

# Semantic Data Integration For the Enterprise

*An Oracle White Paper*  
*June 2007*

**Note:**

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

# Semantic Data Integration for the Enterprise

The Semantic Web defines and links data in such a way that it can be used for more effective discovery, automation, integration, and re-use across various applications.

## INTRODUCTION

The Semantic Web is a project and vision of the World Wide Web Consortium to extend the current Web, so that “information is given well-defined meaning, better enabling computers and people to work in cooperation.”<sup>1</sup> This is important, as the mix of content on the web and in applications built using web architectures is shifting from exclusively human-oriented content to computer-mediated content. In the Semantic Web, data are defined and linked in a way that enables its use for more effective discovery, automation, integration, and re-use across various applications. Toward this end, the W3C has adopted standards and tools such as RDF and OWL to advance the use of semantic technologies.

The data representations defined by Semantic Web initiatives can be seen as the next step in the evolution of data management. One of the challenges of data management is the ability to share and analyze data stored by independent applications. Precursors to the semantic technology such as data exchange formats have maintained the distinction between data and the data describing the data (schema or metadata). Minimizing that distinction in data representation enables semantic technologies to move one step closer to data sharing and integration.

Oracle Database 11g now supports both RDF and OWL data management, affording developers with the industry’s leading software infrastructure for scalable and secure semantic applications. Commercial applications are now using this technology to solve complex problems in defense and national intelligence, life sciences, and geospatial applications.

## MANAGING SEMANTIC DATA MODELS IN ORACLE DATABASE 11g

Oracle Database 11g incorporates native RDF/RDFS/OWL support, enabling application developers to benefit from a scalable, secure, integrated, efficient platform for semantic data management. This semantic database support is part of Oracle Spatial 11g, an option to Oracle Database. Application developers can add meaning to data and metadata by defining a set of terms and the relationships between them. These sets of terms (“ontologies”) enable enhanced query, analysis and actions based on

---

<sup>1</sup> Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

semantic content, rather than simply data values. Ontologies are increasingly used to build applications that utilize domain-specific knowledge. Ontological data sets, often containing 100s of millions of data items and relationships, can be stored in groups of three, or "triples" using the new RDF data model. Oracle Database 11g enables such repositories to scale into the billions of triples, thereby meeting the needs of the most demanding applications.

Some organizations are using semantic approaches to create an information model (the ontology) based on data schema taken from a particular enterprise organization or industry. Individual application database schema are mapped to a standard information model in order to make the meaning of the concepts in different, application-specific data schema explicit and relate them to each other. The resulting information architecture provides a unified view of the data sources in the organization. As shown in Figure 1, application users can begin to query these enterprise semantic (metadata) models, which comprise of RDF data or ontologies. Standard ontologies reconcile queries needing access to heterogeneous data sources and application-specific schema. This results in solutions that have the power to address unique problems facing enterprise and Web based systems:

- data integration across a heterogeneous, expanding set of corporate/public data sources,
- tracking provenance information, and
- modeling probabilistic data and schema.

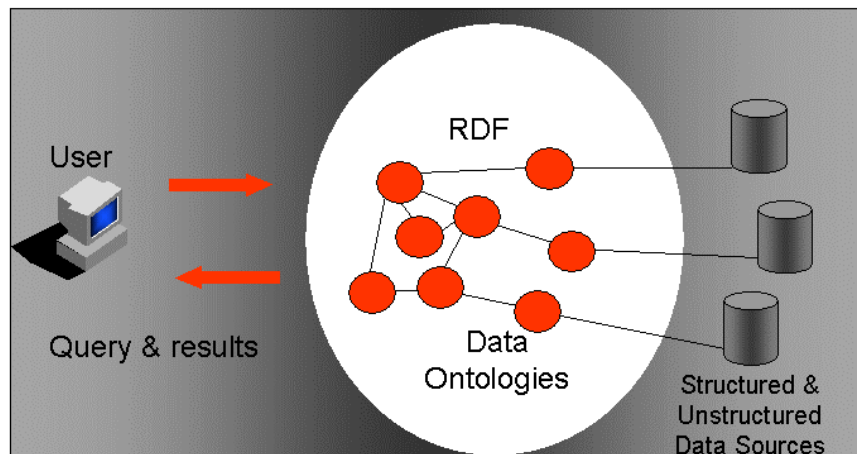


Figure 1. Enterprise Integration Workflow

Managing semantic data models within Oracle Database 11g introduces significant benefits over file-based or specialty database approaches.

