Journal of Biomedical Informatics 46 (2013) 1108–1115

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

An innovative portal for rare genetic diseases research: The semantic Diseasecard

Pedro Lopes *, José Luís Oliveira

DETI/IEETA, Universidade de Aveiro, Portugal

ARTICLE INFO

Article history:
Received 5 April 2013
Accepted 13 August 2013
Available online 21 August 2013

Keywords:
Rare diseases
Biomedical semantics
Data integration
Interoperability
Semantic web

ABSTRACT

Advances in "omics" hardware and software technologies are bringing rare diseases research back from the sidelines. Whereas in the past these disorders were seldom considered relevant, in the era of whole genome sequencing the direct connections between rare phenotypes and a reduced set of genes are of vital relevance.

This increased interest in rare genetic diseases research is pushing forward investment and effort towards the creation of software in the field, and leveraging the wealth of available life sciences data. Alas, most of these tools target one or more rare diseases, are focused solely on a single type of user, or are limited to the most relevant scientific breakthroughs for a specific niche. Furthermore, despite some high quality efforts, the ever-growing number of resources, databases, services and applications is still a burden to this area. Hence, there is a clear interest in new strategies to deliver a holistic perspective over the entire rare genetic diseases research domain.

This is Diseasecard's reasoning, to build a true lightweight knowledge base covering rare genetic diseases. Developed with the latest semantic web technologies, this portal delivers unified access to a comprehensive network for researchers, clinicians, patients and bioinformatics developers. With in-context access covering over 20 distinct heterogeneous resources, Diseasecard's workspace provides access to the most relevant scientific knowledge regarding a given disorder, whether through direct common identifiers or through full-text search over all connected resources. In addition to its user-oriented features, Diseasecard's semantic knowledge base is also available for direct querying, enabling everyone to include rare genetic diseases knowledge in new or existing information systems. Diseasecard is publicly available at http://bioinformatics.ua.pt/diseasecard/.

1. Introduction

Rare genetic diseases research is at the forefront of the most modern personalized medicine endeavors. The rare term broadly defines disorders that affect at most 1 in 2000 patients [1] and the European Organization for Rare Diseases (EURORDIS) estimates that there are approximately 6000–8000 rare diseases, affecting about 6–8% of the population [2]. Within these, about 80% are caused by genetic changes, further strengthening the relations between genotypes and phenotypes associated with these particular conditions [3,4]. Some of these chronic diseases hinder the patients' quality of life and cause serious damage or social disability [5]. Moreover, the low patient number severely obstructs the creation of adequate research cohorts, making it very difficult to coordinate studies capable of generating results in a scientifically-relevant scale [6,7].

In addition to long-term patient care improvements, understanding gene-disease associations is a fundamental goal for bioinformatics research, especially at the rare disease level, where genotype-phenotype connections are typically limited to one or a few more genes. This moves rare diseases research from a relatively minor concern to a major player in a new era of "omics" research [8,9]. Genomics, metabolomics, proteomics or pharmacogenomics, among others, benefit from the focused approach enabled by rare genetic diseases research. A direct consequence of this growing importance is the renewed interest from pharmaceutical companies in this area, which are supporting multiple worldwide initiatives towards improved rare diseases research. On a broader scope, the International Rare Diseases Consortium (IRDiRC) (http://www.irdirc.org/) is leveraging several projects on the field. RD-Connect (http://rd-connect.eu/), RareConnect (https://www.rareconnect.org/), EuRenOmics (http://eurenomics.eu/) and NeurOmics (http://rd-neuromics.eu/) are some of the highlights from IRDiRC sponsorships. On the European level, GEN2PHEN (http://www.gen2phen.org/), EU-ADR [10] (http://eu-adr-project.org/), or EMIF (http://www.imi.europa.eu/content/

* Corresponding author. Address: IEETA, Campus Universitario de Santiago, 3810-193 Aveiro, Portugal.
E-mail address: pedrolopes@ua.pt (P. Lopes).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.006

emif/) projects are actively investing in setting up state of the art research activities for rare diseases stakeholders. A key feature of these projects is their multidisciplinary approach, leveraging on the natural connections between clinicians, researchers, bioinformaticians and patients, who, for once, are active elements of the proposed strategies.

With these miscellaneous initiatives, players and requirements, there is an overwhelming challenge to tackle the wealth of data being made available by next-generation sequencing hardware, omics databases, patient registries, and pharmacovigilance or electronic health records. Multiple new rare diseases research tools are emerging, focusing only a set of particular conditions. Niche databases for the domains of neurological disorders [11] or muscular problems [12], for example, tackle small specific sub-groups. While they provide high quality information and resources, their disease coverage is small and inadequate.

Likewise, platforms such as the Online Mendelian Inheritance in Man database (OMIM) [13,14], the National Organization for Rare Disorders (NORD) website (http://www.rarediseases.org/) or Orphanet, among others, collect and filter available information, with a particular set of users in mind. Where OMIM is more focused on aggregating content for bioinformatics researchers, Orphanet has decade-long pedigree and content far beyond simple research data. It is geared towards clinicians, medical researchers and patients, boasting a large collection of curated clinical information such as patient registries, biobanks and specialized clinician contacts, among others, in multiple languages.

Despite these high quality efforts, entropy is a recurring problem in this field with the ever-growing number of resources, databases, services and applications. Therefore, a new approach is desired, one where everyone is able to quickly access the available knowledge regarding a given set of rare disorders.

Diseasecard addresses these needs by delivering a lightweight holistic perspective over the rare genetic diseases research field. Stemming from a legacy portal [15,16], a crawler-based system that kick-started a new strategy for in-context research, the new semantic Diseasecard version focuses on three fundamental elements:

- The rapid and lightweight access to a comprehensive semantic network of scientifically relevant resources for a given disease, covering multiple domains from proteomics to clinical studies up to medical ontologies.
- The innovative in-context browsing allowing for the eased navigation amongst the multitude of connected resources without leaving the initial research focus.
- The open interoperability layer, making the semantic knowledge base available for everyone to query and access, and enabling the integration of rich rare diseases data in new or existing information systems.

