IEEE Intelligent Systems
www.computer.org/intelligent

# Applying Semantic Web Technologies to Drug Safety Determination

*Susie Stephens, Alfredo Morales, and Matthew Quinlan*

Vol. 21, No. 1
January/February 2006

**Editor: Steffen Staab**
University of Koblenz-Landau
staab@uni-koblenz.de

# Applying Semantic Web Technologies to Drug Safety Determination

**Susie Stephens,** *Oracle*
**Alfredo Morales and Matthew Quinlan,** *Cerebra*

Identifying signals of events that lead to undesirable outcomes is historically one of the most challenging aspects of determining drug safety, both during the drug discovery and development process and once a drug is released to the market. The information required to identify signals resides in disparate and distributed data repositories located in different functional groups and even separate organizations. The Semantic Web provides new capabilities for data integration that exploits explicit semantics and well-defined ontologies. These technologies promise to simplify heterogeneous data integration and allow logic to infer additional insights from the data.

The Oracle RDF Data Model integrated with Cerebra Server is a composite solution that addresses the complexities of mediating information in drug safety. A use case illustrates the situation and how the composite solution might help.

## Drug discovery and development: An overview

Taking a drug to market typically takes companies 8 to 10 years and costs more than $800 million.[1] The process involves identifying a biological target (typically a protein molecule with a natural role in the organism) involved in a disease, designing and synthesizing a lead compound (a small synthetic molecule) that modifies the target's behavior, conducting preclinical tests of the compound's toxicity, and finally, testing the compound's efficacy and safety in clinical trials. Figure 1 provides an overview of the drug discovery and development process.

Science has made significant advances over the last decade—for example, in sequencing the human genome and in the ability to measure the expression of every gene within cells. Despite these advances, productivity in the pharmaceutical industry has declined.[2] Typically, companies enter more compounds into the drug discovery and development process, anticipating more compounds to become drugs in the end. However, this strategy hasn't succeeded. Many compounds have failed to become drugs late in the process owing to safety concerns. Safety and toxicity were the reasons for 46 percent of new drug candidate attrition at Aventis in 2003.[3] The failure of compounds toward the end of the process is costly, given the resources already invested. A 10 percent improvement in predicting failures before beginning large-scale Phase III clinical trials could save approximately $100 million in development costs.[1]

In the face of these challenges, the US Food and Drug Administration made recommendations on improving the efficiency of drug discovery and development, with particular regard to drug safety.[4] The primary recommendations encourage data sharing along the value chain and advocate stronger use of predictive technologies and biomarkers.

Drug discovery and development operate in a series of silos. For example, data generated in preclinical typically remains in preclinical and is accessible only to people working in this area. Furthermore, even when companies submit data to regulatory authorities for assessing a new drug application's safety, the preclinical data might be on separate media from other functions' data. It's unsurprising that pharmaceutical companies and regulatory authorities struggle to gain a comprehensive view of a compound's safety profile.

Data integration across functional areas is a prerequisite for effective drug safety determination. Life sciences data might contain information that signals potential toxicity for drug compounds further along the value chain, but the signal might be missed if data is distributed throughout an organization. Despite extensive clinical trial testing, rare adverse events and unforeseen interactions with coexisting clinical conditions or other drug therapies can easily escape detection. Therefore, post-market surveillance is required to identify such events.

## Data integration challenges

Data integration challenges in functional areas for the life sciences industry are well documented.[5,6] A key challenge is integrating in-house and third-party data. Difficulties stem from the abundance of publicly available data sources of varying quality, inconsistent data formats, data


1541-1672/06/$20.00 © 2006 IEEE
*Published by the IEEE Computer Society*

models, terminology, and identifiers. Furthermore, sources are generating new data at tremendous rates, and data models need to change frequently to reflect advances in scientific understanding.