- **Low Cost of Ownership:** Semantic applications can be combined with other applications and deployed on a corporate level with data stored centrally, lowering ownership costs. Beyond the advantage of central data storage and query, service oriented architectures (SOA) eliminate the need to install and maintain client-side software on the desktop and store and manage data separately, outside of the corporate database.
- **Low Risk:** RDF and OWL models can now be integrated directly into the corporate DBMS, along with existing organizational data, XML and spatial information, and text documents. This results in integrated, scalable, secure, high-performance applications. Customers can choose to deploy on any server platform (UNIX, Linux, or Windows) using existing IT resources to manage these applications.
- **High Value:** Using the Internet, much larger numbers of users can access the application at virtually no additional costs to the organization. This means that all the users that need to access mission critical information can do so 24x365.
- **Performance and Security:** For mission-critical semantic data models, Oracle provides the security, scalability, and performance of the industry's leading database, to manage multi-terabyte RDF datasets and serve communities ranging from tens to tens of thousands of users.
- **Open Architecture:** The leading semantic software tool vendors have announced support for the Oracle Database 11g RDF/OWL data model. In addition, plug-in support is now available from the leading open source tools.

## INTEGRATING HETEROGENEOUS DATA SOURCES

With an unrelenting growth of business information, scientific data, government documents, email messages, and web content, there is also a rich opportunity to integrate and derive new meaning, value, and intelligence from enterprise repositories of business information. Businesses, scientists and government analysts are beginning to build systems that attempt to integrate access to heterogeneous sources of structured and unstructured data. Few of these systems were originally structured to facilitate such cross-domain integration.

Data integration provides specific benefits and challenges for different domain and application areas. We will look at case studies in the following areas:

- Enterprise data integration
- Domain data aggregation (in the life sciences)
- Content aggregation/knowledge management
- Enterprise search

## ORACLE SEMANTIC WEB TECHNOLOGIES

To address many of the new data management challenges introduced earlier in this paper, Oracle Database 11g delivers the industry’s first open, scalable, secure and reliable data management software for RDF and OWL-based applications. This semantic database support is part of Oracle Spatial 11g, an option to Oracle Database. Major enhancements include support for OWL ontologies and improved performance and scalability. These new enhancements ensure that application developers benefit from the scalability of Oracle Database to deploy scalable semantic-based enterprise applications.