In addition, Diseasecard comprises a rich semantic layer, providing future-proof technologies for inference and reasoning over the created knowledge base. Diseasecard is publicly available online at http://bioinformatics.ua.pt/diseasecard/.

2. Methods

Semantic data integration is, in itself, a complex data engineering issue [17], and the life sciences field further increases this complexity [18]. To support Diseasecard's ambitious integration and interoperability features we rely on the COEUS framework [19]. Exploring COEUS flexible integration engine enabled us to simplify the overall platform architecture through the creation of a comprehensive dependency-based resource integration network.

Diseasecard's integration pipeline, including COEUS' use and the internal resource organization, are described in detail next.

2.1. Architecture

To overcome the challenges behind the amount of scattered data for rare diseases, Diseasecard's underlying objective is to collect, connect and deliver access to a network of the most relevant rare diseases scientific resources. To attain this, Diseasecard's knowledge base is constructed from an integration network starting with OMIM's morbid map and spanning through multiple resources, including proteomics data from UniProt [20], InterPro [21], Prosite [22] and Protein Data Bank (PDB) [23] up to ontology data from Medical Subject Headings (MeSH) [24] and International Classification of Diseases (ICD version 10), among many others. These data are obtained from multiple mapping studies [25–27] and genomic name servers, such as GeNS [28] and Bio2RDF [29]. This broad scope results in an extremely rich dataset, where OMIM's rare disorder list is expanded to more than 2 million triples.

To improve its semantic data integration and interoperability features, Diseasecard is built with the COEUS semantic web application framework [30], which heavily influences Diseasecard's architectural design. COEUS delivers a "Semantic Web in a box" approach, enabling the rapid development of new knowledge management systems adopting semantic web technologies [31,32].

By default, the COEUS framework already includes the necessary components to build and launch a new semantic information system from scratch. The platform comprises the tools to acquire and translate knowledge from miscellaneous data sources, and to deliver access to the constructed knowledge base through various interoperable formats. One of COEUS' key caveats is the lack of advanced update methods. Despite this, the trade-off between its semantic integration and interoperability capabilities, and the lack of update features drawback is a positive one, especially considering COEUS' build engine performance.

To complete Diseasecard's architecture we created a dedicated client-side application, to support the agile web workspace; added an indexing engine to improve the efficiency behind the full-text search infrastructure; and added an object-oriented database, to cache data for each rare genetic disease network, to improve workspace access performance. The entire architecture is described in detail in Fig. 1.

2.2. From OMIM's maps to 2 million triples

Diseasecard adopts a targeted warehousing data integration strategy [33,34]. Accordingly, the data import and translation process gathers all data from external resources in a single centralized knowledge base, in opposition to real-time data gathering strategies [35]. Curating a niche warehouse focused on rare genetic diseases knowledge enables Diseasecard's future endeavors on advanced inference and reasoning algorithms. Since we were looking at managing semantic information from the start, Diseasecard's integration process not only collects data per se, but it translates external data into a semantic knowledge base. Creating this new semantic layer leverages a major challenge on how to translate large heterogeneous datasets into a new semantic environment. This problem can be divided in two areas, focusing on the technological challenge and on the logical data modelling challenge.

From the technological perspective, Diseasecard relies on the COEUS framework to perform the data abstractions, triplifying data from the miscellaneous external resources into a unified Diseasecard knowledge environment.

On the integration modelling side, Diseasecard has a custom ontology to integrate OMIM's data (http://bioinformatics.ua.pt/dis-

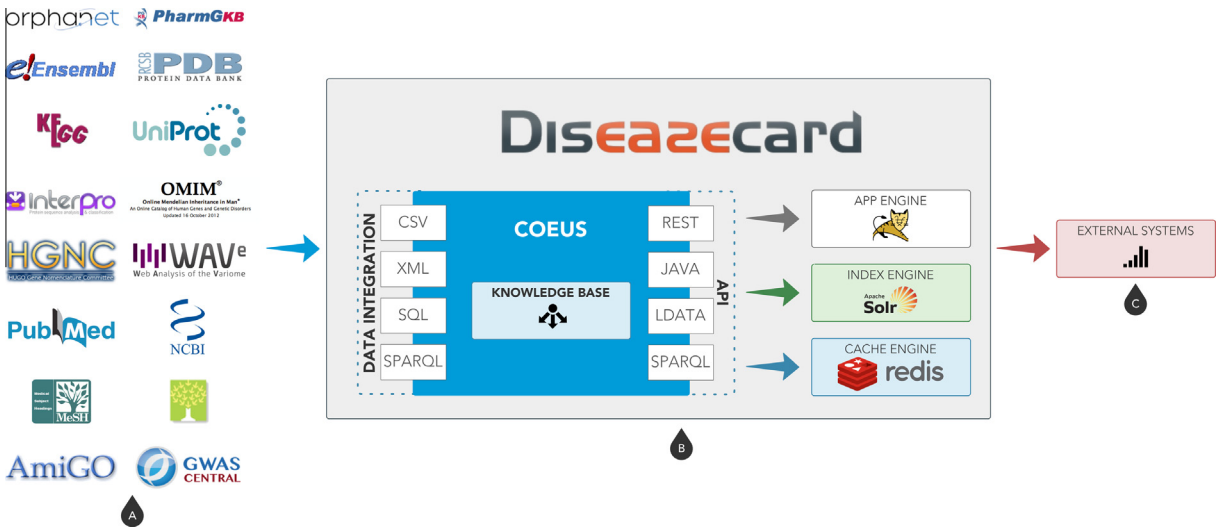

Fig. 1. Diseasecard architecture overview. (A) External resources are identified and configured for integration in Diseasecard's rare genetic diseases semantic network. (B) Diseasecard architecture, highlighting COEUS, with the data integration connectors, the knowledge base, and the interoperability API; the Tomcat server, for web application delivery; the Solr indexing engine, for improved search performance; and the Redis cache engine, for faster disease network access. (C) Any external system can query Diseasecard's knowledge base and use its data without limitations.

easecard/diseasecard.owl). This simple ontology is used to enhance the translation from OMIM's morbid map into a semantic environment. Additionally, following Semantic Web's "reuse instead of rewrite" motto, Diseasecard's data model reuses existing schemas internally. Using COEUS' instance configuration and taking advantage of existing ontologies and models for internal use is enough to organize collected data.

