

Technologyforecast

A quarterly journal
Spring 2009

In this issue

04

Spinning a data Web

20

Making Semantic Web
connections

32

A CIO's strategy for
rethinking "messy BI"

Contents

Features

04 Spinning a data Web

Semantic Web technologies could revolutionize enterprise decision making and information sharing. Here's why.

20 Making Semantic Web connections

Linked Data technology can change the business of enterprise data management.

32 A CIO's strategy for rethinking "messy BI"

Take the initial steps toward your next-generation data architecture.

Interviews

16 Traversing the Giant Global Graph

Tom Scott of BBC Earth describes how everyone benefits from interoperable data.

28 From folksonomies to ontologies

Uche Ogbuji of Zepheira discusses how early adopters are introducing Semantic Web to the enterprise.

40 How the Semantic Web might improve cancer treatment

M. D. Anderson's Lynn Vogel explores new techniques for combining clinical and research data.

46 Semantic technologies at the ecosystem level

Frank Chum of Chevron talks about the need for shared ontologies in the oil and gas industry.

Departments

02 Message from the editor

52 Acknowledgments

56 Subtext

Message from the editor



In the middle of a downturn, the nature of industry change becomes more apparent. Take newspapers, for example. After 146 years, the *Seattle Post-Intelligencer* stopped its presses in March to publish solely on the Web. In Denver, the *Rocky Mountain News* closed entirely, just weeks short of its 150th anniversary. The *Detroit Free Press* has come up with an alternative strategy of offering home delivery only three days a week.

Newspapers certainly haven't been immune to change since the advent of the Web, but the current economic downturn clearly has accelerated the pace of that change. In an April 2009 article, *The Wall Street Journal* even revived the phrase "creative destruction" to describe the latest series of large-city newspaper closures.

The *Journal's* main point? Newspapers must reinvent themselves the way they did when radio and television ushered in the era of real-time reporting. The *Journal's* own strategy, developed and executed by Managing Editor Barney Kilgore beginning in 1941, has been to "explain what the news meant," rather than just report the news.

Large enterprises in many different industries face a similar challenge. The new reality is not only the necessity to do more with less—it's the need to respond to permanent changes in the economy in a more meaningful way because of the downturn. To be able to respond, businesses must come to terms with deeply rooted problems they may have ignored while the economy was growing and the future seemed more predictable.

Among the most critical of these problems are information gaps. The 1,100 CEOs that PricewaterhouseCoopers surveyed in 2008 said the most acute information gaps were in the areas of customer needs and business risk. The natural instinct of most CIOs is to respond to information gaps by loading their data warehouses with more data to generate new reports using expensive, centralized resources. But most enterprises are already flooded with hundreds of reports that are rarely used. Adding more will simply create more information clutter.

What CIOs may be missing is that CEOs want more context about the reports they already have—context that "explains what the data mean," in *The Wall Street Journal* parlance. But the creation of meaning is an active, not a passive, process. It is created by exploring context-specific linkages inherent in enterprise data. Today's business intelligence (BI) and reporting

systems are not designed for this on-the-fly creation of meaning. These systems lack the capability to capture and manage the semantics of the business in a more dynamic, scalable way.

In this issue of the *Technology Forecast*, we examine the emerging technologies and methods being used to directly engage with the meaning and context of a business—its semantics. During the next three to five years, we forecast a transformation of the enterprise data management function driven by explicit engagement with data semantics.

The lead article, “Spinning a data Web,” takes a detailed look at the semantic techniques that enable a Web where documents as well as individual data elements are linked. The result is an ability to filter data sets more effectively and pull more relevant information out of the aggregate.

“Making Semantic Web connections” makes clear that better information context, rather than pure accuracy, empowers better decision making. To provide that context, more “mapmakers” are needed in business units to link information domains. Linked Data technologies will take advantage of the network effect and gain the interest and involvement of people who have never been directly involved before.

