Towards pharmacogenomics knowledge discovery with the semantic web

Michel Dumontier and Natalia Villanueva-Rosales

Submitted: 1st August 2008; Received (in revised form): 10th November 2008

Abstract
Pharmacogenomics aims to understand pharmacological response with respect to genetic variation. Essential to the delivery of better health care is the use of pharmacogenomics knowledge to answer questions about therapeutic, pharmacological or genetic aspects. Several XML markup languages have been developed to capture pharmacogenomic and related information so as to facilitate data sharing. However, recent advances in semantic web technologies have presented exciting new opportunities for pharmacogenomics knowledge discovery by representing the information with machine understandable semantics. Progress in this area is illustrated with reference to the personalized medicine project that aims to facilitate pharmacogenomics knowledge discovery through intuitive knowledge capture and sophisticated question answering using automated reasoning over expressive ontologies.

Keywords: pharmacogenomics; semantic web; XML; OWL; ontology design; data integration

INTRODUCTION
Individuals respond differently to drugs sometimes unpredictable results. Personalized medicine aims to provide the safest and most effective therapeutic strategy based on genetic and physiological factors [1]. Genetic variability is recognized as a major contributor to therapeutic response in patients as it affects pharmacokinetics (how the body metabolizes a drug) and pharmacodynamics (how the drug acts upon the body). Already, clinical trials now consider a limited set of genetic variants so as to determine which subsets of the population react favorably or adversely to drug treatments [2]. These trials generally pursue pharmacogenetic strategies to study the drug response with respect to variation within a single gene, but with the availability of high throughput technologies, pharmacogenomic studies consider the modulation of a drug response across the entire genetic complement of an individual. Pharmacogenetic and pharmacogenomic information are largely locked up in the biomedical literature, but significant efforts such as that of the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) [3] are underway to curate this information into structured knowledge. Crucial to the success of personalized medicine will be the development of decision support tools that can not only capture this complex knowledge, but also answer questions and make predictions about optimal therapeutic treatments given the background of an individual. Realizing this aspect of knowledge management requires the integration of Health Care and Life Sciences (HCLS) so as to facilitate knowledge discovery across these diverse and wide ranging areas.

KNOWLEDGE DISCOVERY
Knowledge discovery often involves a guided search using implicit information between the large volumes of data to derive new knowledge. For instance, we might be interested in determining whether several personalized drug treatments act on pathways in a conflicting manner. The process of knowledge discovery necessitates, in the very least, data integration and semantic annotation that enable more sophisticated data mining. This article addresses these tasks by focusing on ontology design, population and question answering.

Corresponding author. Michel Dumontier, Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6. Tel: 613-520-2600 x4194; Fax: 613-520-3539; E-mail: michel_dumontier@carleton.ca

Dr. Michel Dumontier is a professor of Bioinformatics at Carleton University. His research interests involve the development of a computational platform for personalized medicine.

Natalia Villanueva-Rosales is a PhD candidate in Computer Science. She holds an MSc in AI and Computer Science bachelor degrees. Her main interests include knowledge discovery and the Semantic Web.
Data integration in the HCLS has been enormously facilitated by the use of the eXtensible Markup Language (XML), which provides a validation procedure for document structure based on a corresponding XML schema (XMLS). HCLS researchers have capitalized on XML to provide data exchange standards (Table 1). These standards have enabled the development of XML-compliant tools to submit pharmacogenetic data [4], automatically populate relational databases [5], create and run biochemical simulations [6, 7] and visualize interaction networks [8].

OWL ontologies provide a better representation

While XML documents can be queried via path expressions (XPATH), the lack of semantics precludes any other querying based on inference (e.g. to ask for all molecules and include proteins in the result, knowing that a Protein is a type of Molecule). In contrast, formal knowledge representation languages being developed within the context of the World Wide Web Consortium (W3C), as a part of the global Semantic Web initiative [9], include the Resource Description Framework (RDF), RDF Schema (RDFS) and highly expressive Web Ontology Language (OWL [10]). These not only build on the syntactic interoperability of XML, but also add machine understandable semantics that support symbolic data mining methods. In particular, OWL-DL is a variant that is based on a family of description logics (DL) that facilitates the description of complex concepts from simpler ones with an emphasis on decidability of reasoning tasks [11]. In other words, a feature of DL is that reasoning tasks terminate after a finite amount of time and that the inferences drawn are valid and complete. OWL-DL reasoners (e.g. Pellet [12], FaCT++) [13]) offer a number of useful services including consistency checking (for non-contradictory statements), classification (computing the complete ‘is a’ class hierarchy) and realization (classifying individuals into their most specific category). Together, these reasoning tasks allow question answering about explicit or inferred knowledge across vastly different domains of knowledge [14–17]. An introduction to the OWL language can be found here [18]. Thus, OWL-DL offers a promising framework for the design of highly expressive ontologies to support knowledge representation and reasoning for pharmacogenomics.