Data integration between functional areas presents additional challenges of interpretation and semantics. Different functional areas frequently operate at different abstraction levels—for example, proteins might be referenced by individual splice variant, protein family, or protein complex. Acronym collision is also common—for example, the acronym PCR means *polymerase chain reaction* to molecular biologists and *phosophocreatinine* to endocrinologists.

Integrating data between user communities historically required costly and time-consuming specialized code to transform data, compounded for the global life sciences community by its many specialized dialects and consequently difficult data-exchange scenarios.

## Semantic Web for life sciences

The Semantic Web provides for defining and linking data to enable its more effective discovery, automation, integration, and reuse. It promises to help companies integrate disparate data both within and across functional areas, a promise that is generating strong interest from the life sciences community—for example, the World Wide Web Consortium (W3C) has established its first Semantic Web interest group to focus on Health Care and Life Sciences (www.w3.org/2001/sw/hcls). The W3C standards recommendations that underlie the Semantic Web include RDF and OWL.

The life sciences community has a rich history of using common vocabularies for annotating and integrating knowledge bases.[7,8] Vocabularies are increasingly available as OWL-based ontologies, providing a formal representation of knowledge about specific domains. Examples include the Gene Ontology (www.geneontology.org), BioPax (www.biopax.org), and the Unified Medical Language System (www.nlm.nih.gov/research/umls). Abstracting domain knowledge to an ontology layer avoids extensive reliance on custom procedural programming and the need to rewrite legacy code whenever a model, schema, or policy changes.

RDF, OWL, and other Semantic Web technologies in development combine to provide a more effective mechanism to integrate data across drug discovery and development functions, promising a more supportive environment for earlier detection of safety-related issues. Eric Neumann has described early candidate areas for integration using Semantic Web technologies.[9]

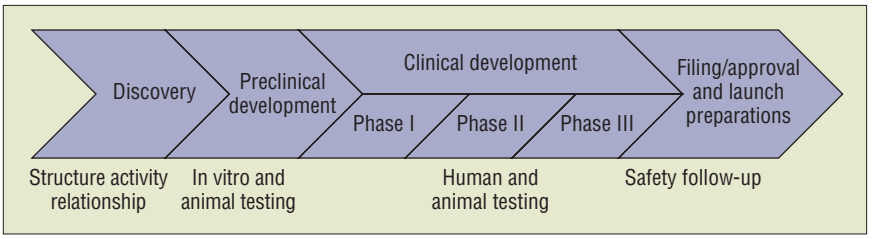


**Figure 1. The traditional process to discover and develop a new drug.**

Semantic Web standards have matured to the point where commercial software companies have implemented solutions. The latest release of the Oracle Database provides support as a repository for RDF-based information, including OWL. Oracle's support generated much attention from a range of organizations that wanted to manage RDF data in a secure, scalable, and highly available environment. In addition, the Oracle Database now provides a single framework for managing and querying relational, XML, and RDF data. Cerebra has brought to market a scalable and robust solution for semantic information mediation and metadata interpretation. Cerebra Server provides highly optimized, decidable, and provable reasoning for OWL knowledge bases.

## Oracle RDF Data Model

In Oracle Database 10g Release 2, a new object type stores RDF data.[10] The RDF data structure is made up of a *triple*, where each triple represents a statement about a resource using a subject-predicate-object syntax. This functionality builds on the Oracle Spatial Network Data Model, which was designed for storing, indexing, and querying graph data within the relational database management system. Oracle is the first mainstream commercial database to provide direct and native support for Semantic Web technologies. The implementation's primary goal was to provide a scalable, secure, and highly available data management platform for RDF-based applications.

The RDF Data Model consists of several key tables. RDF_MODEL$ is the system-level table for storing information on all of the RDF models in the database. RDF_VALUE$ is the table that stores the text values for each part of the triple and generates a unique VALUE_ID for the text entry. RDF_NODE$ stores the VALUE_ID for text values that participate in statements' subjects or objects, and RDF_LINK$ stores the triples for all of the RDF models in the database.