Diseasecard's lightweight integration approach means that, for each individual, such as a UniProt protein or an OMIM entry, we only need to store its identifier. Hence, we can reuse the identifier term from the Dublin Core ontology [36]. As such, each individual has a **dc:identifier** data property, matching a string with the external identifier. Another example is the **rdfs:label** property, obtained from the Resource Description Format (RDF) schema ontology that is used to label each individual [37]. External LinkedData references are also included, establishing direct connections to external individuals. For instance, UniProt published interfaces are linked through the **rdfs:seeAlso** property. Despite this over-simplification, new relationships amongst integrated data are autonomously generated. Whereas in a CSV file we have a set of columns with text, with the move to a semantic environment all data are interconnected, generating a richer dataset. The same is true for SQL databases where foreign key relationships and table/column names are mapped to new properties, resulting in more metadata and more relationships.

Starting with OMIM's morbid map, which has around 6300 entries related to a gene map with about 14,200 entries, Diseasecard's engine expands the integration network, generating new knowledge, and collecting pointers for the resources mentioned in Table 1, as detailed in the following section.

2.3. Semantic integration network

Diseasecard's knowledge base covers miscellaneous resources within the rare genetic diseases research domain. The knowledge base is obtained from a dependency graph, the semantic integration network, where the acquisition of data from external sources is defined. Starting with OMIM's morbid map, Diseasecard loads information about rare disorders and HUGO Gene Nomenclature Committee (HGNC) gene symbols. The integration engine then proceeds to expand the list of integrated individuals into new sources. For instance, using the OMIM accession number, Diseasecard obtains the associated UniProt and Orphanet identifiers. Likewise, from the integrated HGNC symbols, Diseasecard obtains identifiers for GWAS Central, Clinical Trials and Ensembl databases.

Table 1
List of resources integrated in Diseasecard's knowledge base, comprising the entity (for the tree-based navigation interface), the resource name, the origin of the integration mapping and the original resource URL.

Entity	Resource	Origin	Resource URL
Disease	OMIM	Morbid map	http://www.omim.org/
	Orphanet	OrphaData	http://www.orpha.net/
Drug	PharmGKB	GeNS	http://www.pharmgkb.org/
Literature	Pubmed	Morbid map	http://www.ncbi.nlm.nih.gov/pubmed/
Locus	Ensembl	Ensembl	http://www.ensembl.org/
	Entrez	Entrez	http://www.ncbi.nlm.nih.gov/gene/
	GeneCards	UniProt	http://www.genecards.org/
	HGNC	Gene map	http://www.genenames.org/
Ontology	Gene Ontology	UniProt	http://amigo.geneontology.org/
	ICD10	OrphaData	http://www.who.int/classifications/icd/en/
	MeSH	UniProt2MeSH	http://www.nlm.nih.gov/mesh/
Pathways	KEGG	UniProt	http://www.genome.jp/
	Enzyme	UniProt	http://enzyme.expasy.org/
Protein	InterPro	UniProt	http://www.ebi.ac.uk/interpro/
	PDB	UniProt	http://www.pdb.org/
	PROSITE	UniProt	http://prosite.expasy.org/
	STRING	UniProt	http://string-db.org/
	UniProt	UniProt	http://www.uniprot.org/
Study	Clinical Trials	Clinical Trials	http://www.clinicaltrials.gov/
	GWASCentral	GWASCentral	https://www.gwascentral.org
Variome	LSDB	GEN2PHEN	http://gen2phen.org/
	WAVe	HGNC	http://www.bioinformatics.ua.pt/WAVe/

Fig. 2 displays a visual overview over Diseasecard's complete semantic integration network, highlighting the multiple connections amongst resources and how each is extended to generate more data inputs into the knowledge base.

A sample example for semantic integration is the translation of UniProt and PDB identifiers into Diseasecard's knowledge base. With UniProt becoming a major source for data mappings, it is used in Diseasecard to establish connections among diseases, proteins and external entities. This integration branch starts with the aggregation of UniProt entries associated with each particular OMIM code (using COEUS' CSV connector to translate UniProt search results). Next, Diseasecard uses COEUS' XML connector to load XPath query results and generate new triples. For instance, to create new PDB individuals and their respective connections, the "//entry/dbReference[@type = "PDB"]" XPath query is performed on each UniProt entry XML.

When Diseasecard retrieves the semantic network for each disease, the association graph between OMIM, UniProt and PDB entries is traversed.

2.4. Implementation

Using COEUS imposes some restrictions on the technologies used in the Diseasecard platform. As the framework provides a solid Java-based backend solution, we opted to deploy the Diseasecard client-side web environment also within an Apache Tomcat server (http://tomcat.apache.org/).

Diseasecard's indexing engine is built on top of Solr (http://lucene.apache.org/solr/). This Lucene-based search engine enables indexing the resources connected in Diseasecard's knowledge base and searching them with a notable performance.

Initial tests unravelled a slow response time for the disease network generation tasks. The complexity behind the SPARQL queries retrieving all identifiers associated with a given rare genetic disease reduces the web application usability. Hence, Diseasecard uses an object-oriented database, Redis (http://redis.io/), to store a cached version of the knowledge network for each disease entry. With this, the performance increased ten-fold from an average page loading time of 4.5 s to an almost instant 300 ms for the rare diseases workspaces.

For Diseasecard's web interface, a combination of JavaScript algorithms with modern CSS and HMTL5 technologies was used. The main JavaScript library used is jQuery (http://jquery.com/), with several plugins for cookie management and tree displays. The JavaScript InfoVis Toolkit (http://philogb.github.io/jit/) is used to display the central workspace hypertree. The outcome of these technological implementation choices is a responsive and agile web application, further improving the final user experience.