### Oracle Database 11g RDF/OWL Semantic Data Store

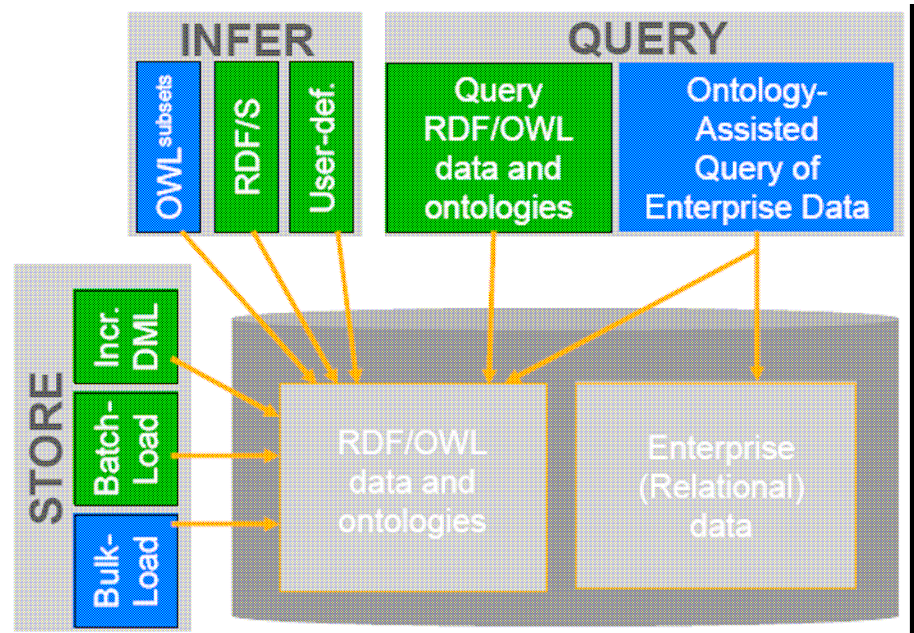


Figure 2. Oracle Database 11g Semantic Data Store

“The future of information navigation will enable enterprises to access both structured and unstructured data seamlessly to find the exact information they need. Oracle’s support of the RDF Data Model in its new 11g software is a major step toward this vision. Combined with Siderean’s Seamark Navigation Server, customers can now deliver a new generation of applications, where users navigate uniformly through all digital information, leveraging the inter-relationships of content to pinpoint results.”

—Bradley Allen,  
CTO and Founder, Siderean Software

The Oracle Database 11g semantic database features enable:

- Storage, Loading, and DML access to RDF/OWL data and ontologies
- Inference using OWL and RDFS semantics and also user-defined rules
- Querying of RDF/OWL data and ontologies using SPARQL-like graph patterns embedded in SQL
- Ontology-assisted querying of enterprise (relational) data

#### Storage, Loading, and DML access to Semantic Data

Oracle Semantic Data Store allows storage, loading and DML access to RDF/OWL models. Each model is an RDF/OWL graph consisting of directed

labeled edges. The edge is labeled by a predicate and connects a subject node to an object node. A subject node must be a URI or a blank node, a predicate must be a URI, and an object node must be a URI, blank node, or literal.

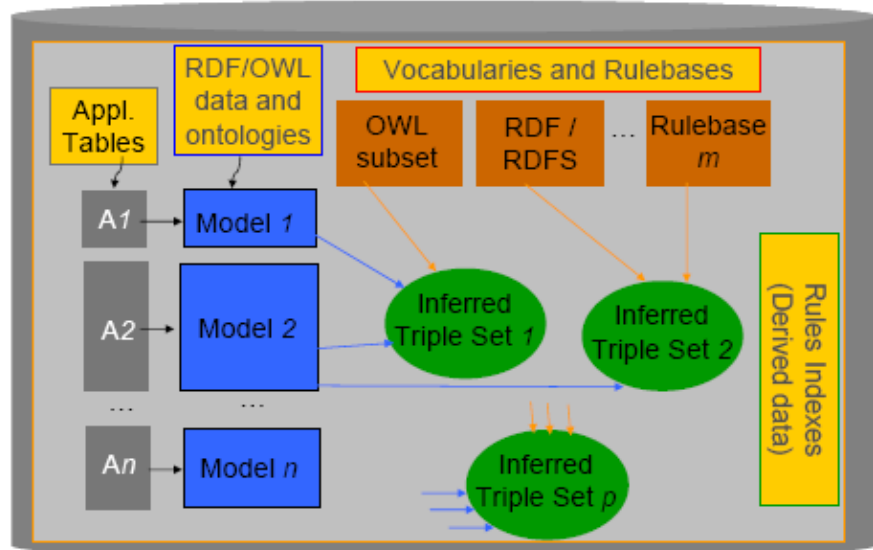


Figure 3. Storage structures in Oracle Database Semantic Data Store

**Advantages of the Semantic Web include the ability to integrate heterogeneous data through common explicit semantics, the expression of rich and well-defined models of systems, the formal annotation of findings and interpretations, the ability to embed models and semantics directly within online publications, the application of logic to infer new insights, the ability to search based on term meaning, and in summary enabling data to be machine-processable.**