An ontological buffet

Ontologies explicitly describe and relate objects using formal, logic-based representations that a machine can understand. Ontologies already play an important role in managing HCLS terminology. The Open Biomedical Ontologies (OBO) is a portal of biological/medical ontologies that includes the popular Gene Ontology (GO) [19]. By providing a standardized vocabulary, OBO-controlled vocabularies and taxonomies are used in the annotation of biological information, which helps make information more accessible for computer interpretation. Through the OBO Foundry effort [20], OBO ontologies are being re-factorized for conformance to basic principles and mapped to the Basic Formal Ontology (BFO), an ontology that provides distinction between objects and processes and can be linked using basic relations. Together, they provide an enhanced platform to describe and semantically annotate domain specific knowledge, and enable making queries that retrieve information from diverse domains.

Table 1: Selected XML-based data representations for the HCLS. XMLS provides a basic document structure and inherent validation procedure that facilitates data sharing and supports tool interoperability in the HCLS

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Chemical Markup Language (CML) [9]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene expression</td>
<td>Microarray and Gene Expression Markup Language (MAGE-ML) [10]</td>
</tr>
<tr>
<td></td>
<td>Gene Expression Omnibus (GEO) [11]</td>
</tr>
<tr>
<td>Genotype</td>
<td>NCBI Database of Genetic Variation (dbSNP) [12]</td>
</tr>
<tr>
<td>Genotype/phenotype</td>
<td>Haplotype Map (HapMap) [13]</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>Pharmacogenetic phenotypes</td>
</tr>
<tr>
<td></td>
<td>Pharmacogencic phenotypes</td>
</tr>
<tr>
<td>Biochemical simulation</td>
<td>Systems Biology Markup Language (SBML) [16]</td>
</tr>
<tr>
<td>Clinical</td>
<td>Clinical Laboratory Procedure (CLP) [17]</td>
</tr>
<tr>
<td></td>
<td>Clinical Data Interchange Standard (CDISC) [18]</td>
</tr>
</tbody>
</table>
work is that it begins to describe an extensive knowledge representation containing 70 core classes and incorporates over 40,000 concepts across seven ontologies that span genomic variation, methods, disease, phenotypes and chemicals. It has been applied in knowledge discovery scenarios involving genotype–phenotype relationships in a familial hypercholesterolemia dataset [22]. However, a major challenge with SO-Pharm is that the sheer number of classes and relations makes it computationally expensive to use and leads to significantly higher complexity for knowledge composition. This will challenge users (researchers, clinicians and patients) in asserting knowledge or making routine queries.

To minimize the overall complexity, we sought to create a simple, but as effective knowledge representation for pharmacogenomics knowledge [23]. Our approach identified key required terminology and relations from use cases as well as the PharmGKB XML schema (XMLS). The resulting Pharmacogenomics Ontology (PO) identifies 40 core concepts spanning drugs, genotypes, phenotypes and drug treatments. We then created a knowledge base by populating the ontology using PharmGKB web services and demonstrated its utility in answering sophisticated questions about pharmacogenomics knowledge.

PO design
Our approach to ontology design [14, 24] essentially consists of the following steps.

Step 1. Define the scope and requirements of the ontology
The scope and requirements of an ontology are largely defined by a set of use cases, or usage scenarios. They help establish the interests of a wide range of individuals, in our case, that of a researcher, a doctor or a patient. For instance, a researcher would be interested in the specific gene–drug–outcome relationship, while a doctor would be interested in dosage given the disease and genotype. Use cases would also indicate how the users might interact with the system. Here, a good system would allow users to capture knowledge in a fairly natural manner, and would also allow them formulate queries in a manner that is somewhat intuitive. Thus, use cases define the standards to which an ontology may be evaluated. Our use cases also identify the kinds of questions that a user might ask.