3. Results

Diseasecard's knowledge base contains around 2 million triples, built from OMIM's maps and the expanded rare genetic diseases network. These triples establish about 500 thousand connections to more than 100 thousand unique resources, which are entirely indexed by Diseasecard's engine. From these numbers we can infer that, on average, each unique resource is present in 5 single disease networks.

Additionally, each rare disorder has, on average, around 24 connections to external resources. As expected, disease resources from OMIM represent the biggest slice of individuals with around 18 thousand entries for more than 11 thousand HGNC entries. Another interesting result stems from Ontology mappings, as MeSH and ICD terms are the least represented concepts in Diseasecard's knowledge base.

The entirety of these data are stored in Diseasecard's semantic knowledge base, made available for end-users through an innovative in-context web workspace and to developers through an advanced semantic interoperability layer, which are detailed next.

3.1. Semantic knowledge base

One of the main premises behind the creation of a new Diseasecard version lied in the need to better explore the powerful technologies pushed forward by the Semantic Web paradigm. With these, Diseasecard is able to construct a rich and comprehensive semantic knowledge base for rare genetic diseases. Its knowledge infrastructure extends the capabilities of the majority research platforms by making the collected knowledge interoperable and future-proof.

With this knowledge network the door is open for inferring new relationships amongst connected resources and for reasoning over gathered data in search for previously uncovered connections. In the future, these methods can be used to enrich Diseasecard's knowledge base with new annotations, and to federate knowledge discovery through multiple databases with the publicly available SPARQL endpoints.

It is equally important to note that the lightweight integrative approach adopted in Diseasecard re-uses existing ontologies to describe data. Consequently, data from the knowledge base is easier to integrate by third parties and to connect using LinkedData technologies.

3.2. In-context research

Diseasecard is a unique alternative for exploring biomedical rare diseases information in a centralised web-based workspace. Along with direct access to diseases' workspaces through any of the integrated resources identifiers, full-text searching enables

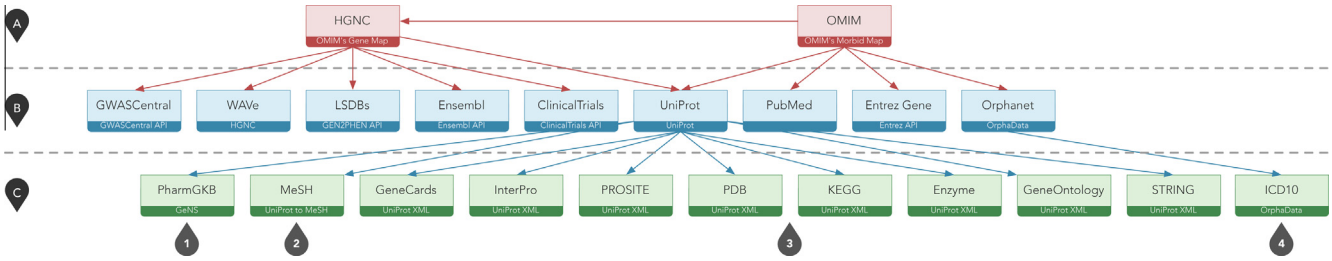

Fig. 2. Diseasecard's integration network overview. (A) The integration process starts with OMIM's morbid map and constructs the OMIM and HGNC individuals, which will be used to expand other resources. (B) On a second level, the Orphanet's OrphaData is used to load Orphanet mappings; UniProt is used for disease-gene-protein associations; ClinicalTrials, GWASCentral, WAVe, LSDBs and Ensembl mappings are obtained from their respective APIs using HGNC and OMIM identifiers. (C) The third and final level loads: (1) PharmGKB entries from GeNS for each UniProt entry; (2) MeSH terms are loaded from the results of previous research; (3) GeneCards, InterPro, Prosite, PDB, KEGG, Enzyme, Gene Ontology and STRING entries from direct UniProt queries; and (4) ICD10 mappings to Orphanet entries from OrphaData.

querying all the web pages for all the resources integrated in Diseasecard's knowledge base.

3.2.1. Search

For a more comprehensive access, Diseasecard has a powerful search feature comprising three components: browsing, identifier search and full-text search.

With Diseasecard's browsing feature, users can browse all entries by their starting letter – http://bioinformatics.ua.pt/diseasecard/browse. This displays the OMIM accession number, the disease name and the number of available connections in the generated network.

The identifier search, selected by default, searches through the extended identifier network. This network includes accession numbers for resources listed in Table 1. In addition to search boxes, using the query string to perform searches is also available. For instance, http://bioinformatics.ua.pt/diseasecard/search/id/HTT will retrieve all entries where the HTT string identifier is present.

At last, full-text search is the most powerful search mechanism. This method, selectable in the home page search button, searches web pages for the connected resources. This results in detailed access to a restricted set of locations, where users filter more specific queries such as author names, publication titles, protein sequences or more complex disease descriptions.

3.2.2. Web workspace

As mentioned, Diseasecard's semantic integration network starts with OMIM's morbid map. Consequently, disorders can be directly accessed using their unique OMIM identifier. For instance, OMIM's "Huntington Disease" entry (OMIM #143100) can be explored in Diseasecard at http://bioinformatics.ua.pt/diseasecard/entry/143100. The displayed web workspace has two key features: the navigation tree and map – Fig. 3(A), and the LiveView browsing – Fig. 3(B).

The navigation tree and map are two complementary alternatives for exploring each network. The left sidebar displays the disease navigation tree to quickly access all links with a familiar metaphor. The central area displays a circular navigation map, pointing to all individual identifiers. Both the navigation tree and map trigger the Live View feature. This opens the external resource application within Diseasecard, allowing users to browse the multiple collected connections without leaving the initial context.

3.3. Semantic interoperability

With Diseasecard's knowledge base built, several interoperability services are enabled by default. Hence, Diseasecard's access API includes two main data access alternatives: a SPARQL endpoint [38] and a LinkedData interface [39]. These two options allow flexible output formats, thus facilitating the data integration from Diseasecard's platform in external applications.

SPARQL is the most advanced query language available and enables distributed reasoning and inference, as well as combining Diseasecard's data with other federated SPARQL endpoints.

