Chapter 2

THE ENTERPRISE SEMANTIC WEB

Technologies and Applications for the Real World

Susie Stephens[14]

Oracle, 10 Van de Graaff Drive, Burlington, MA 01803, USA – susie.stephens@gmail.com

1. INTRODUCTION

In recent years, much progress has been made in developing ideas and tools to enable the growth of the Semantic Web. The core standard recommendations have reached a level of maturity that allows them to be widely adopted. A range of open-source and commercial software tools is now available. Commercial organizations are also increasingly using Semantic Web technology. However, the Semantic Web still has some distance to go before it reaches a point of widespread adoption.

There are many benefits that would be attained with the extensive adoption of the Semantic Web vision. For example, it would become easier for individuals to find information of interest on the Web and perform computation on that data. Yet, it is expected that many of the initial implementations of Semantic Web technologies will occur within commercial enterprises, and that these will pave the way for the creation of more general applications that operate on the Semantic Web. Consequently, it is very important that state of the art Semantic Web tools are able to meet the scalability, availability, and reliability requirements of commercial organizations.

This chapter reviews some of the business and technology challenges that companies are facing today, and describes how a number of these difficulties could be overcome with the use of Semantic Web tools and technologies. An

[14] Currently at Lilly Corporate Center Indianapolis, Indiana 46285, USA.

overview is then given on the state of the art of these tools and technologies. Use cases are provided that describe real world implementations and areas in which there would be significant benefit in the deployment of Semantic Web technology based solutions. Further, the chapter highlights why it is expected that mainstream adoption of the Semantic Web will initially occur within an enterprise setting.

2. BUSINESS AND TECHNOLOGY DRIVERS OF THE SEMANTIC WEB

Commercial organizations are always under pressure to perform financially. They need to ensure that they meet shareholder expectations while operating under increasingly stringent regulatory controls. Modern companies need to ensure that they can respond rapidly to changing business conditions, as the world continues to move at an ever-quickening pace. Companies are beginning to recognize that information technology can provide a strategic advantage in responding to key business drivers.

Companies today have a growing interest in being able to integrate all data related to the core components that drive their success. Consolidating information about key concepts − such as employees, customers, competitors, operations, and products − enables them to make decisions based on a comprehensive understanding of their business environment. Achieving this vision, however, is no easy task.

Organizations need to integrate not only structured data, but also the increasing volume of unstructured data that they collect and generate. Such unstructured data can take the form of text documents, emails, and presentations. As this unstructured data includes valuable, and sometimes hard-won knowledge, it is especially critical that these resources can be accessed and effectively used across the organization.

Many industries are moving toward more collaborative business models. For example, in the life sciences industry, it is common for companies to outsource various components of their pipeline from drug discovery to clinical evaluation. It is necessary for companies to have flexible data architectures so they can integrate data from collaborators with internally generated data. This becomes a real challenge when companies have many partners and have to manage many interrelated projects.

Increasingly, companies conduct business in many countries around the globe. Ensuring that their operations meet the legal requirements of each country is a complex and dynamic challenge. In addition to archiving information related to those operations, companies may have to record the context in which information was collected or generated to demonstrate that they have met all compliance requirements.

Integrating data across departments also comes with many challenges. It is common for different departments within companies to use their own specialized vocabularies and to use data about the same business elements, but described at different levels of granularity.

While it is theoretically possible for companies to build comprehensive relational models of their data, this approach is somewhat inflexible. For example, it does not enable data to be easily shared with partners during collaborations. The complexity of such a data representation and the difficulties involved in reliably managing any changes to the model would not leave companies well positioned to rapidly adapt to business change. As business conditions continue to evolve quickly, it is imperative that organizations build enough flexibility into their architecture to ensure that it enables, rather than hinders, their ability to respond rapidly to change.

Increasingly, companies are recognizing that their data is their most important asset and that they must be able to control its use. Consequently, they do not want it locked into software that uses proprietary schemas or Application Programming Interfaces (APIs). They want to be able to access the data and use it in ways they did not plan for when their information systems were initially designed. They also need to be able to integrate their own data with data that was created completely independently.

Semantic Web technologies and design principles provide a framework that should make it possible for enterprises to continue to achieve their business goals in the face of change and an increasing dependence on effective use of business data.

3. THE ENTERPRISE SEMANTIC WEB

Resource Description Framework (RDF) is one of the core Semantic Web recommendations from the W3C (Manola, Miller, et al. 2004). RDF represents data using subject-predicate-object triples (also known as ‘statements’). This triple representation connects data in a flexible piece-by-piece and link-by-link fashion that forms a directed labeled graph. This is a very simple, flexible, robust, and expressive model for representing data. As RDF does not have a rigid data structure, it provides a strong framework for seamlessly incorporating new, even unexpected, data with heterogeneous structure. Further, a graph structure simplifies the modeling of data, as it tends to more closely mirror the real world than other data representations that are commonly used.

The components of each RDF statement are identified using Uniform Resource Identifiers (URIs). It is the use of URIs that gives the Semantic Web a fundamental benefit over other technologies. As URIs can be made globally unique, each occurrence of the same identifier means the same thing. The use of URIs enables people and software to know precisely what it is that is being referred to. When a fact is exerted against a URI, there is no possibility of any ambiguity. As there is no ambiguity, it becomes possible to aggregate all data that refers to a given resource. Further, it becomes simpler to integrate data sources that were created independently, as it is not necessary for the data to contain the same values. This makes it simpler for organizations to selectively reuse data.

The approach taken to identify entities using the Semantic Web is very different from the way relational databases identify entities. In a database, all of the identifiers are local. For example, there may be information within a database that refers to the concept of a "purchase order." However, without understanding the schema of the database or seeing the database application, it would not be possible to know if these were purchases that the user of the database wishes to make or whether these are orders from customers. It, therefore, would not be possible to automate the integration of such data with purchase order data from other sources. With the current pace of acquisitions, it would be helpful for commercial enterprises to be able to take advantage of well-defined global identifiers to streamline integration of such business data.