Some of the questions we considered included: What is the most effective drug treatment for an individual with a given genetic profile that suffers from a particular disease? Which drugs yield side effects? Which drugs lead to variable outcomes depending on the genetic profile? The quality of the ontology depends on its ability to satisfy the variety of usage scenarios.

Step 2. Create primitive ontology from necessary concepts
Ontologies describe concepts and relations, for which the intent and meaning is clearly expressed. Essential concepts include drugs, drug treatments, genes, gene variants, SNPs, alleles, drug–gene interactions, drug-induced side effects, biomedical measures and clinical outcomes. Each is assigned an English name (annotated with rdfs:label) and provided with clear and precise definition (annotated with rdfs:comment). Concepts defined in the ontology will be identified by italicized prefix po:, such that po:Drug indicates the concept of a drug, as defined in the PO.

The PO [25] includes concepts that are hierarchically organized based on set inclusion or the ‘is a’ criteria. For example, every gene–drug interaction is a type of drug interaction. During this process, we ensure non-redundancy of concepts (synonymy) because each has distinguishable characteristics from its parents or siblings. In line with general normalization techniques [26], ontological terms are normally assigned to have a single parent.

An important aspect of pharmacogenomics is the variety of measurements used to infer health and drug response of individuals or groups of individuals. The PharmGKB-controlled vocabulary defines five main categories for semantic annotation of biomedical literature: Clinical outcome, Pharmacodynamics and drug responses, Pharmacokinetics, Molecular and cellular functional assays and Genotype. We formalized the relationships and incorporated these into a biomedical measure ontology [27]. The ontology contains 123 classes based on a variety of measures including those that are biophysical, clinical, cognitive, genotyping, metabolic, pharmacokinetic, pharmacodynamic, taxonomic and tissue-based. The ontology also distinguishes measures whose values are obtained from some procedure (e.g. a Hamilton depression score) from those measures whose values that are calculated from a comparison of values (e.g. a difference in Hamilton depression score). Measured values are represented with a single
datatype property named `hasValue`. Units [28] for measured values are assigned with the `hasUnit` object property.

We examined the PharmGKB XMLS [29] to ensure that additional associated knowledge could also be captured. First, we generated an OWL schema ontology from the PharmGKB XMLS using a generic XML2OWL stylesheet from the ReDeFer Project [30]. To conform with an entity treatment of qualities and measures, we transform each XML datatype into an object property and corresponding class using an additional XML style sheet [31]. We then mapped the XMLS entities to the OWL ontology and discovered that we also required support for entities such as pathways and publications. Such concepts were then added to the primitive ontology.

**Step 3. Map to an upper level ontology**
Upper level ontologies promise increased semantic interoperability and coherency by identifying the basic type of domain entities while imposing restrictions on the relationships that these entities may hold. Current upper level ontologies include the BFO [32], Suggested Upper Merged Ontology (SUMO) [33], General Formal Ontology (GFO) [34] and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [35]. Of these, we selected the BFO for its simplicity in distinguishing qualities, dispositions, objects and processes. Of course, many of these foundational ontologies have similar representations and efforts have been made to compatible ontological entities.

Classes defined in the primitive layer of the PO were mapped to these BFO classes in the complex layer (see Step 6). For example, the `po:BiomedicalMeasure` is a type of `bfo:Quality`, a `po:DrugTreatment` is a type of `bfo:Process` and a `po:Drug` is a type of `bfo:Object`.

**Step 4. Assign relations between concepts and attributes**
In addition to the is a hierarchy, a critical part of the modeling process is to establish which relations exist between entities. We use as a first resource the Basic Relation Ontology (BRO) [36] to hierarchically organize basic object relations [32, 37] for use in OWL ontologies. The BRO provides object-process, object-quality, mereological, spatial and temporal relations and features an `brosisRelatedTo` symmetric super-property that yields a reciprocal relationship between any pair of entities, hence allowing the easy discovery of any directly related entity.