The LinkedData interfaces provide quick access to all data for a given resource and the use of these resources, through their URIs, in any external context. The URI for accessing Huntington's disease data in Diseasecard's knowledge base is http://bioinformatics.ua.pt/diseasecard/resource/omim_143100.

In the discussion section we highlight how these methods can be used to enrich an existing system with Diseasecard's rare genetic diseases knowledge.

4. Discussion

The concept behind Diseasecard is a unique approach towards comprehensive access to rare genetic diseases research knowledge. To demonstrate Diseasecard's innovative features, we detail in the following sections a couple real-world scenarios regarding Huntington disease.

Starting with an end-user oriented scenario, we highlight the variety of relevant questions that can be answered with

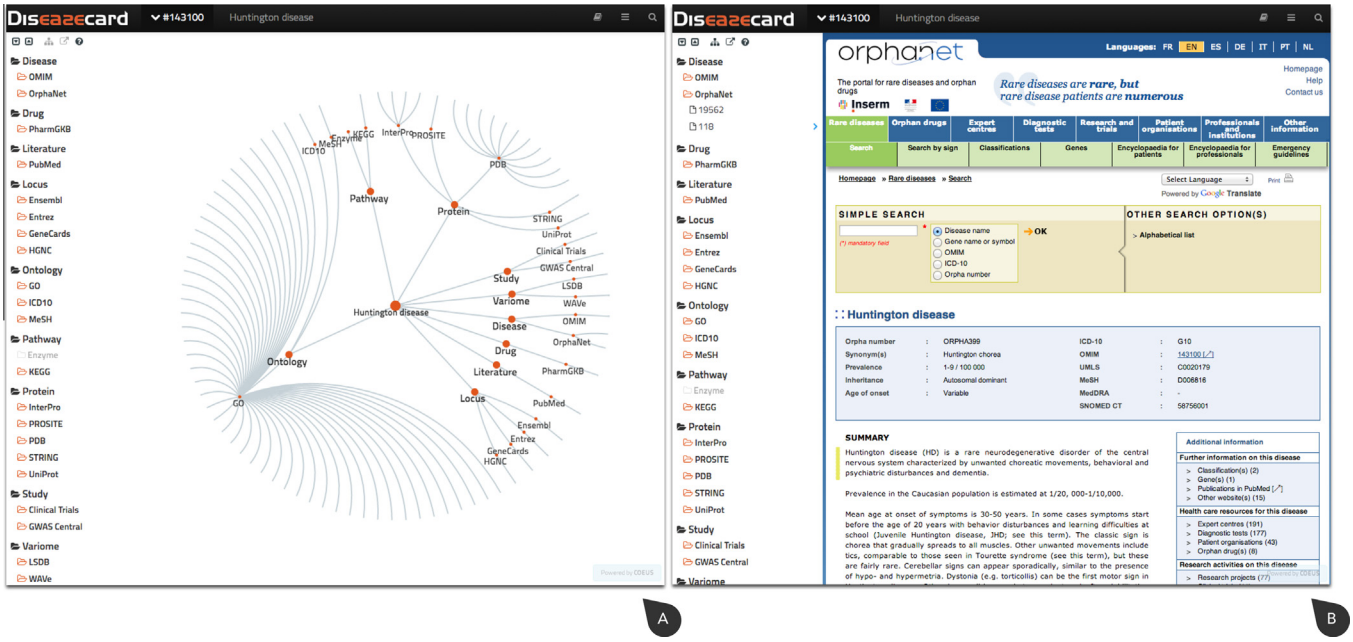

Fig. 3. Diseasecard's workspace for Huntington's disease. (A) Initial display for Diseasecard's workspace for Huntington's disease. Both the left navigation tree and the central hypertree provide direct access to the collected disease network. These tree-based navigation strategies reflect the Entity-Concept-Item structure behind Diseasecard's configuration and result in a familiar interaction metaphor for the application users. (B) Loading external resources in Diseasecard using LiveView opens the associated location in the workspace, empowering in-context research. LiveView highlights Diseasecard's connectedness features, as external resources are not hidden or replicated, they are linked. Furthermore, this strategy overcomes traditional drawbacks for original resource creators, maintaining content accreditation and ownership.

Diseasecard for clinicians, researchers and patients. In comparison with the aforementioned systems, such as Orphanet, NORD or OMIM, none offers such a broad scope of directly accessible knowledge.

On top of end-user features, Diseasecard also provides an interoperability layer. Henceforth, we emphasize how its knowledge can be integrated in external systems through simple methods.

4.1. Exploring Huntington's disease knowledge

Huntington disease is an "autosomal dominant neurodegenerative disorder with midlife onset characterized by psychiatric, cognitive, and motor symptoms" [40]. Once a patient is diagnosed with Huntington's disease, he has an average 12–15 years until death [41]. Along with its genetic profile, Huntington's disease affects about 5–10 people in 100 thousand [42]. To fully demonstrate Diseasecard's capabilities we setup three use cases, targeting users with distinct needs, in the context of Huntington's disease.

This study starts when a patient is seeking for diagnostic and starts being monitored by his general practitioner, who is traditionally not familiarized with Huntington's disease, but suspects this is the patient condition. Consequently, the clinician wants to learn about this disorder, searching for answers for the following questions:

1. What are Huntington's disease main features?
2. Are there any ICD terms for this disease?
3. What laboratories perform genetic tests for this disease?
4. What are the most relevant Huntington's disease publications?

Once the clinician has a better understanding concerning Huntington's disease, he proposes his patient for genetic analysis in a nearby institute. However, researchers working with this patient may also be unaware of the deep genotype characteristics of this disorder. To further understand the genomic and proteomic scope of this disease, the researcher starts by exploring answers to multiple questions. Some of these are:

5. What are the underlying genes associated with Huntington's disease?
   a. What are the gene names?
   b. What are known gene mutations?
6. What are the proteins coded by these genes?
   a. What is the 3D structure of these proteins?
7. In what pathways are the genes for this disease involved?

