymé, S., 2006. Use of a directory of specialized services and guidance in the healthcare system: The example of the Orphanet database for rare diseases. *Revue d'épidémiologie et de santé publique*, *54*(1), pp. 41–53. Retrieved from http://europepmc.org/abstract/MED/16609636

Pasquier, C., 2008. Biological data integration using Semantic Web technologies. *Biochimie*, *90*(4), pp. 584–594.

Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., and Ayme, S., 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, *33*(5), pp. 803–808. http://doi.org/10.1002/humu.22078

Robinson, P.N., and Mundlos, S., 2010. The human phenotype ontology. *Clinical Genetics*, *77*(6), pp. 525–534.

Schieppati, A., Henter, J.I., Daina, E., and Aperia, A., 2008. Why rare diseases are an important medical and social issue. *The Lancet*, *371*(9629), pp. 2039–2041. http://doi.org/http://dx.doi.org/10.1016/S0140-6736(08)60872-7

Seoane-Vazquez, E., Rodriguez-Monguio, R., Szeinbach, S.L. and Visaria, J., 2008. Incentives for orphan drug research and development in the United States. *Orphanet journal of rare diseases*, *3*(1), p.1.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M. and Cooper, D.N., 2003. Human gene mutation database (HGMD®): 2003 update. *Human mutation*, *21*(6), pp.577-581. http://doi.org/10.1002/humu.10212

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I.G., Hansson, M.G., Peter-Bram, A., Patrinos, G.P., Dawkins, H. and Ensini, M., 2014. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *Journal of general internal medicine*, *29*(3), pp.780-787. http://doi.org/10.1007/s11606-014-2908-8

Via, M., Gignoux, C. and Burchard, E.G., 2010. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med*, *2*(3).

Wastfelt, M., Fadeel, B., and Henter, J.I., 2006. A journey of hope: Lessons learned from studies on rare diseases and orphan drugs. *Journal of Internal Medicine*, *260*(1), pp. 1–10. http://doi.org/http://dx.doi.org/10.1111/j.1365-2796.2006.01666.x

Wilkinson, M.D., Vandervalk, B. and McCarthy, L., 2009, December. SADI Semantic Web Services-,cause you can't always GET what you want!. In*Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific*(pp. 13-18). IEEE.

Yuille, M., van Ommen, G.J., Bréchot, C., Cambon-Thomsen, A., Dagher, G., Landegren, U., Litton, J.E., Pasterk, M., Peltonen, L., Taussig, M. and Wichmann, H.E., 2008. Biobanking for europe. *Briefings in bioinformatics*,*9*(1), pp.14-24. http://doi.org/10.1093/bib/bbm050