The above description of the benefits of using URIs for integrating information assumes that URIs are being shared and reused. However, this need not be the case, as anyone can generate a URI to describe a concept. If multiple URIs are discovered to represent the same concept, it is possible to use constructs within Web Ontology Language (OWL), such as sameAs or inverseFunctional, to enable such URIs to be used interchangeably. As RDF provides a common framework that is grounded in the Web it allows mappings between data sources to be shared, thus enabling the network effect for data integration.

It is simpler if the same URI is always used to describe the same concept. Within a single company it should be relatively easy to maximize coordination in the assignment of URIs, and this is one of the reasons why the Semantic Web approach is expected to flourish within the enterprise at a relatively early stage.

OWL is another core standard recommendation from W3C for the Semantic Web (McGuinness and van Harmelen 2004). OWL is a more expressive language than RDF. It provides more ways to define terms as classes and instances, and allows users to define relationships between them, which is useful for modeling real-world objects. The data expressed in OWL can be queried, checked for consistency, and have new relationships inferred based on the complex definitions it allows. A further important standard is SPARQL, which is the query language for RDF and OWL. SPARQL contains capabilities for querying by triple patterns, disjunctions, conjunctives, and optional patterns. As a data access language, it is suitable for local and remote use (Prud'hommeaux and Seaborne 2005).

4. SOFTWARE FOR THE SEMANTIC WEB

For companies to fully benefit from the Semantic Web approach, new software tools need to be introduced at various layers of an application stack. Databases, middleware, and applications must be enhanced in order be able to work with RDF, OWL and SPARQL. As application stacks become Semantic Web enabled, they form a flexible information bus to which new components can be added.

As the Semantic Web continues to mature, there is an increasing range of data repositories available that are able to store RDF and OWL. The earliest implementations of RDF repositories, or triple stores as they are also known, were primarily open source. The technology providers adopted varying architectural approaches. For example, 3store (http://threestore.sourceforge.net/) employs an in-memory representation, while Sesame (http://sourceforge.net/projects/sesame/) and Kowari (http://kowari.org) use a disk-based solution. More recently a number of commercial triple stores have become available. These include RDF Gateway from Intellidimension (http://www.intellidimension.com/), Virtuoso from OpenLink (http://virtuoso.openlinksw.com), Profium Metadata Server from Profium (http://www.profium.com/index.php?id=485), and the Oracle Database from Oracle (http://www.oracle.com/technology/tech/semantic_technologies). An interesting trend is that solutions from both Oracle and OpenLink enable RDF/OWL to be stored in a triple store alongside relational and XML data. This type of hybrid solution enables a single query to span multiple different data representations.

Although the ability to store data in RDF and OWL is very important, it is expected that much data will remain in relational and XML data formats. Consequently, tools have been developed that allow data in these formats to be made available as RDF. The most commonly used approach for mapping between relational representations and RDF is D2RQ (http://sites.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/), which treats non-RDF relational databases as virtual RDF graphs. Other toolkits for mapping relational data to RDF include SquirrelRDF (http://jena.sourceforge.net/SquirrelRDF/) and DartGrid (http://ccnt.zju.edu.cn/projects/dartgrid/dgv3.html). Alternative approaches for including relational data within the Semantic Web include FeDeRate (http://www.w3.org/2003/01/21-RDF-RDB-access/) that provides mappings between RDF queries and SQL queries, and SPASQL (http://www.w3.org/2005/05/22-SPARQL-MySQL/XTech.html) which eliminates the query-rewriting phase by modifying mySQL to allow parsing SPARQL directly within the relational database server.

It is also important to be able to map XML documents and XHTML pages into RDF. W3C will soon have a candidate recommendation, Gleaning Resource Descriptions from Dialects of Languages (GRDDL) (http://www.w3.org/2004/01/rdxh/spec), that enables XML documents to designate how they can be translated into RDF, typically by indicating the location of an XML Stylesheet Translations (XSLT). Work is also underway to develop mapping tools from other data representations to RDF (http://esw.w3.org/topic/ConverterToRdf). The ability to map data into RDF from relational, XML, and other data formats, combined with the data integration capabilities of RDF, makes it exceptionally well positioned as a common language for data representation and query.

If a company has been able to consistently assign URIs across all of their data sources, then integrating data is straightforward with RDF. However, if different departments initially assigned different URIs to the same entity, then OWL constructs can be used to relate equivalent URIs. Several vendors provide middleware designed for such ontology-based data integration. These include OntoBroker (http://www.ontoprise.de/content/e1171/e1231/), TopBraid Suite (http://www.topquadrant.com/tq_topbraid_suite.htm), Protégé (http://protege.stanford.edu/), and Cogito Data Integration Broker (http://www.cogitoinc.com/databroker.html).

A number of organizations have developed toolkits for building Semantic Web applications. The most commonly used is the Jena framework for Java-based Semantic Web applications (http://jena.sourceforge.net/). It provides a programmatic environment for reading, writing, and querying RDF and OWL. A rule-based inference engine is provided, as well as standard mechanisms for using other reasoning engines such as Pellet (http://pellet.owldl.com/) or FaCT++ (http://owl.man.ac.uk/factplusplus/). Sesame is another frequently used Java-based Semantic Web application toolkit that provides support for both query and inference (http://www.openrdf.org/about.jsp). A further commonly used application is the Haystack Semantic Web Browser that is based on the Eclipse platform (http://haystack.csail.mit.edu/staging/eclipse-download.html). There are also many programming environments that are available for developing Semantic Web applications in a range of languages (http://esw.w3.org/topic/SemanticWebTools).