At last we need to consider the patient perspective, of someone who was recently diagnosed with an unknown disease, and that is looking into understanding what is happening in his organism. The patient searches for answers for the following questions to learn more about Huntington's disease:

8. Where can I get a description of this disease?
9. Are there any clinical trials open for this disease?
   a. What are the results of previous clinical trials?
10. Are there any patient registries or biobanks for Huntington's disease?

Finding the answers for these questions is not a straightforward process. Without a tool like Diseasecard, clinicians, researchers and patients will loose precious hours browsing Google, Wikipedia, OMIM, Orphanet or UniProt. This is a rather inefficient and ineffective endeavors. Diseasecard's streamlines this workflow. Starting by typing "Huntington" in Diseasecard's homepage search box, the entry for Huntington's disease appears almost immediately on the top results (http://bioinformatics.ua.pt/diseasecard/entry/143100). From there, the disorder workspace provides quick access to multiple web resources where users can find the answers to their questions:

1. Huntington's disease clinical features aptly start with the mention of classic signs of progressive chorea, rigidity and dementia. Clicking the disease name on Huntington's disease entry loads the disorder OMIM page in LiveView, where this information is highlighted.
2. On Diseasecard's navigation map and tree (Ontology node), the ICD10 node shows one link to the ICD version 10 "G10" term, entitled "Huntington disease".
3. Orphanet is the best resource to find genetic testing laboratories and is also linked in Diseasecard (Disease node). Orphanet lists around 180 diagnostic testing laboratories covering almost the entire Europe, from Portugal to Finland.
4. Pubmed is the key resource for relevant scientific publications. Diseasecard links directly to Pubmed's search engine (Literature node), where this publication list can be retrieved.
5. HGNC genes are loaded from OMIM's morbid map, thus playing a key role in Diseasecard's integration network. From the disease workspace, we can access HTT HGNC page (Locus node), Huntington's disease approved gene.
   a. From the previous page, we learn that HTT gene is denominated "huntingtin".
   b. WAVe is a gene-centric web portal collecting links for multiple locus specific databases [43]. WAVe access is provided in Diseasecard (Variation node) and it lists two locus-specific databases for the HTT gene, where we can find the variants in an LOVD [44] instance curated by Weilleke van Roon-Mom.
6. Huntington's disease proteins can be inferred from underlying genes of this disease. In Diseasecard's navigation interface, we learn that UniProt entry P42858 (HD_HUMAN) is associated with Huntington's disease (Protein node).
   a. Like UniProt, PDB is available in Diseasecard and this resource includes multiple 2D and 3D models portraying this protein.
7. Using KEGG (Pathway node), Diseasecard delivers easy access to the Huntington's disease pathway.
8. Like the clinical features, a disorder description is available in Huntington's disease OMIM entry (Disease node).
9. Clinical Trials detail studies and analysis over a variety of cohorts. The Study node lists two open Clinical Trials (as of early 2013), NCT01597128 at the University of Kentucky, USA; and NCT01065220 at the Medical University of Vienna, Austria.
   a. The Clinical Trial NCT00491842 has already finished and the collected data is also available in Diseasecard.
10. The previously mentioned Orphanet database also lists available patient registries and biobanks for Huntington's disease, easing the tasks of accessing and contacting these sites spread throughout Europe.

Diseasecard's focused environment provides a broad amount of connections where the answers to critical clinical and research questions can be answered. In comparison to using multiple applications, Diseasecard's always in-context browsing environment dynamically improves the end users data exploration workflow.

4.2. Integrating Diseasecard's data

The European Huntington disease network is a large-scale patient registry, with multiple centres spread throughout Europe focused on creating a wide patients community [45]. Taking in

account the amount and variety of knowledge that is offered to patients, clinicians and researchers of this database, and, additionally, the scope of knowledge that could be available, the inclusion of external data provided by Diseasecard is a welcome addition.

Despite being a closed system, we can assert from the various documentation available that this patient registry could be improved with the introduction of further clinical-oriented information. This information can be in the form of related ICD terms or links to Orphanet database entries, both containing relevant information for clinical practice. For instance, ICD classifications are already widely used in multiple hospital information systems [46–48].

The following SPARQL query can be sent to Diseasecard's endpoint, at http://bioinformatics.ua.pt/diseasecard/sparql, retrieving a unified list of ICD and MeSH terms, for the OMIM entries regarding Huntington's disease (#143100) and the "huntingtin" gene (*613004).

```
PREFIX coeus: <http://bioinformatics.ua.pt/coeus/>
PREFIX diseasecard: <http://bioinformatics.ua.pt/
  diseasecard/resource/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT DISTINCT ?icd {

  ?item dc:title ?icd.
  ?item coeus:hasConcept
  diseasecard:concept_ICD10.
  ?item coeus:isAssociatedTo ?orpha.
  ?orpha coeus:isAssociatedTo ?omim.
  ?omim coeus:hasConcept
  diseasecard:concept_OMIM.
  { ?omim diseasecard:omim ''143100'' }
  UNION
  { ?omim diseasecard:omim ''613004'' }
}
ORDER BY ASC(?icd)
```

The results list the ICD10 identifier "G10" as the term matching Huntington's disease. The query can be tested at http://bioinformatics.ua.pt/diseasecard/api/sparqler/. While SPARQL may have a steep learning curve, its advanced features make it the most complete query language for accessing any semantic knowledge base. Furthermore, results can be obtained in CSV, XML or JSON, making the use of these data in any programming language very straightforward.

While these mappings are also provided through several other services, Diseasecard's interoperability API offers a bigger variety of identifiers than most common systems. Querying UniProt or PDB identifiers is a similar process to querying the detailed ICD and Orphanet entries.

4.3. Future perspectives

Diseasecard can be seen as an initial step towards a comprehensive integrative "omics" suite. This will make it a key player in future large-scale research projects, acting as a channel for delivering a rich set of connections to rare genetic diseases knowledge.