The article “A CIO’s strategy for rethinking ‘messy BI’” asserts that BI and related systems aren’t bad, but were designed for only a small part of the information needs businesses have today. To address the remaining needs, CIOs must lead the effort inside business units to build and map information domains.

In addition to these feature articles, this issue of the *Technology Forecast* includes interviews with four executives and technologists from companies at the center of Linked Data research, development, and deployment:

- Tom Scott of BBC Earth talks about his company’s deployment of semantic technologies at the bbc.co.uk/programmes and bbc.co.uk/music sites.
- Uche Ogbuji of Zepheira describes department-level Linked Data initiatives and how grassroots efforts can lead to companywide successes.
- Lynn Vogel of M. D. Anderson discusses the ways and means of semantic technology R&D from a research hospital perspective.
- Frank Chum of Chevron shares insights on how the oil and gas industry is moving Linked Data from pilot to production.

Please visit pwc.com/techforecast to find these articles and other issues of the *Technology Forecast*. If you would like to receive future issues of the *Technology Forecast* as a PDF attachment, you can sign up at pwc.com/techforecast/subscribe.

And as always, we welcome your feedback on this issue of the *Technology Forecast* and your ideas for where we should focus our research and analysis in the future.



Paul Horowitz
Principal
Technology Leader

paul.j.horowitz@us.pwc.com

Spinning a data Web

Semantic Web technologies could revolutionize enterprise decision making and information sharing. Here's why.



Linked Data is all about supply and demand. On the demand side, you gain access to the comprehensive data you need to make decisions. On the supply side, you share more of your internal data with partners, suppliers, and—yes—even the public in ways they can take the best advantage of. The Linked Data approach is about confronting your data silos and turning your information management efforts in a different direction for the sake of scalability. It is a component of the information mediation layer enterprises must create to bridge the gap between strategy and operations. (See the Winter 2008 issue of the *Technology Forecast* for more specifics on the role of and necessity for the information mediation layer in enterprises.)

When you hear about the Semantic Web, don't just think about what's on the other end of a Google search. The issues that the World Wide Web has with data semantics and data silos are simply Web-scale versions of what enterprises have been struggling with for years.

The term "Semantic Web" says more about how the technology works than what it is. The goal is a data Web, a Web where not only documents but also individual data elements are linked. That's why the effort to encourage adoption of Semantic Web techniques is called the Linked Data Initiative.

(See <http://linkeddata.org/> for more information.)

PricewaterhouseCoopers believes a Web of data will develop that fully augments the document Web of today. You'll be able to find and take pieces of data sets from different places, aggregate them without warehousing, and analyze them in a more straightforward,

powerful way than you can now. And don't let the term "Web" fool you into thinking this approach applies only to Web-based information; the underlying technology also applies to internal information and non-Web-based external information. In fact, it can bridge data from anywhere—including your data warehouse and your business partners.

This article provides some background on the technology behind Linked Data, a first semantic step to the data Web. It focuses on how to build on the metadata and ontology technologies that already exist for data analytics purposes. To achieve the data Web, organizations will have to make their own contributions to it—not just by providing access to data, but by exposing and making explicit the context that's now only implicit in column and row headers, in cubes, in inscrutable metadata, or on Web pages. To share this context across domains, organizations will need the context to have a breadth and universality that it doesn't have in its current closed environment.

Optimizing the use of data—not just internally, but throughout the digital ecosystem—is increasingly important. Enterprises continue to be consumed with a subset of what they could be analyzing. To break that mold, they will need to collaborate in a more disciplined way. Their future business agility will depend on their ability to focus on techniques that optimize sharing rather than maintaining silos. That's why a standards-based approach makes sense. In a digital ecosystem, the assets of others can benefit you directly, and vice versa. It's about supply and demand.