5. USE CASES

This section of the chapter describes some use cases of Semantic Web technologies within an enterprise setting. The first three examples are hypothetical, but describe how the Semantic Web could play an important role within a services industry, manufacturing, and government. The next use case describes the deployment of Semantic Web technologies within a life sciences company. The final use case highlights the use of the Semantic Web for content search within a technology company. The use cases section is completed with a brief description of some other application areas where it is expected that there will be significant benefit in the adoption of Semantic Web technologies.

5.1 Enhancing effectiveness of recruitment services

One of the main strengths of the Semantic Web is its ability to integrate disparate data. This use case describes how the technology can be of benefit to recruitment services.

Typically, when a company is looking to hire a person with a particular skill set, they will start a search by engaging the recruiting agencies with which the company works. Each recruiting agency will then perform a search of their proprietary database to identify the best possible candidates for the position. Once the candidates have been identified, the company that is hiring has to compile and manage a list of all appropriate candidates from all of the recruiting agencies.

It would be more efficient if the company that was hiring was able to perform a single query across all of the recruiter databases, thereby retrieving information about all candidates in a consistent format. This would save time in describing the job opening to each agency and minimize potential miscommunications between the company and the recruiting agency. It would also allow the hiring company to directly identify the candidates they believe would be the best match for the role, rather than having to engage in a back-and-forth dialogue about candidates with the recruiting agency. Further, it would become possible to identify internal and external candidates using the same query.

This use case describes how it would be possible to uniformly access all of the recruiter's databases using Semantic Web technologies. Importantly, for the hiring company, this approach would enable querying across different recruiter databases depending on which agencies are currently under contract.

In order to enable querying across multiple recruiter databases, the hiring company would encourage all of its recruitment agencies to make a subset of their data available in RDF, using a common vocabulary and exposing this data by publishing a SPARQL endpoint. It likely that the hiring company would specifically request that it wants to be able to identify job titles, employers, location, qualifications, and willingness to relocate. Figure 2-1 shows a subset of such a schema.

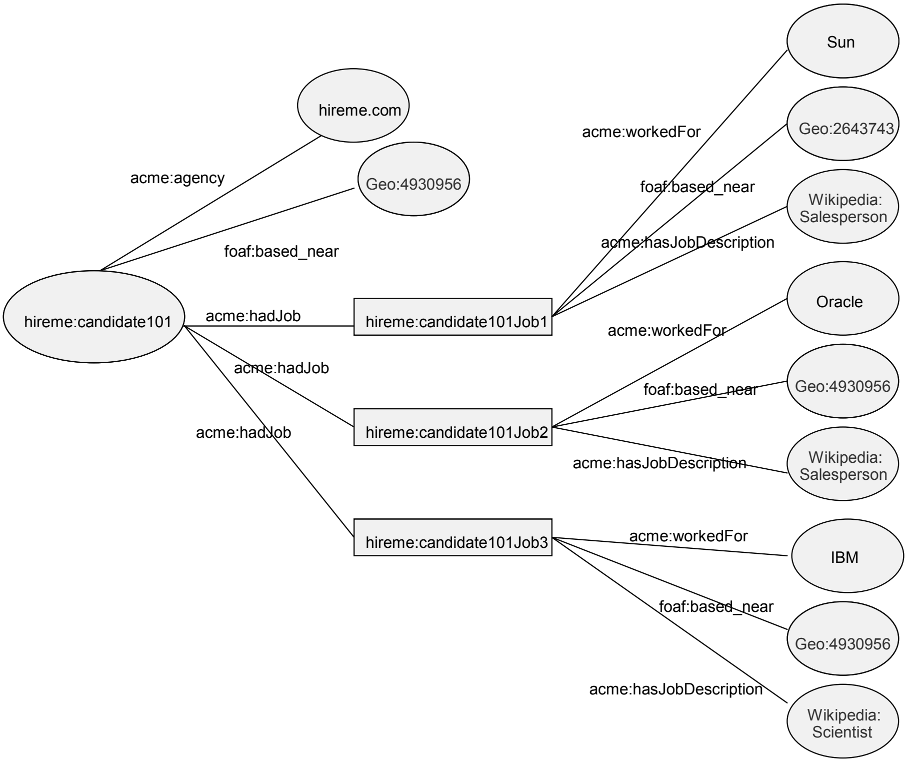


Figure 2-1. The diagram highlights the schema used to represent information about an individual job candidate. The name of the hiring agency that the candidate is working with is shown, along with the location of the agency. Information regarding all three of the candidate's past positions is also shown

Recruitment agencies have two options for making their data available as RDF. They could convert some of their existing relational data into RDF and then store it in a triple store such as Sesame. Alternatively, they could map the relational data to RDF using a technology such as D2RQ. Recruitment agencies would typically use the mapping approach, because this would only require them to map the data once; whereas if they choose to convert their data to RDF, they would need to keep updating the RDF store as their database changed.

The hiring company would likely specify that, in order to query across the data, they would like the agencies to use widely shared URIs. For example, they may require agencies to use the Wikipedia list of occupations (http://en.wikipedia.org/wiki/Occupationlist) to describe the roles candidates have held, and Geonames (http://www.geonames.org/) to describe the geographical location of the positions.

It is likely that a recruiting agency will not store all of its data in the representation required by the hiring company. For example, a U.S.-centric recruiting agency may only store zip codes for location, while the hiring company may require Geoname URIs to be used. In this case, the recruiting agency would need a method to convert a zip code to a Geoname. A Web Service such as GetInfoByZip (http://www.webservicex.net/uszip.asmx) could be used to convert the zip code into an address, and then a Web Service provided by Geonames could be used to convert the addresses into a Geoname URI (http://www.geonames.org/export/geonames-search.html).