A key feature of the RDF Data Model is that it stores nodes only once, regardless of the number of times they participate in triples. The model always creates a new link when a user inserts a new triple. When a user deletes a triple from the RDF Data Model, the model deletes the corresponding link, but it deletes the nodes attached to this link only if no other connections to them exist. The RDF Data Model supports

- reification (treating RDF statements as resources so that users can assert statements about them),
- heterogeneous data types (object column values may be of different data types), and
- multiple representations of the same scalar value.

The RDF_MATCH table function lets users embed a graph query in a SQL query. This architecture provides the option of further processing the arbitrary graph pattern by joining it to relational or XML data in the database. RDF_MATCH provides support for inferencing based on RDF, RDF Schema, and user-defined rules.

The RDF Data Model is already integrated with numerous RDF- and OWL-based applications. However, we will focus on the integration with Cerebra Server.

## Cerebra Server

The Cerebra Suite provides a comprehensive and collaborative environment for representing, persisting, managing, and reasoning about OWL Description Logic (DL) ontologies.

Implementing a series of highly optimized tableaux algorithms, Cerebra Server makes it possible to load, manipulate, and draw inferences from knowledge expressed in

```
<owl:Class rdf:about="http://www.cdc.gov/nchs/icd9/dx#_250">
  <rdfs:subClassOf rdf:resource="http://www.cdc.gov/nchs/icd9/dx#Diagnosis"/>
  <rdfs:comment>This represents diagnosis of Diabetes Melitus</rdfs:comment>
</owl:Class>
<owl:Restriction>
  <owl:onProperty rdf:resource="#age"/>
  <owl:allValuesFrom rdf:resource="http://custom.com/datatype#adult_age_range"/>
  <owl:equivalentClass rdf:resource="#Adult_Patient"/>
  <rdfs:subClassOf rdf:resource="#Patient"/>
</owl:Restriction>
<owl:Class rdf:ID="Adult_Patient"/>
<owl:Restriction>
  <owl:onProperty rdf:resource="#has"/>
  <owl:allValuesFrom rdf:resource="http://www.cdc.gov/nchs/icd9/dx#_250"/>
  <owl:equivalentClass rdf:resource="#Adult_Patients_with_Diabetes"/>
  <rdfs:subClassOf rdf:resource="#Adult_Patient"/>
</owl:Restriction>
<owl:Class rdf:ID="Adult_Patients_with_Diabetes"/>
```

**Figure 2. Sample representation of an OWL Description Logic ontology.**

OWL, providing insight into the results of the inference process through XQuery, a standard XML technology.

A key capability of Cerebra Server is to act as an information mediation service, brokering data from disparate sources using OWL ontologies that describe domains, terminologies, business processes, and the data sources associated with them. Cerebra Server provides the interpretation necessary to ensure accurate contextual information and disambiguate queries issued against federated information. The ontologies act as a metadata layer on top of existing systems and provide unified, cohesive views into legacy information.

## Integrating Cerebra Server with the Oracle RDF Data Model

Cerebra Server has been integrated with the Oracle RDF Data Model to let it invoke SQL to perform RDF_MATCH queries against data stored in the Oracle RDF Data Model. The integration enables combining RDF inferencing capabilities that the Oracle RDF Data Model provides with OWL inferencing capabilities that Cerebra Server provides.

The combined reasoning enables applications to query the semantic mediation infrastructure at the appropriate fidelity and completeness levels. To illustrate, the OWL sample in figure 2 defines an Adult_Patient_with_Diabetes as an Adult_Patient (itself defined as a Patient with an age property in a defined range) who has a specific diagnosis. RDF triples may describe John Doe as a member of Patient with a specific diagnosis and age.