This semantic data store manages the complexity arising from repeated usage of typically long URIs and literal values across triples by using a normalized storage architecture. This leads to space-efficient storage, and scalable and performant loading of RDF/OWL data. Equivalence between multiple lexical representations of the same value point (e.g., “0010”<sup>^</sup>xsd:integer and “10”<sup>^</sup>xsd:positiveInteger) are supported as well. Convenient DML access to RDF/OWL models is provided through the familiar concept of database view objects.

**Native Inferencing using OWL, RDFS, and user-defined rules**

The ability to draw inferences from existing data using the precision and rigor of mathematical logic (e.g., Description Logic) is probably the most important property that distinguishes semantic data from others. New Oracle Database 11g enhancements include a native inference engine for efficient and scalable inferencing using major subsets of OWL. This OWL inferencing engine makes the existing native inferencing for RDF, RDFS, and user-defined rules (used for additional specialized inferencing capabilities) more efficient and scalable. Inferencing may also be done using any combinations of these various entailment regimes.

**Querying Semantic Data**

RDF/OWL data can be queried using SQL. The SEM\_MATCH table function, which can be embedded in a SQL query, has the ability to search for an arbitrary pattern against the RDF/OWL models, and optionally, data inferred using RDFS, OWL, and user-defined rules (see below). The SEM\_MATCH function has been

designed to meet most of the requirements identified by W3C in SPARQL for graph querying and RDFS inferencing.

The ability to embed a graph-pattern match query in a SQL query has several advantages: 1) It allows users to specify a query against RDF/OWL graphs as a graph-pattern match query, thereby avoiding the need to manually translate what is naturally a graph query into a relational query; 2) The results returned from one or more graph-pattern match queries embedded in a SQL query can be further processed using the powerful SQL constructs (e.g., aggregate functions) and/or can be joined with other relational tables; 3) Ability to automatically rewrite the graph-pattern match query into a SQL subquery that gets transplanted into the outer SQL query avoids staging of intermediate results and allows leveraging the power of Oracle SQL optimizer, leading to efficient query processing.

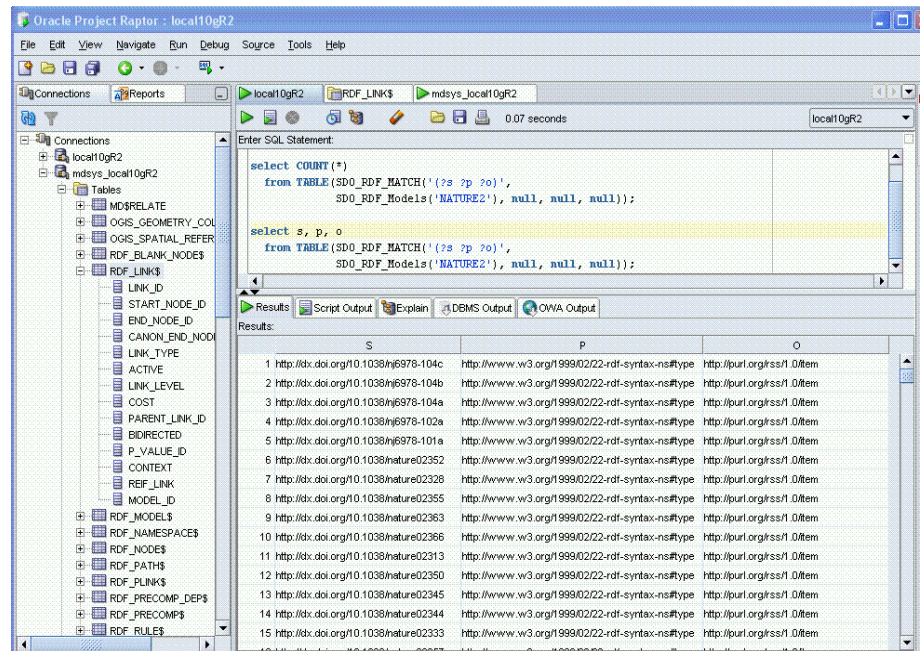


Figure 4. The screenshot above illustrates the structure of an RDF model, SQL query script and query result on an Oracle RDF data model. The ontology used is the Gene Ontology. On top right panel, it shows a simple RDF graph pattern match queries. On bottom right panel, it shows the triples (subject, property, object) of query results.