The recruiting agency would also need to create some URIs. For example, they would need to create a URI for the agency itself, which would have properties that include telephone number, address, and contact person. The agency could be a foaf:organization (http://xmlns.com/foaf/0.1/#term_Organization), and the contact could be a foaf:Person. They would also need to create a URI for each candidate, which could also be a foaf:Person.

Once all of the agencies have made their data available in RDF via a SPARQL endpoint, it would become straightforward for the hiring company to query for candidates of interest across all agencies. It would also be fairly straightforward for the hiring company to add a new recruitment agency to their list of partners.

Going forward there are many reasons why the hiring company may decide to extend its Semantic Web infrastructure to include OWL. The hiring company may decide there are too many terms that describe similar skill sets, so choose to simplify its queries by using an ontology of occupations. For example, it would be useful if there were a class that described candidates with legal skills, thus avoiding the need to search for candidates with job titles that include lawyer, chief council, solicitor, and legal representative. OWL would also be useful for more advanced querying in which, for example, searches are performed for candidates who have worked at two Fortune 500 companies. OWL would be needed for such a query, as RDF does not provide support for cardinality constraints. Further, ontologies could be used to extend search capability so that they include geographical regions as well as towns.

5.2 Enhancing agile manufacturing

Food manufacturers operate in an environment in which there is severe cost pressure. It is one of the few industries in which purchase decisions are made based on price differences of a few pennies. These cost pressures have resulted in manufacturers shifting from an approach in which equipment is dedicated to the production of a single product, to one in which equipment can be used for multiple products. There has also been a trend toward the reuse of ingredients across different products and greater variety in size of packaging. The shift to equipment that can be used for many purposes puts pressure on manufactures to produce the right product, in the right quantity, and at the right time. This requires a better understanding of consumer demand, the ability to rapidly source ingredients from suppliers, and the need to switch products runs with minimal downtime.

Food manufacturing is also an industry that is facing increasing pressures to be agile. The industry needs to be able to recognize the latest food craze, and to modify the ingredients in products and labeling accordingly. It is also an industry that needs to respond to regulatory agencies such as the U.S. Food and Drug Administration regarding the labeling of products that claim to have health benefits (http://www.cfsan.fda.gov/~dms/lab-ssa.html). Further, the industry must respond to increasing regulations regarding the use of ingredients that are common allergens (http://www.cfsan.fda.gov/~dms/alrgqa.html).

In order to effectively manage these business drivers, companies use a number of enterprise applications. One of the key applications has been manufacturing resource planning software that is designed to handle sales order management, inventory control, accounts receivable and payable, purchasing, payroll, and other front-office functions. Another critical application is the manufacturing execution system. This software analyzes performance and allows dynamic decision making about production floor activities, including process control, work in process, throughput, downtime, changeover time, scrap, and other functions and parameters. Scheduling software is commonly used to identify any potential bottlenecks or critical disruptions in production by taking into account factors such as changeover time, ingredient or package continuity, and parameter changes.

Traditionally within organizations, enterprise resource planning applications have been implemented to support business processes. These applications have evolved over recent years from large monolithic systems, to a more agile Web Services-based approach. However, even with the new approach, it is frequently difficult to integrate such disparate applications, as they often support different underlying data models.

The Semantic Web provides the ability to represent concepts within an ontology. Architecturally, the ontology is used at the interface between the application and the database. With this approach, databases focus on the persistent storage and indexing of data. The application code could be dedicated to implementing the business logic, which it does using terms defined by the ontology, thus creating a unified data model across multiple databases.

The benefits of such an approach include the fact that ontologies name and organize the key concepts within the organization, which have a tendency to change relatively slowly. Software engineering is, therefore, more robust against change when ontologies are employed. Having the concept definitions, constraints, and relationships in the ontology makes it easier for data to be reused by applications for which it was not originally intended. It is also easier for people who understand the business to help work on developing ontologies than it is for them to create database schemas. Further, it is relatively simple to extend the capabilities of an ontology by adding to its constituent relationships, or by linking to a different ontology altogether.

There already exist a number of food ontologies that a company could use (http://www.schemaweb.info/schema/SchemaDetails.aspx?id=61). If a company used an ontology to define all of its ingredients, as well as defining all of its products in terms of those ingredients, it would make it possible to re-classify products as necessary. Recently there have been regulations imposed within the United States that require finer-grained labeling of products that contain common allergens, such as nuts. By adding new definitions of product classes based on ingredients named in these regulations, automated classification would make it easier to identify which products require such labeling. Similarly, in order to support marketing, classes could be defined that capture the balance of ingredients that meet government regulations for when health claims may be made, or which satisfy the latest popular weight loss diets. As regulations or fads change, these definitions could be modified or augmented. Ontology reasoning services could then identify existing products that can be marketed in new ways.

By taking advantage of Semantic Web technologies it would be much simpler for companies to incorporate new data that is deemed relevant. For example, a company may decide that the weather is a strong predictor of the popularity of certain types of food. With a Semantic Web approach it could be easier to incorporate this new data source into decision-making.

5.3 Identification of patterns and insights in data

Semantic Web technologies can be used to find patterns and insights across data that originates from many sources. This approach has much applicability to projects relating to national security as facts and correlations of interest would frequently only be observed when disparate information is brought together.

Within government agencies much information is captured in the form of reports. Text mining approaches could be pursued to extract core information from these documents. With natural language processing the extracted data is commonly in the form of a triple. It is straightforward to take the extracted triples, convert them into RDF, and store them in an RDF repository. The latest release of the Oracle Database could be used as the triple store as it provides support for RDF (Murray 2005).

Once the data has been loaded into the system it forms a directed labeled graph. Users can then query the data using Oracle's SQL extensions for RDF, which makes it simple for a user to query for all information that relates to a particular individual or place. It also becomes possible for a user to follow a chain of links within the data. Importantly for this use case, it further enables a user to find pairs of individuals who know each other, where their combined activities may lead someone to suspect that suspicious activities were being undertaken.