To answer a query "return all Patients" requires a simple lookup of triples from the RDF store. To fully answer a query "return all Adult_Patients_with_Diabetes" or "return Patients with a Diagnosis" requires preprocessing the query using OWL inferencing before retrieving all relevant instances. The level of inferencing that individual use cases require has trade-offs: full OWL-DL inferencing might lead to slower responses but complete answers, and a lower inferencing level might be faster but in some cases less complete. The integration with the Oracle Database provides the ability to select the required level of inferencing when querying Cerebra Server with the Oracle RDF Data Model. Furthermore, the integration simplifies mapping between Cerebra's RDF and OWL objects and the Oracle object-relational storage, as well as integrating RDF data with other enterprise data. Figure 3 highlights the product integration.

Oracle and Cerebra plan a tighter integration between the products in the future, which would let Cerebra Server treat RDF stored in Oracle Database 10g as instances of classes in the ontology. This would provide a highly scalable instance repository for reasoning. Plans also include a closer alignment of indexing rules expressed in the RDF Data Model and axioms in the

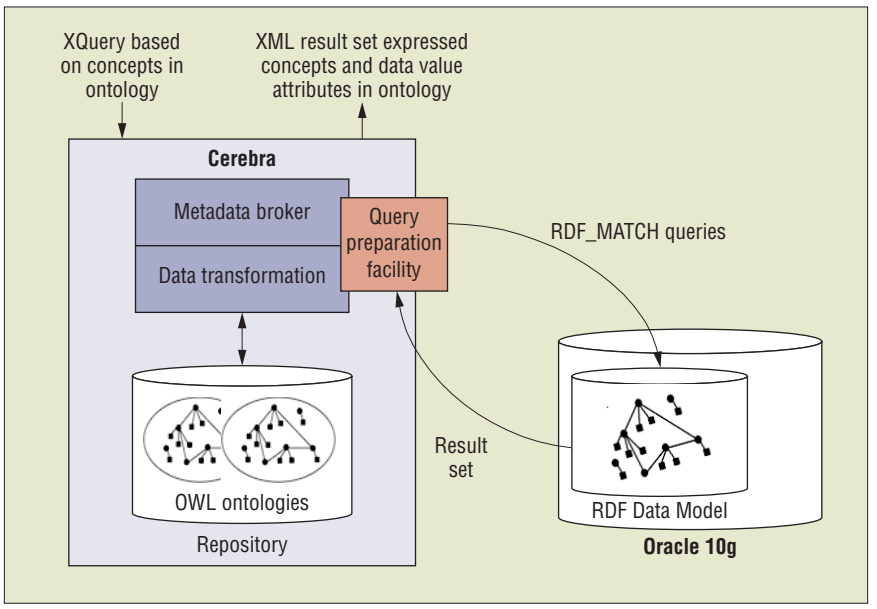


**Figure 3. Cerebra Server mediation of the Oracle RDF Data Model.**

ontologies that Cerebra Server manages, providing a mechanism to further reconcile the knowledge represented in the two systems.

## Drug safety use case

As compounds move along the drug discovery and development pipeline, companies make decisions at several points as to whether to pursue the compound further. To pass through a decision gate a compound must meet a series of criteria that a particular functional area predetermines.[11] More effective decision-making regarding a drug's safety profile would use all known information regarding the compound, target, and patient group. OWL inferencing (for term disambiguation and reconciliation for integrating data, and for complex definitions and classification), complemented where appropriate by rules-based inferencing, could help guide decisions on continuing to pursue a compound or withdrawing a drug from the market. The goal is to minimize the high rate of adverse events and medical errors.[12]

Figure 4 shows an example of a rule that uses data from many functional groups to help guide a physician on the best course of action when a patient in a clinical trial reports a skin rash. The rule shows many characteristics relating to the compound, its interaction with a target, and related toxicity data. No single condition in the rule would make the compound appear ideal, yet none is individually strong enough to warrant halting the trial. However, seeing all conditions in parallel might necessitate a stronger reaction. Table 1 shows details about the rule parameters.

Both the life sciences and healthcare industries use rules heavily. Partners Healthcare currently has 23,000 active clinical rules.[13] Many rules are used to identify safety signals in data. This is partly because the tolerance of adverse events differs according to the therapeutic condition—for example, childhood asthma compared to cancer. Rules also integrate data, guide portfolio investment decisions, and help companies meet international regulatory requirements. Rules can be generated from machine learning approaches, such as decision trees, or constructed by specialists with expertise in their field. Ideally, organizations within the life sciences community could share rules to help minimize adverse drug reactions.