“The entry of Oracle into the Semantic Web space has already made a big splash, and rightly so. This isn't a big-name player passing off some veneer over an old product as something new; this is a genuine new capability, done well.”

—Dean Allemang,  
Chief Semantic Technology Consultant,  
TopQuadrant

### Ontology-assisted Querying of Enterprise (Relational) Data

Queries can extract more information out of relational data if the relational data is associated with ontologies in the domain of the relational data. For example, if a column in a relational table contains names of diseases, a query asking for match on ‘Immunodeficiency Syndrome’, will be able to retrieve rows containing the value ‘AIDS’ if we interpret the values in that column in the context of the NCI Cancer Ontology [NCI] which states that ‘AIDS’ is a type of ‘Immunodeficiency Syndrome’. The new Oracle Database 11g enhancements include support for new Semantic Operators (similar to those described in [Das et al., VLDB 2004]) that allows ontology-assisted querying of relational data in an efficient manner through the use of Oracle’s Extensibility Framework.

### CONCLUSION

The need to search and and derive greater business knowledge from existing database repositories and applications has become a high priority in many industries. Semantic technologies are being added to enterprise solutions to accommodate new techniques for discovering relationships across different database, business applications and Web services.

Oracle’s semantic web technologies constitute the industry’s first open, scalable, secure and reliable data management platform for RDF and OWL-based applications. New object types have been defined to manage semantic data in Oracle Database 10g . Based on a graph data model, RDF triples are persistent, indexed, and queried, similar to other object-relational data types. Oracle database capabilities to manage semantics expressed in RDF and OWL ensure that application developers benefit from the scalability of the Oracle database to deploy high performance enterprise applications.

### FURTHER INFORMATION

For additional information, white papers, sample code please visit the Semantic Technologies section of the Oracle Technology Network (OTN) at:  
[http://www.oracle.com/technology/tech/semantic\\_technologies](http://www.oracle.com/technology/tech/semantic_technologies)

The following two specific documents have information about some of the use cases mentioned in this paper:

"University of Texas Health Science Center at Houston Deploys Public Health Preparedness Framework with Oracle and TopQuadrant --- Integrated Semantic Web Solution Allows Intuitive Health Data Navigation for Public Health Information Exchange and Improved Decision Making," Oracle Press Release, 19 Feb 2007, [http://www.oracle.com/corporate/press/2007\\_feb/UT%20Houston-TopQ.html?rssid=rss\\_ocom\\_pr](http://www.oracle.com/corporate/press/2007_feb/UT%20Houston-TopQ.html?rssid=rss_ocom_pr)

"Pharma Stuck on Semantic Web", Wendy Wolfson, Bio-IT World, 15 Nov 2006, Report from Oracle OpenWorld 2006, San Francisco, <http://www.bio-itworld.com/issues/2006/nov/oracle-openworld/>

|



Semantic Data Integration for the Enterprise

June 2007

Author: Xavier Lopez, Souripriya Das

Contributing Authors: Melliyal Annamala, Jay Banerjee, Jean Ihm, Jayant Sharma, Jim Steiner

Oracle Corporation  
World Headquarters  
500 Oracle Parkway  
Redwood Shores, CA 94065  
U.S.A.

Worldwide Inquiries:  
Phone: +1.650.506.7000  
Fax: +1.650.506.7200  
[oracle.com](http://oracle.com)

Copyright © 2007, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission. Oracle, JD Edwards, PeopleSoft, and Retek are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.