The Oracle Database also provides support for a range of graph analytics, and network constraints (Stephens, Rung, et al. 2004). This functionality is available to RDF data through a Java API. Users can either build in-house applications that work against the API, or use applications that are already integrated such as Cytoscape (Shannon, Markiel et al. 2003) or Tom Sawyer (http://www.tomsawyer.com). Graph visualization and analysis tools are well suited to represent networks of individuals. Such applications would allow users to identify who knows who, recognize sets of individuals that form groups, and find the well-connected individuals who form key hubs both within and between groups.

This use case has focused on analyzing information that was originally contained within government reports. However, it would also be possible to incorporate data that is in other representations into the analysis. One approach would be to convert existing relational or XML data into RDF, and to then load it into the triple store for incorporation into the analysis. The other approach would be to use the power of SQL to perform queries that span relational, XML and RDF data.

In the next release of the Oracle Database support will also be provided for OWL. There will be native, forward-chaining based inference for an expressive subset of OWL Description Logic (basic constructs, property characteristics, class comparisons, individual comparisons, and class expressions). This would allow users to performed more advanced inferencing over data within the Oracle Database. The next release will also include new SQL semantic operators to enhance the query of relational data using ontologies.

5.4 Integration of heterogeneous scientific data

Many pharmaceutical companies are interested in the data integration capabilities promised by the Semantic Web (http://www.w3.org/2004/10/swls-workshop-report.html). This is because drug discovery and development is a very expensive and time-consuming process. To get a drug from bench to market averages 5,000 screened compounds, 15 years and nearly $1 billion (Wolfson 2006). Companies want to minimize costly late-stage attrition by identifying and eliminating drugs that do not have desirable safety profiles or sufficient efficacy as early on as possible (Lesko and Woodcock 2004). It is also important for companies to be aware of competitive offerings or patents that may reduce the market potential of drug candidates. In order to be able to make such decisions, companies need to have an integrated view of their data (Stephens, Morales et al. 2006).

The integration of data, however, has proven to be far from straightforward. Data commonly originates in different departments in which varying terminologies are used. Further, the data itself is very heterogeneous in nature, and consists of data types that include chemical structures, biological sequences, images, biological pathways, and scientific papers. Many companies have attempted to create data warehouses that contain all of this data, but many have found this approach lacking the flexibility required within a scientific discipline. Consequently, companies are exploring alternative approaches to data integration.

A number of pharmaceutical companies are now exploring the use of Semantic Web technologies as a framework for integrating heterogeneous data. Most early projects are focused on integrating data within drug discovery, although ultimately companies are striving to provide a unified view of data from the laboratory bench to clinical observations (Payne, Johnson et al. 2005).

Semantic Web implementations in pharmaceutical companies have involved the development of tools that provide insight into the key entities encountered within drug discovery (Wolfson 2006). These entities typically include genes, proteins, compounds, samples, diseases, projects, and companies. When a user interrogates a particular instance of an entity it becomes possible to view all other information that relates to the object. Frequently the results of such queries display a set of entities as a graph in order to assist the user in visualizing and navigating among the relationships between the entities. This approach enables scientists to discover information as they navigate through available knowledge, rather than necessarily having to have a specific query in mind at the outset. When ontologies are used in combination with such a tool, it enhances the user's ability to retrieve all information of interest even if specialist terminology was originally used to record the data. When querying, it can also help by recognizing identical terms used in different contexts. For example, an ontology would have different concepts for GSK the company versus GSK the protein. Upon querying for GSK, the tool would recognize the ambiguity, and could offer the user a choice of browsing by company or by protein.

There are a number of areas in which the pharmaceutical industry is interested in further exploring the use of the Semantic Web. One additional area of interest is in identifying whether a common biological pathway relates two separately occurring observations, for example, a change in brain pathology and a change in body temperature. This could occur by integrating an ontology that relate clinical conditions to symptoms, and another ontology that relates clinical conditions to possibly modulated pathways. Another area of interest is in the use of ontologies with images. For example, the annotation of medical images by terms in a well-designed anatomy ontology, such as the Foundational Model of Anatomy ontology (http://sig.biostr.washington.edu/projects/fm/), would make possible the retrieval of all relevant images. Moreover, it would be very interesting if image sections could be overlaid with terms from an ontology that effectively defined a coordinate system. This would allow biological mashups in which, for example, a gene expression pattern could be identified as having originated from cells within a certain part of an image.