A vital enhancement for future Diseasecard developments regards the inclusion of deeper semantic relationships amongst aggregated data. Whereas all individuals are connected through similar predicates in the current version, future iterations will comprise new rich connections between particular individuals, obtained from data mining workflows [49] and new scientific discoveries [50,51]. New metadata will improve Diseasecard's genomics perspective, with annotations for relationships regarding diseases, genes and proteins interactions [52]; and the clinical perspective, with annotations mined from electronic medical records [53] and specialized patient registries [54].

5. Conclusions

We presented Diseasecard, a portal for rare diseases researchers, clinicians and patients. Diseasecard features a rich semantic knowledge base to deliver a lightweight holistic perspective over the wealth of genetic diseases information stemming from the growing number of "omics" research projects.

Diseasecard's results are significant in at least three major respects. (1) The use of semantic web technologies to collect connections to the most relevant resources regarding rare diseases is a pivotal step in the future of data integration and interoperability. (2) The available in-context research features – full text search and LiveView augmented browsing – enable full access to external resources within each disease workspace. (3) At last, making all data available for further inclusion in other systems, whether through LinkedData or the SPARQL endpoint, empowers other developers to enrich their systems with a myriad of connections to the most relevant rare diseases resources.

The new Diseasecard represents a milestone towards semantic interoperable rare diseases knowledge, and is publicly available online at http://bioinformatics.ua.pt/diseasecard/.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under Grant agreement No. 200754 – the GEN2PHEN project, and under Grant agreement No. 305444 – the RD-Connect project.

References

[1] Nabarette H, Oziel D, Urbero B, Maxime N, Aymé S. Use of a directory of specialized services and guidance in the healthcare system: the example of the Orphanet database for rare diseases. Rev d'épidémiol santé publique 2006;54:41.
[2] EURORDIS. What is a rare disease? 2012.
[3] Aronson JK. Rare diseases and orphan drugs. Br J Clin Pharmacol 2006;61:243–5. <http://dx.doi.org/10.1111/j.1365-2125.2006.02617.x>.
[4] Wastfelt M, Fadeel B, Henter JI. A journey of hope: lessons learned from studies on rare diseases and orphan drugs. J Intern. Med. 2006;260:1–10. <http://dx.doi.org/10.1111/j.1365-2796.2006.01666.x>.
[5] Seoane-Vazquez E, Rodriguez-Monguio R, Szeinbach SL, Visaria J. Incentives for orphan drug research and development in the United States. Orphanet J Rare Dis 2008;3:33.
[6] Schieppati A, Henter JI, Daina E, Aperia A. Why rare diseases are an important medical and social issue. Lancet 2008;371:2039–41. <http://dx.doi.org/10.1016/S0140-6736(08)60872-7>.
[7] Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 2010;31:631–55. <http://dx.doi.org/10.1002/humu.21260>.
[8] Zhang L, Pei Y-F, Li J, Papasian CJ, Deng H-W. Improved detection of rare genetic variants for diseases. PLoS ONE 2010;5:e13857. <http://dx.doi.org/10.1371/journal.pone.0013857>.
[9] Zhang L, Pei Y-F, Li J, Papasian CJ, Deng H-W. Efficient utilization of rare variants for detection of disease-related genomic regions. PLoS ONE 2010;5:e14288. <http://dx.doi.org/10.1371/journal.pone.0014288>.
[10] Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol. Drug Saf. 2011;20:1–11. <http://dx.doi.org/10.1002/pds.2053>.
[11] Gowthaman R, Gowthaman N, Rajangam MK, Srinivasan K. Database of neurodegenerative disorders. Bioinformation 2007;2:153.
[12] Aartsma-Rus A, Van Deutekom JC, Fokkema IF, Van Ommen GJB, Den Dunnen JT. 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 2006;34:135–44. <http://dx.doi.org/10.1002/mus.20586>.
[13] 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:D514–7. <http://dx.doi.org/10.1093/nar/gki033>.