**Step 5. Import ontologies and add complex class descriptions**
The complex layer (i) imports other ontologies so as to extend the set of available entities for knowledge curation and (ii) adds restrictions on class membership to ensure correct knowledge engineering. Currently, the complex layer [38] for the PO imports the NULO ontology [39] for upper level entities and relations as well as the biomedical measure ontology. The NULO ontology defines the domain and range values in terms of BFO concepts e.g. the `bro:hasParticipant` relation has a domain of `bfo:Occurrent` and a range `bfo:Continuant`, such that the type of the subject must be of `bfo:Occurrent` and the type of the object must be of `bfo:Continuant`. The NULO-constraints ontology [40] additionally restricts how these relations may be used. For instance, it defines that a valid parthood relationship may only occur between `bfo:Occurrent` or between `bfo:Continuant`, but not between these types. Further, we can apply OWL2 property characteristics such as reflexive, irreflexive, asymmetric, disjoint roles and role chains to augment the reasoning capability [41]. In particular, the `bro:hasParticipant` role chain (`bro:hasPart o bro:hasParticipant → bro:hasParticipant`), infers that the participants of sub-processes are also participants the process whole. Thus, this greatly simplifies pharmacogenomic knowledge representation by not requiring us to explicitly name all the participants of process whole (e.g. a drug treatment), when they will be named in process parts (e.g. gene–drug interaction). The resulting ontology focuses on drug treatment (Figure 1).

A complex class description places additional constraints on class membership (with necessary conditions) and creates new opportunities to infer class membership (with necessary and sufficient conditions). The complex layer here largely applies disjunction and existential restrictions. For instance, we can specify that a `po:DrugGeneInteraction` is a `po:Process` that necessarily has at least one `po:Drug` and one `po:Gene` as participants and that is different from a `po:DrugTreatment`. While certain restrictions may be deemed necessary, they may not necessarily be sufficient. For instance, a process in which a drug inhibits a transcription factor from binding at a gene likely would not be classified as a gene–drug interaction. However, it is possible to define the
necessary and sufficient conditions for certain concepts. For instance, a `po:GeneVariant` is fully defined by as a `po:Gene` that `bro:isVariantOf` some `po:Gene`. The reasoner can then identify all such genes and automatically classify them, which could be useful in asking questions about whether there are known SNPs for a gene.

**PO population**

Ontology population involves the instantiation of ontological entities from data via semantic annotation. Typically, this process uses predefined mappings between the ontology and the sources, i.e. between an OWL ontology and an XMLS, database schema, web services description, etc. For example, an XML element from a gene database could contain the value ‘HIF1alpha’ that would in turn be semantically annotated as an instance of a gene class as defined in an OWL ontology.

The most valuable resource for pharmacogenetics and pharmacogenomics knowledge is PharmGKB [3]. PharmGKB accepts submissions from its network of associates and also undertakes curation to identify the role of genes and their gene variants and classifies them in terms of pharmacokinetic, pharmacodynamics and clinical outcomes.

We populated the PO with genes, drugs and diseases with known pharmacogenomic relationships (as identified by PharmGKB) using web services of PharmGKB. The population was performed with following steps: (1) assigning names and namespaces, (2) asserting class membership and (3) assigning object and datatype properties. Class membership and the assertion of object and datatype properties were guided by the mappings between the PharmGKB database schema and the PO, which were encoded in the parsers. In the assignment of unique names, we assigned an arbitrary base (http://www.pharmgkb.org/) Universal Resource Identifier (URI) for PharmGKB entities that was obtained by querying PharmGKB web services. Publications identified by NCBI PMIDs were assigned a base identifier (http://purl.org/obo/owl/PMID#PMID_) favored by the OBO community. The populated ontology [42] contains nearly 4300 individuals (genes, diseases and drugs) having some pharmacogenomics importance (Table 2).

In populating the ontology using PharmGKB web services, it was often possible to correctly assign class membership, but the lack of explicit semantics between the entities makes it challenging to correctly assign an ontological relationship. Hence, in the absence of further information, we could only annotate with our most generic `bro:isRelatedTo` property. In such cases, it is not possible to query the knowledge base using more specific relations.

Although the schema ontologies obtained from the transformation of the PharmGKB represent the

![Figure 1: Drug treatment is the focus in the PO. The PO is built around the drug treatment, which is composed in part by drug–gene interactions and drug-induced side effects. These processes identify specific entities (genes, drugs, diseases and measures) as their participants. OWL2 role chains are used to infer relations between the drug treatments and the participants of drug–gene interactions and drug-induced side effects.](https://academic.oup.com/bib/article-abstract/10/2/153/184167)
structure of the data and lack of domain semantics, they can be used to populate the ontologies with mappings at the ontology level. For instance, we can automatically populate `po:Disease` from the XML data by asserting its equivalence to `pgkbxsd:Disease`. Similarly, the semantics of ontology schema object properties can be augmented by asserting these as object properties in the PO. In this way, we can semantically annotate concepts and data stored in the XML representation. Similarly, RDF sources can be integrated into OWL ontologies instance data. Hence, integration of data from essential pharmacogenomics resources (Table 3) can be executed automatically.