5.5 Optimizing enterprise search and navigation

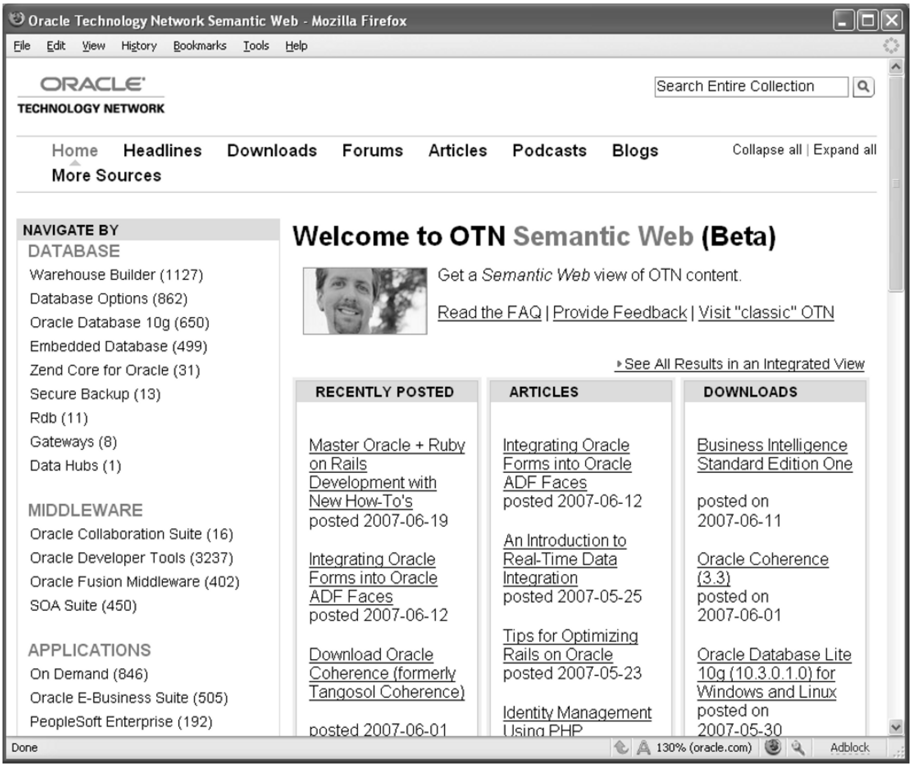
Content search is an area in which several companies are using Semantic Web technologies. Leading commercial vendors in this space include Siderean Software (http://www.siderean.com/) and Endeca (http://endeca.com/). The following use case describes how Semantic Web technologies have been harnessed to improve search within a large corporate enterprise.

The Oracle Technology Network (OTN), part of the Oracle.com Web site, is the main source of technical information for the Oracle developer community. The Web site provides access to product documentation, notifications of product releases, software download, blogs, podcasts, and a discussion forum. The richness, complexity and dynamism of this information have made it challenging for traditional search and navigation techniques to provide effective access to and discovery of information of interest. Oracle has worked with Siderean Software to apply Semantic Web technologies to the Web site to help to address this problem. The Web site is available at: http://otnsemanticweb.oracle.com

The solution is based on the integration of Siderean's Seamark Navigator and Oracle Secure Enterprise Search. This combination of technologies enhances information access by aggregating multiple sources of content and providing it via a single portal. The application enables a common approach for searching and browsing over the rich multi-media data. Such unified access to information is valuable in helping users to identify all information of interest with a single request. Users also have the ability to personalize their environment by selecting the feed items they want to incorporate and the layout of the information. In addition, many of the information types support data visualizations, including tag clouds, contributor clouds, and timelines. These visualizations have been designed to help guide the user by suggesting ways to find the precise information that they are looking for. Figures 2-2 and 2-3 show screen snapshots of the capabilities on the enhanced OTN Web site.

To keep the Web site up to date, every hour the application pulls content of interest to the Oracle developer community from multiple RSS feeds. Although each item in an RSS feed contains some metadata that describes it, often it is not sufficiently detailed or organized. The Seamark Metadata Assembly Platform enhances this metadata by using a combination of pattern-matching techniques to identify the subject matter of an item. Once the subject matter has been determined by Seamark, terms from an Oracle proprietary taxonomy are used to record the concepts and entities to which the item refers. At this point, XSLT is used to convert the XML based RSS syntax into RDF. This approach provides the annotation required to support the rich discovery experience based on dynamic navigation of entities and their relationships. For example, since terms are taken from an ontology, parent terms are more general than their children. Consequently, sets of search results whose metadata include different child terms can be grouped together and labeled using the parent concept. This makes it possible for a user to understand more results at once, and it makes it easier for them to narrow those results to a precise topic of interest.

The search interface also provides the ability to create customized RSS feeds based on user-defined queries. As terms for the concepts and entities in metadata are standardized, feeds constructed using this technology tend to be more targeted, with fewer inappropriately included items. Since the same method of associating metadata is applied to all media types, these feeds can deliver relevant content from diverse media platforms.


Oracle Technology Network Semantic Web - Mozilla Firefox
File Edit View History Bookmarks Tools Help
ORACLE TECHNOLOGY NETWORK
Search Entire Collection
Home Headlines Downloads Forums Articles Podcasts Blogs More Sources
Collapse all | Expand all
Done
130% (oracle.com)
Adblock

NAVIGATE BY

DATABASE
Warehouse Builder (1127)
Database Options (862)
Oracle Database 10g (650)
Embedded Database (499)
Zend Core for Oracle (31)
Secure Backup (13)
Rdb (11)
Gateways (8)
Data Hubs (1)

MIDDLEWARE
Oracle Collaboration Suite (16)
Oracle Developer Tools (3237)
Oracle Fusion Middleware (402)
SOA Suite (450)

APPLICATIONS
On Demand (846)
Oracle E-Business Suite (505)
PeopleSoft Enterprise (192)

Welcome to OTN Semantic Web (Beta)

Get a *Semantic Web* view of OTN content.