[14] Amberger J, Bocchini C, Hamosh A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat 2011;32:564–7. <http://dx.doi.org/10.1002/humu.21466>.
[15] Oliveira JL, Dias GMS, Oliveira IFC, Rocha PDNSd, Hermosilla I, Vicente J, et al. DiseaseCard: a web-based tool for the collaborative integration of genetic and medical information. In: 5th International symposium, ISBMDA 2004: biological and medical data analysis; 2004. p. 409–17. <http://dx.doi.org/10.1007/978-3-540-30547-7_41>.
[16] Oliveira I, Oliveira J, Sanchez J, Lopez-Alonso V, Martin-Sanchez F, Maojo V, et al. Grid requirements for the integration of biomedical information resources for health applications. Methods Inf Med 2005;44:161–7.
[17] Gardner SP. Ontologies and semantic data integration. Drug Discovery Today 2005;10:1001–7. <http://dx.doi.org/10.1016/S1359-6446(05)03504-X>.
[18] Pasquier C. Biological data integration using Semantic Web technologies. Biochimie 2008;90:584–94. <http://dx.doi.org/10.1016/j.biochi.2008.02.007>.
[19] Lopes P, Oliveira JL. COEUS: "semantic web in a box" for biomedical applications. J Biomed Semantics 2012;3:1–19. <http://dx.doi.org/10.1186/2041-1480-3-11>.
[20] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). Nucleic Acids Res 2005;33:D154–9. <http://dx.doi.org/10.1093/nar/gki070>.
[21] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. New developments in the InterPro database. Nucleic Acids Res 2007;35:D224–8. <http://dx.doi.org/10.1093/nar/gkl841>.
[22] Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 2010;38:D161–6. <http://dx.doi.org/10.1093/nar/gkp885>.
[23] Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, et al. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 2013;41:D475–82. <http://dx.doi.org/10.1093/nar/gks1200>.
[24] Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc 2000;88:265–6.
[25] Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, Pohlenz H-D, et al. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 2005;21:418–20. <http://dx.doi.org/10.1093/bioinformatics/bti010>.
[26] Mottaz A, Yip Y, Ruch P, Veuthey A. Mapping proteins to disease terminologies: from UniProt to MeSH. BMC Bioinformatics 2008;9(suppl. 5):S3. <http://dx.doi.org/10.1186/1471-2105-9-S5-S3>.
[27] MiličićBrandt M, Rath A, Devereau A, Aymé S. Mapping orphanet terminology to UMLS. In: Peleg M, Lavrač N, Combi C, editors. Artificial intelligence in medicine. Berlin Heidelberg: Springer; 2011. p. 194–203. <http://dx.doi.org/10.1007/978-3-642-22218-4_24>.
[28] Arrais J, Pereira J, Oliveira JL. GeNS: A biological data integration platform. In: Brojack B, editor. ICBB 2009, international conference on bioinformatics and biomedicine. Venice: WASET, World Academy of Science, Engineering and Technology; 2009.
[29] Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008;41:706–16.
[30] Lopes P, Oliveira JL. COEUS: a semantic web application framework. In: Proceedings of the 4th international workshop on semantic web applications and tools for the life sciences: ACM; 2011. p. 66–73. <http://dx.doi.org/10.1145/2166896.2166915>.
[31] Lopes P, Oliveira JL. A semantic web application framework for health systems interoperability. In: Proceedings of the first international workshop on managing interoperability and complexity in health systems: ACM; 2011. p. 87–90. <http://dx.doi.org/10.1145/2064747.2064768>.
[32] Lopes P, Oliveira JL. Towards knowledge federation in biomedical applications. In: Proceedings of the 7th international conference on semantic systems: ACM; 2011. p. 87–94. <http://dx.doi.org/10.1145/2063518.2063530>.
[33] Zhu Y, An L, Liu S. Data updating and query in real-time data warehouse system. In: Computer science and software engineering, 2008 international conference on 2008. p. 1295–7. <http://dx.doi.org/10.1109/CSSE.2008.78>.
[34] Reddy SSS, Reddy LSS, Khanaa V, Lavanya A. Advanced techniques for scientific data warehouses. In: International conference on advanced computer control, ICACC2009. p. 576–80. <http://dx.doi.org/10.1109/ICACC.2009.145>.
[35] Chaudhuri S, Dayal U. An overview of data warehousing and OLAP technology. ACM Sigmod Record 1997;26:65–74. <http://dx.doi.org/10.1145/248603.248616>.
[36] Weibel S. The dublin core: a simple content description model for electronic resources. Bull Am Soc Inform Sci Technol 1997;24:9–11. <http://dx.doi.org/10.1002/bult.70>.
[37] Pan J, Horrocks I. RDFS(FA) and RDF MT: Two Semantics for RDFS. In: Fensel D, Sycara K, Mylopoulos J, editors. The semantic web – ISWC 2003. Berlin (Heidelberg): Springer; 2003. p. 30–46. <http://dx.doi.org/10.1007/978-3-540-39718-2_3>.
[38] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics 2007;8:S2. <http://dx.doi.org/10.1186/1471-2105-8-S3-S2>.
[39] Bizer C. The emerging web of linked data. Intell Syst IEEE 2009;24:87–92. <http://dx.doi.org/10.1109/mis.2009.102>.
[40] Vonsattel JPG, DiFiglia M. Huntington disease. J Neuropathol Exp Neurol 1998:57.
[41] Walker FO. Huntington's disease. The Lancet, vol. 369. p. 218–28. <http://dx.doi.org/10.1016/S0140-6736(07)60111-1>.
[42] Driver-Dunckley E, Caviness J. Huntington's disease. In: Schapira AHV, editor. Neurology and clinical neuroscience. Mosby Elsevier; 2007. p. 879–85.
[43] Lopes P, Dalgleish R, Oliveira JL. WAVe: web analysis of the variome. Hum Mutat 2011:32. <http://dx.doi.org/10.1002/humu.21499>.
[44] Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. LOVD v. 2.0: the next generation in gene variant databases. Hum Mutat 2011;32:557–63. <http://dx.doi.org/10.1002/humu.21438>.
[45] Orth M. Network TEHsD. Observing huntington's disease: the European Huntington's disease network's REGISTRY. J Neurol Neurosurg Psychiatry 2011;82:1409–12. <http://dx.doi.org/10.1136/jnnp.2010.209668>.
[46] Hougland P, Nebeker J, Pickard S, Van Tuinen M, Masheter C, Elder S, et al. 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); 2008.
[47] Stausberg J, Hasford J. Identification of adverse drug events: the use of ICD-10 coded diagnoses in routine hospital data. Dtsch Arztebl Int 2010;107:23–9. <http://dx.doi.org/10.3238/arztebl.2010.0023>.
[48] Stausberg J, Hasford J. Drug-related admissions and hospital-acquired adverse drug events in Germany: a longitudinal analysis from 2003 to 2007 of ICD-10-coded routine data. BMC Health Services Res 2011;11:134. <http://dx.doi.org/10.1186/1472-6963-11-134>.
[49] Campos D, Matos S, Oliveira J. Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics 2013;14:54. <http://dx.doi.org/10.1186/1471-2105-14-54>.
[50] Arrais J, Oliveira J. Using biomedical networks to prioritize gene-disease associations. Open Access Bioinf 2011;3:123–30.
[51] Rosa N, Correia MJ, Arrais JP, Lopes P, Melo J, Oliveira JL, et al. From the salivary proteome to the OralOme: comprehensive molecular oral biology. Arch Oral Biol 2012. <http://dx.doi.org/10.1016/j.archoralbio.2011.12.010>.
[52] Nunes T, Campos D, Matos S, Oliveira JL. BeCAS: biomedical concept recognition services and visualization. Bioinformatics 2013. <http://dx.doi.org/10.1093/bioinformatics/btt317>.
[53] Oliveira JL, Lopes P, Nunes T, Campos D, Boyer S, Ahlberg E, et al. The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. Pharmacoepidemiol Drug Saf 2012. n/a–n/a. <http://dx.doi.org/10.1002/pds.3375>.
[54] Bellgard MI, Macgregor A, Janon F, Harvey A, O'Leary P, Hunter A. A modular approach to disease registry design: Successful adoption of an internet-based rare disease registry. Hum Mutat 2012. <http://dx.doi.org/10.1002/humu.22154>.