### Pharmacogenomics of depression

We sought to capture the pharmacogenomics of depression by extending our PharmGKB (ontology + PharmGKB data + curated data). We manually curated over 40 publications, of which many were initially identified by PharmGKB curators, and augmented with more recent literature. As of 1 January 2008, the pharmacogenomics of depression ontology [43] contains statements from 11 publications involving 45 genes/gene variants, 57 drugs annotated with 19 classes of antidepressants, 45 drug treatments, 47 drug–gene interactions, 29 clinical outcomes, 10 drug-induced side effects and 8 gene–disease interactions. Thus, this work extends the OWL-based PharmGKB by adding specific knowledge with respect to the therapeutic outcomes of drugs given specific genetic variants.

### Querying pharmacogenomics of depression

The addition of this new knowledge makes possible semantic question answering of some of our core competency questions brought on by our use-case requirements. Using Protegé 4, a freely available ontology management system, we show three queries that satisfy some of our use-case requirements.

(i) A psychiatrist diagnoses an elderly patient with depression based on the results of the Hamilton depression method. Considering the physical condition of the patient, postural hypotension must be avoided (a common side effect of many antidepressants). Using a semantic web application (Figure 2) that allows questions to be posed in the form of the Manchester Syntax [44], she imports the web-based ontology [43], enables the reasoner and asks for all drugs that have been used to treat depression (pgkb:PA447278) with a postural hypotension side effect of less than 5% of cases:

(Query 1) "drug" that ’is participant in’ some (‘drug treatment’ that ’is related to’ value 'Postural Hypotension' and ’is related to’ some (‘side effect rate’ that ’has value’ some double[<5] and hasUnit value percent))

Of the five drugs that are used to treat depression (Amitriptyline, Citalopram, Desipramine, Fluoxetine, Nortriptyline and Venlafaxine), only Nortriptyline (pgkb:PA450657) exhibits no side effect for postural hypertension. To answer the question, the reasoner invokes reasoning over a number of expressive elements of the ontology: (a) the `bro:hasParticipant role chain` infers drugs involved in

### Table 2: Statistics on ontology population from PharmGKB web services. The PO was populated from publications having been annotated with some pharmacogenomic importance using web services of PharmGKB

<table>
<thead>
<tr>
<th>Class</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>1396</td>
</tr>
<tr>
<td>SNP</td>
<td>847</td>
</tr>
<tr>
<td>Drug</td>
<td>269</td>
</tr>
<tr>
<td>Disease</td>
<td>156</td>
</tr>
<tr>
<td>Phenotype</td>
<td>75</td>
</tr>
<tr>
<td>Publication</td>
<td>71</td>
</tr>
<tr>
<td>Biomedical measure</td>
<td>127</td>
</tr>
<tr>
<td>Pathway</td>
<td>18</td>
</tr>
<tr>
<td>Other</td>
<td>1307</td>
</tr>
<tr>
<td>Total</td>
<td>4266</td>
</tr>
</tbody>
</table>

### Table 3: Essential pharmacogenomics resources. In building a truly diverse PharmGKB, existing resources not represented with semantic web technologies (RDF/OWL) will have to be transformed using special parsers (flat files) or with XML stylesheets for XML data. RDF data may be used to directly instantiate OWL ontologies

<table>
<thead>
<tr>
<th>Information</th>
<th>Resource</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharmacogenomics</td>
<td>PharmGKB [3]</td>
<td>XML(S)/web services</td>
</tr>
<tr>
<td>Pharmacogenomics</td>
<td>DrugBank [52]</td>
<td>Flat File</td>
</tr>
<tr>
<td>Chemical</td>
<td>PubChem [53]</td>
<td>XML(S)</td>
</tr>
<tr>
<td>SNP</td>
<td>dbSNP [53]</td>
<td>XML(S)</td>
</tr>
<tr>
<td>Gene/protein</td>
<td>RefSeq [53]</td>
<td>XML</td>
</tr>
<tr>
<td>Gene/protein</td>
<td>UniProt [54]</td>
<td>RDF</td>
</tr>
</tbody>
</table>
a drug treatment when this information is actually specified as part of a drug–gene interaction in which the drug is involved and (b) that ‘is related to’ is a super-property of all properties including has participant, which is used to assert the relationship between the drug treatments and the diseases as well as drug treatments and side effect rates.