Read the FAQ | Provide Feedback | Visit "classic" OTN

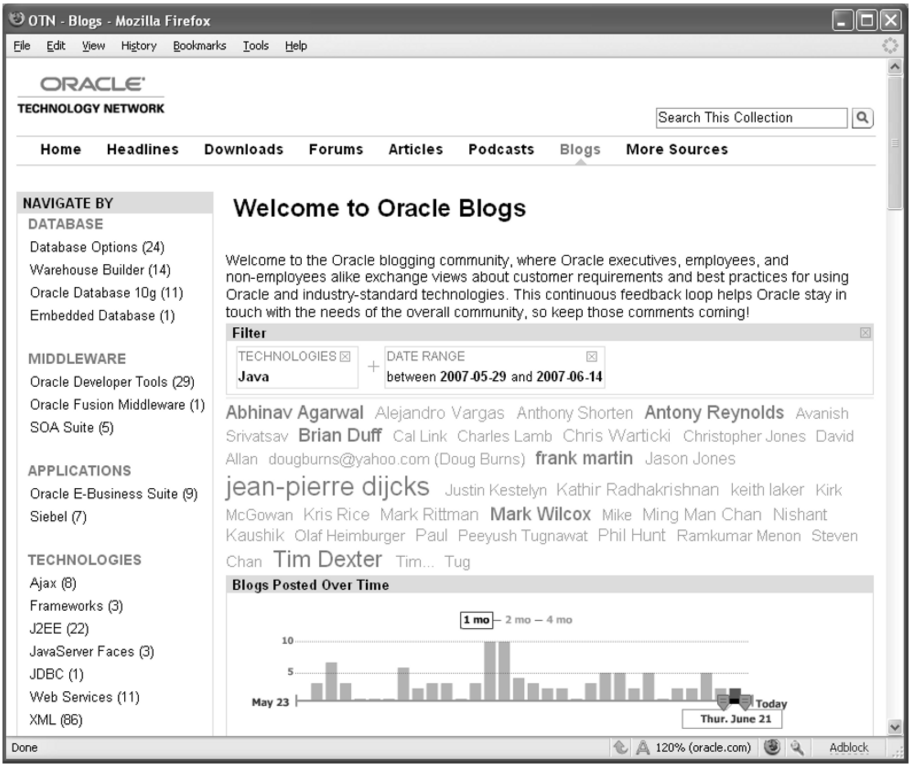
See All Results in an Integrated View

RECENTLY POSTED	ARTICLES	DOWNLOADS
Master Oracle + Ruby on Rails Development with New How-To's posted 2007-06-19	Integrating Oracle Forms into Oracle ADF Faces posted 2007-06-12	Business Intelligence Standard Edition One posted on 2007-06-11
Integrating Oracle Forms into Oracle ADF Faces posted 2007-06-12	An Introduction to Real-Time Data Integration posted 2007-05-25	Oracle Coherence (3.3) posted on 2007-06-01
Download Oracle Coherence (formerly Tangosol Coherence) posted 2007-06-01	Tips for Optimizing Rails on Oracle posted 2007-05-23	Oracle Database Lite 10g (10.3.0.1.0) for Windows and Linux posted on 2007-05-30
	Identity Management Using PHP	

Figure 2-2. Oracle OTN Web site

The Figure 2-2 provides a screen snapshot of the Oracle OTN Web site. Dashboard portlets have been selected by a user to enable simultaneous navigation over multiple repositories. The panel on the left contains facets that are available for navigation, and highlights the number of items available for each link.

Integration and deployment of the solution was very fast and straightforward as both Seamark and Secure Enterprise Search are available as Web Services. The system was designed such that at query time Seamark calls the Secure Enterprise Search API, and then displays the results in the Seamark Relational Navigator framework. The total integration and deployment time for the OTN solution was about eight weeks. Expected future enhancements include support for additional data types, more personalization options, and more innovative visualizations.



NAVIGATE BY

DATABASE
- Database Options (24)
- Warehouse Builder (14)
- Oracle Database 10g (11)
- Embedded Database (1)

MIDDLEWARE
- Oracle Developer Tools (29)
- Oracle Fusion Middleware (1)
- SOA Suite (5)

APPLICATIONS
- Oracle E-Business Suite (9)
- Siebel (7)

TECHNOLOGIES
- Ajax (8)
- Frameworks (3)
- J2EE (22)
- JavaServer Faces (3)
- JDBC (1)
- Web Services (11)
- XML (86)

Welcome to Oracle Blogs

Welcome to the Oracle blogging community, where Oracle executives, employees, and non-employees alike exchange views about customer requirements and best practices for using Oracle and industry-standard technologies. This continuous feedback loop helps Oracle stay in touch with the needs of the overall community, so keep those comments coming!

Filter

TECHNOLOGIES: Java + DATE RANGE: between 2007-05-29 and 2007-06-14

Abhinav Agarwal Alejandro Vargas Anthony Shorten Antony Reynolds Avanish Srivatsav Brian Duff Cal Link Charles Lamb Chris Warticki Christopher Jones David Allan dougburns@yahoo.com (Doug Burns) frank martin Jason Jones jean-pierre dijcks Justin Kestelyn Kathir Radhakrishnan keith laker Kirk McGowan Kris Rice Mark Rittman Mark Wilcox Mike Ming Man Chan Nishant Kaushik Olaf Heimburger Paul Peeyush Tugnawat Phil Hunt Ramkumar Menon Steven Chan Tim Dexter Tim... Tug

Figure 2-3. Oracle OTN Web site – Blogs

In Figure 2-3 shows the Oracle OTN Web site. It now supports a range of visualization capabilities alongside the Semantic Web search and navigation functionality. The screen snapshot shows the results of a search for blogs that mention Java, and that were written between May 29, 2007 and June 14, 2007. The tag cloud shows that Jean-Pierre Dijcks and Tim Dexter have written many of the blogs that meet these requirements. The bar chart depicts the number of blogs that were written on each day over the selected time period. A list of the search results is shown towards the bottom of the page. It is now possible to view blogs according to multiple dimensions that include author, time, and topic.

This search solution was implemented by an enterprise wanting to enhance search and navigation for their customers. In the process of reaching this goal, Semantic Web technologies were used. As a byproduct, the data is now available in RDF, and can be queried through a SPARQL interface. This is, therefore, a good example of how Semantic Web technologies employed to achieve enterprise goals can pave the way for creation of more general applications on the Semantic Web.

An interesting extension to this capability would be to broaden the search and navigation to include business applications. To achieve this objective, it would be necessary to extend the ontology to include all of the entities represented within the business applications. For example, transactions within an inventory tracking system could be provided as RSS feeds with metadata recording the items being added to or removed from the inventory. If the inventory feeds were added to the ones monitored by the portal, it would enable the existing search and navigation facilities to explore inventory as well. This approach would become increasingly interesting as data is incorporated from additional business applications.