We could use the Oracle RDF_MATCH capability to run the rule in figure 4 against data in the RDF Data Model provided all data were unified prior to running the query. Cerebra Server, when used with the Oracle RDF Data Model, would link multiple ontologies and disambiguate queries over federated information, as a mediation service. For example, when asked for all "rashes," the mediation service would use OWL restriction classes and inference to interpret that, in a local vocabulary, a rash is an inflammation of the skin. The mediation service would then retrieve the appropriate data regardless of whether the original modeler deemed the feature "dermatitis" or "rashes."

Conceptually, this process uses three ontology layers. The first layer is the data ontology, providing an interface between the database schema and the domain ontology. This layer isn't required if data is stored in the Oracle RDF Data Model. The second layer is the domain ontology, describing life sciences and healthcare domains. In our example, a subset of the SNOMED (Systematized Nomenclature of Medicine) terminology could represent diagnosis and procedures, while a separate ontology represents portions of LOINC (Logical Observation Identifiers, Names, and Codes) to describe laboratory findings. The Gene Ontology could represent genetic information, and an in-house ontology could represent the chemical structure information. With these two layers in place, it becomes possible to query data through the concepts in the ontologies instead of the traditional instance-matching approach, allowing queries to be more easily specialized and generalized. The third layer, the application ontology, describes a more specialized ontology based on either a particular application or a third-party world view. We can construct new concepts from, or map to, concepts in the domain ontology. We can then dynamically reclassify instances in the database on the basis of the axioms in the application ontology.

```
IF compound has >90% structural similarity to a failed compound
  AND compound binds to target with more than 5 SNPs
  AND therapeutic index is low
  AND histology indicated >5% incidence of liver necrosis in preclinical
  AND ALT reading is >2x above normal in Phase I
  AND therapeutic dose is >30 mg in Phase II
  AND >80% of patients with Cytochrome P450 2DE report skin rash in Phase III
THEN consider immediately stopping trials for those patients
```

**Figure 4. Rule for detecting safety signals that spans data sets along the pharmaceutical value chain. SNP is single nucleotide polymorphism; ALT is Alanine Aminotransferase.**

**Table 1. Parameters used along the drug discovery and development pipeline to determine a potential drug's toxicity risk.**

| Parameter | Significance | Department |
|---|---|---|
| Structural similarity of compounds | Similarity to a compound that failed owing to toxicity would indicate an increased risk. | Cheminformatics |
| SNP (single nucleotide polymorphism) | A SNP indicates DNA sequence variation. High variability might increase the range of response to the drug and therefore increases risk. | Bioinformatics |
| Therapeutic index | The ratio between the lethal dose of a drug for 50% of a population compared to the effective dose for 50% of the population. Low ratios indicate more risk. | Preclinical |
| Liver necrosis | Liver damage is the most common side effect of drugs. Liver necrosis would indicate cause for concern. | Preclinical |
| ALT (Alanine Aminotransferase) | A measure of liver toxicity. | Phase I clinical |
| Therapeutic dose | The higher the dose required, the more risk. | Phase II clinical |
| Cytochrome P450 2DE | Variant of a common enzyme that frequently metabolizes drugs. | Phase III clinical |

Ensuring drug safety is of paramount importance to the life sciences industry. It's critical that drugs are able not only to achieve the desired result but also to do so without harmful side effects. By identifying problems as early as possible in the drug discovery and development process, life sciences companies can avoid drug safety sagas, such as a recent example that concerned COX-2 inhibitors.[14] Unfortunately, drug safety problems are often revealed only during clinical trials or occasionally after marketing. These challenges are becoming more acute as medicines are targeted to defined patient populations.