(ii) Genotyping results indicate that the patient is homozygous (C/C) at the 3435 position of the ABCB1 gene (ABCB1_3435_C). She tries to find any Nortriptyline treatments involving this genotype with known side effect rates of postural hypotension:

(QUERY 2) ‘drug treatment’ that ‘is related to’ value ‘Postural Hypotension’ and ‘has participant’ value ‘Nortriptyline’ and ‘has participant'value’ABCB1_3435_C

Two results are returned, in which one (NortriptylineABCB1Treatment1) corresponds to the homozygous genotype of our patient and the other to a heterozygous genotype (NortriptylineABCB1Treatment2). As part of this drug treatment, the specific interaction between Nortriptyline and the ABCB1 gene variant is captured via a ‘drug metabolism by gene’ interaction. As in the previous query, the reasoner infers that the gene variant is a participant in the drug treatment via the hasParticipant role chain.

(iii) Most drugs, including Nortriptyline, are metabolized by the Cytochrome P450 family of drug metabolizing enzymes. The psychiatrist determines the genotype of the patient to be CYP2D6*4/CYP2D6*6, a known poor metabolizer variant. The recommended dose

Figure 2: Using Protegé 4 to ask a question. Users may interact with an OWL knowledge base using built-in DL Query Tab of Protegé 4. Queries are formulated as class expressions using the Manchester OWL syntax. In this screenshot, we ask for all (depression) drugs that where the rate of postural hypotension is less than 5%. Using the included Pellet OWL DL reasoner, the sole drug that satisfies the query is that of Nortriptyline.
may also be obtained by querying the knowledge base:

(QUERY 3) ‘dose recommendation’ that ‘has participant’ value CYP2D6*4 and ‘has participant’ value CYP2D6*6 and ‘has participant’ value ‘Nortriptyline’

The recommended dose is found to be 103 mg of Nortriptyline for this genotype. Although the representation of dose in the ontology involves at a minimum the therapeutic compound and the amount, it is nevertheless possible to augment this information with the genetic background that it corresponds to.

While these questions represent a small sample of questions that could be asked of the KB, it does illustrate that the representation allows questions to be asked from the perspective of any ontological entity. That is to say, we may ask about the relationships that are held by drug treatments, drugs, recommended doses, genetic variants, or any other entity. This gives extreme flexibility in question answering, provided that the knowledge is consistently represented, as we have done so here.

DISCUSSION
Towards personalized medicine
The development of a reasoning capable knowledge base for pharmacogenomics is an important step in realizing personalized medicine. Interest in personalized medicine will only grow with the recent introduction of rapid, low-cost genotyping methods to the clinical setting. While motivated to model the pharmacogenomics of depression, our framework is sufficiently generic to be useful for a wide variety of pharmacogenomics studies. A flexible PharmGKB may be useful from multiple perspectives: a clinical researcher may want to formulate or test hypotheses with respect to the mechanism of drug interaction, a doctor and patient will seek therapeutic options, a clinical trial management company will want to aggregate, mine and prepare their knowledge for FDA drug approval. Hence, establishing methods for knowledge acquisition and knowledge discovery in pharmacogenomics is of great importance.

Building the ontological base
The OBO boasts an increasing large collection of terminological resources, built by both groups and individuals that are of interest to the pharmacogenomics community. These include, but are not limited to, the three axes of the GO (biological process, molecular function and cellular component), human anatomy (CARO, FMA), human disease (DOID), phenotype (PATO, MP), chemicals/drug roles (CHEBI), molecular interactions (MI) and biomedical investigation (OBI). Other resources may also prove useful. For instance, SNOMED-CT [45] contains clinical terminology for use in electronic health records, and hence would provide valuable terminology for annotating the bedside aspect of pharmacogenomic medicine. Towards the molecular side of things, pharmaceutical researchers intent on describing the mechanism of drug action might be inclined to look at our work on representing chemical structure and function [46, 47].