5.6 Additional applications of the Semantic Web

The previous section highlight in detail some use cases of the Semantic Web. There are, however, several other application areas in which Semantic Web technologies are expected to provide significant benefits. A brief overview is provided of three such areas:

- **Compliance and Regulation**. There is a growing and increasingly complex set of regulatory requirements to which companies must adhere, including Sarbanes-Oxley, Health Insurance Portability and Accountability Act (HIPAA), and Basel II. The regulatory organizations typically define their policies independently, yet companies must apply all of them to their data. If Semantic Web technologies were used to model the policies related to the regulations, it would make it simpler for enterprises and legislators to merge policies and, thereby, keep abreast of the ramifications of complex interacting policies. In addition, many regulations require organizations to be able to trace and verify data movements and relationships. These requirements are also well suited to the capabilities of the Semantic Web.

- **Event Driven Architecture.** Semantic Web technologies could also aid our ability to more rapidly perceive, process, analyze and act in response to changes in data. These capabilities would be especially valuable when responding rapidly to an emergency, when unexpected relationships between data may need to be explored. For example, if there was a flood, it may be necessary to examine data according to terrain elevation. Another likely use of event driven architectures would be with sensor information. Monitoring signals from Radio Frequency Identification (RFID) tags that include URIs could greatly enhance the effectiveness of condition monitoring, maintenance planning, and the documentation of the technical status of systems and components on offshore platforms or ships, based on using networked devices for remote access during maintenance planning or for *in situ* access during inspection.
- **Service Oriented Architecture Metadata.** The Semantic Web will likely play an important role within Service Oriented Architectures (SOA). The ability of the Semantic Web to assign metadata will help with true dynamic service discovery, invocation, and composition. The Semantic Web could, therefore, improve the inherent flexibility of a SOA infrastructure.

6. CONCLUSIONS

This chapter reviews some of the business and technology challenges that companies are facing today, and describes how a number of these difficulties could be overcome with the use of Semantic Web technologies. However, in order for Semantic Web technologies to be adopted within an enterprise setting, there must be tools that are available that support the scalability, availability, and reliability requirements of such companies. To assess this, the chapter summarizes the state of the art in software tools and technologies, including both open-source and commercial products. A number of use cases are provided that describe real-world implementations and examples of situations in which there would be significant benefit to implementing Semantic Web technology-based solutions. Importantly, the chapter describes why it is likely that mainstream deployment of the Semantic Web will occur first within an enterprise setting and how that can then pave the way for the creation of more general applications of the Semantic Web.

7. QUESTIONS FOR DISCUSSION

Beginner:
1. How could a company generate URIs from existing unique identifiers?
2. What would you use as the namespace for your company?

Intermediate:
1. How could a food company make use of recipe collections on the Web to propose new products?
2. What are some of the limitations of Wikipedia's list of occupations? How would you address some of these limitations?

Advanced:
1. What other business applications do you think would benefit most from the adoption of Semantic Web technologies?
2. What biological or business mashups would you envision to be of value?

Practical exercises:
1. How would you write a SPARQL query to identify the recruitment agency that represents hireme:candidate10021?
2. What facets would you select for navigating the Oracle OTN Web site?

8. SUGGESTED ADDITIONAL READING

- Davies, J., Studer, R., and Warren, P. *Semantic Web Technologies: Trends and Research in Ontology-based Systems.* Chichester, UK; John Wiley & Sons, 2006. A thorough overview of the Semantic Web and interesting use cases.
- Antoniou, G. and van Harmelen, F. *A Semantic Web Primer.* Cambridge, MA; MIT Press, 2004. This book provides a good introduction to the Semantic Web.

9. ACKNOWLEDGEMENTS

I would like to acknowledge Alan Ruttenberg for his contribution to the development of the use cases, and for proof reading the manuscript. I would like to express thanks to Ivan Herman for his feedback. I would also like to thank Brian Anderson for his input to the enterprise search use case.

10. REFERENCES

Lesko, L.J., J. Woodcock (2004). Translation of Pharmacogenomics and Pharmacogenetics: a regulatory perspective. Nature Reviews Drug Discovery. **3**: 763-769.

Manola, F., E. Miller, et al. (2004) RDF Primer. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-primer/

McGuinness, D. L. and F. van Harmelen (2004) OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/

Murray, C. (2005) Oracle Spatial. Resource Description Framework (RDF) 10g Release 2 (10.2). http://download-west.oracle.com/otndocs/tech/semantic_web/pdf/rdfrm.pdf

Payne, P. R. O., S. B. Johnson, et al. (2005) Breaking the Translational Barriers: the value of integrating biomedical informatics and translational research. Journal of Investigative Medicine. 53(4): 192-200.

Prud'hommeaux, E., A. Seaborne (2005) SPARQL Query Language for RDF. W3C Working Draft 21 July 2005. http://www.w3.org/TR/rdf-sparql-query

Shannon, P., A. Markiel, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13: 2498-2504.

Stephens, S., J. Rung, et al. (2004) Graph Data Representation in Oracle Database 10g: Case studies in Life Sciences. IEEE Data Engineering Bulletin 27: 61-67.

Stephens S.M., A. Morales, et al. (2006) Application of Semantic Web Technology to Drug Safety Determination. IEEE Intelligent Systems. 21: 82-86

Wolfson, W. (2006) Oracle OpenWorld 2006: Pharma Stuck on Semantic Web. Bio-IT World. http://www.bio-itworld.com/issues/2006/nov/oracle-openworld/

Springer

http://www.springer.com/978-0-387-48530-0

The Semantic Web
Real-World Applications from Industry
Cardoso, J.; Hepp, M.; Lytras, M.D. (Eds.)
2008, XX, 308 p. 60 illus., Hardcover
ISBN: 978-0-387-48530-0