The life sciences industry can use Semantic Web technologies to integrate data more effectively across all drug discovery and development business units, thereby providing a more supportive environment for the early detection of safety-related issues. Effective integration would enable genomic data and patient profiles to be more easily related to safety, thus providing

- a simpler framework for determining risk-benefit for individual patients in particular treatment regimens,
- a better mechanism to distribute new data relating to safety throughout the organization, and
- a better decision-making environment to determine which drugs to pursue.

Furthermore, Semantic Web inferencing capabilities enable an intelligent decision support system or autonomous agent to reason about combined domain-specific and industry-specific knowledge and act on the conclusions drawn from this inferencing process. ❐

## Acknowledgments

We thank the Oracle Spatial Development group for the implementation of the Oracle RDF Data Model. We also thank Otto Ritter of AstraZeneca for his assistance with this article.

## References

1. L.J. Lesko and J. Woodcock, "Translation of Pharmacogenomics and Pharmacogenetics: A Regulatory Perspective," *Nature Rev. Drug Discovery*, vol. 3, no. 9, 2004, pp. 763–769.
2. J.J. Herbst and K. Dickinson, "Automated High-Throughput ADME-Tox Profiling for Optimization of Preclinical Candidate Success," *Am. Pharmaceutical Rev.*, vol. 8, 2005, pp. 96–101.
3. Goldman Sachs, "An Analysis of Late Stage Success Rates: Biotech Products versus Pharmaceuticals," *Parexel's Pharmaceutical R&D Statistical Sourcebook 2005/2006*, M.P. Mathieu, ed., 2005, pp. 194.
4. US Food and Drug Administration, *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*, white paper, Mar. 2004; www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.
5. B. Donnelly, "Data Integration Technologies: An Unfulfilled Revolution in the Drug Discovery Process?" *Drug Discovery Today: Biosilico*, vol. 1, no. 1, 2003, pp. 59–63.
6. D. Searls, "Data Integration: Challenges for Drug Discovery," *Nature Rev. Drug Discovery*, vol. 4, no. 1, 2005, pp. 45–58.
7. Gene Ontology Consortium (M.A. Harris, et al), "The Gene Ontology (GO) Database and Informatics Resource," *Nucleic Acids Research*, vol. 32, 2004, pp. D258–D261.
8. S. de Coronado et al., "NCI Thesaurus: Using Science-Based Terminology to Integrate Cancer Research Results," *Medinfo 2004*, 2004, pp. 33–37.
9. E. Neumann, "Finding the Critical Path: Applying the Semantic Web to Drug Discovery and Development," *Drug Discovery World*, vol. 6, 2005, pp. 25–33.
10. C. Murray, *Oracle Spatial Resource Description Framework (RDF), 10g Release 2 (10.2)*, Oracle, 2005, http://download-west.oracle.com/otndocs/tech/semantic_web/pdf/rdfrm.pdf.
11. J.F. Pritchard et al., "Making Better Drugs: Decision Gates in Non-Clinical Drug Development," *Nature Rev. Drug Discovery*, vol. 2, no.7, 2003, pp. 542–553.
12. M. Pirmohamed et al., "Adverse Drug Reactions as Cause of Admission to Hospital: Prospective Analysis of 18 820 Patients," *British Medical J.*, vol. 329, 2004, pp. 15–19.
13. J. Glaser, "New Medicines: Can Innovation and Safety Coexist?" MIT-CBI Forum, 2005, http://cbi.mit.edu/forum05.php.
14. S. Frantz, "How to Avoid Another 'Vioxx,'" *Nature Rev. Drug Discovery*, vol. 4, no.1, 2005, pp. 5–7.

**Susie Stephens** is a principal product manager for Life Sciences at Oracle. Contact her at susie.stephens@oracle.com.

**Alfredo Morales** is a senior consultant at Cerebra. Contact him at alfredo.morales@cerebra.com.

**Matthew Quinlan** is director of product marketing at Cerebra. Contact him at matthew.quinlan@cerebra.com.