Most OBO-controlled vocabularies are collections of hierarchically organized terminology, but their representation makes use of directed acyclic graphs, whose syntax and semantics are woefully inadequate. While there exists a formal mapping [48] from OBO to OWL, the direct transformation may lead to erroneous inferences because the OBO resource never committed to OWL semantics in the first place. Thankfully, the OBO Foundry effort aims to re-factor OBO ontologies to ensure compatible representations with the normal representation of (first order) logic-based ontologies. However, OBO Foundry is not, at this time, adopting OWL as the language for the representation of ontologies, and hence will not necessarily be using OWL reasoners to check consistency of the ontologies produced. Thus, while the OBO Foundry effort should certainly improve the accuracy of the resulting ontologies, the correctness of the ontology cannot be guaranteed.

In adopting OWL users gain the distinct benefit of interoperable ontologies stemming from the W3C standardization of the language syntax and semantics. Currently, OWL is being upgraded with increasingly expressive features that are not found in other languages, and that major tool providers already have support for. As early adopters, we expect that OWL ontologies will that support the representation of scientific knowledge [49] will form an important cornerstone of the semantic web.

The eternal battle: class versus instance
A major challenge in modeling pharmacogenomics knowledge is whether to represent an entity as a class or as an instance. This initial approach involves a
minimal ontological commitment, in the sense that PharmGKB data and curated data are treated as instances of basic types of pharmacogenomic entities. This is a necessary approach because relations between the class entities must be represented as an existential restriction (for every instance of the class, there exists at least one instance of another class expression that it is related to) or universal restriction (for every instance of the class, the only instances related to the instance are herein specified). In most cases, the knowledge that is to be represented does not fall into such strong statements. Nevertheless, this does not preclude the assertion of more specific class membership to existing ontologies of types. For instance, the drug instance, Nortriptyline, has been semantically annotated as a type of tricyclic anti-depressant, from a drug ontology. In addition, we and others (OBO Foundry/OBI) are currently investigating ways to represent specifications of protocols, or in our case, specifications of therapeutic prescriptions to be represented as classes. Such specifications could then be realized during real instances of drug treatments involving doctors, patients, drugs, etc., and therefore lead to the capture of day-to-day events.

Question answering
There are several ways to retrieve information from OWL knowledge bases. Class description queries, also known as DL Queries, are constrained to the concepts, properties and individuals contained in the ontology. Here, we represented our queries using the Manchester OWL syntax, which are English phrase-like sentences. As the semantic web matures, we expect new interfaces and languages to be used to formulate queries. Efforts to develop and promote controlled natural languages for OWL starting to bear fruit among core developers. In contrast, more powerful graph languages such as nRQL [50] or SPARQL-DL [51] create the means in which more sophisticated queries can be formulated by the expert user.

CONCLUSION
In this article, we show that the semantic web framework enables data integration and semantic annotation, two key components of knowledge discovery. While there exists significant debate in terms of best practices for the design of ontologies, we believe that simple, intuitive conceptualizations are motivated by clear requirements stemming from diverse sets of use cases and provide a persuasive means to evaluate the utility of resulting ontologies. In turn, the use of standardized formal knowledge representation languages such as OWL makes possible a dependable representation of information, whose syntax and semantics lead to a consistent interpretation. With a populated ontology in hand, new approaches to mine the knowledge base found on the semantic web will uncover new patterns in the information, which in turn can be captured as expressive ontologies. Our works to extend PharmGKB using the OWL framework with the curated information on the pharmacogenomics of depression demonstrates how third parties can contribute and augment to a growing structured knowledge base. Clearly, the semantic web forms the basis for a promising research area in pharmacogenomics knowledge discovery.

Key Points
- Knowledge discovery involves data integration and semantic annotation, which is made possible with ontologies designed using expressive knowledge representation languages such as OWL-DL.
- Reasoning-capable knowledge bases for pharmacogenomics facilitate knowledge discovery by enabling sophisticated queries over curated knowledge.
- Built with semantic web tools and technologies, the PharmGKB provides exciting possibilities to improve our collective understanding of the pharmacogenomics of human diseases.

FUNDING
CONACYT scholarship (#150581 to N.V.R.); NSERC Discovery Grant (to M.D.).

References
5. Mungall CJ, Emmert DB. A Chado case study: an ontology-based modular schema for representing