The appeal of data federation

If these themes sound familiar, it's because they echo discussions that have gone on for decades. In the early to mid-1980s, claims about relational databases were comparable to claims made now for the Semantic Web—the scale was just smaller. Since then, relational databases—in which tables, or “relations,” can be joined and queried together—have scaled up considerably. They've shouldered the burden of enterprise data analysis, which usually focuses on transactional data. For this reason, their heritage doesn't immediately lend itself to incorporating non-transactional data.

Even so, relational database management systems have remained resilient. Alternatives have been introduced, including object databases (which speed processing by using objects or containers of data or instructions) and Extensible Markup Language (XML)—a file format and data serialization method focused on industry-specific dialects. But relational databases predominate, in part because they're well understood. Investment has continued to pour into their extension and modification; as a result, they have been extensively refined. Because of this refinement, they're fast. Within the boundaries of an organization, relational databases do a lot of heavy lifting, particularly for reports that require speed and precision.

Data warehouses held the promise of a singular, unified approach to all enterprise data, but are becoming justifiable only for high-value, consistent streams or batches of data that require great precision. In a data warehouse environment, the integration task is difficult because each database-to-database link is essentially a custom-built connection in which extract, transform and load (ETL) and associated data profiling processes require much care, expertise, and investment.

Relational data models never were intended for integration at the scale enterprises now need. Relational data management soaks up IT resources that should be dedicated elsewhere. Plus, traditional databases create silos because relational data are specific to the database system implementation. Relational databases can be sources to tap into, but for Web-scale many-to-many sharing with easier connections to more sources, the data model needs an additional dimension—to be designed for reuse by others. Most alternatives to relational databases do not go far enough toward true Web-scale data federation. Although Semantic Web techniques can't compete yet with relational techniques on the basis of pure speed, the payoff from semantic techniques is in Web-scale data federation.

RDF and the Semantic Web

The next step toward true Web-scale data federation must be the Resource Description Framework (RDF), which incorporates lessons from XML and improves on relational data. RDF is one of the primary Semantic Web standards, which also include RDF Schema (RDFS), Web Ontology Language (OWL), and the Semantic Protocol and RDF Query Language (SPARQL). RDF, RDFS/OWL, and SPARQL are among a number of well-thought-out standards that take the page-to-page relationships of the document Web and use them to establish a method for relationships between things. This approach is much more granular and potentially powerful than the document Web.

Fundamental to RDF are global identifiers called Universal Resource Identifiers (URIs). URIs are supersets of Universal Resource Locators (URLs), or ordinary Web addresses. URIs are more specific in a Semantic Web context than URLs, often including a hash that points to

Imagine if every data element you needed had a fixed address you could point to. When you were confident of the source and its relevance, you'd just make the connection to that source and be done with it. That's the inspiration behind URIs.

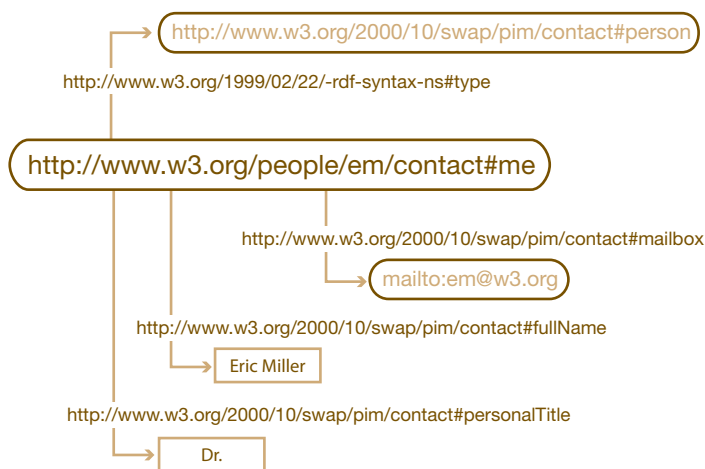


Figure 1: An example of an RDF graph and its computer code

This RDF graph represents machine-readable statements about Eric Miller (formerly head of the Semantic Web Group at the W3C, now CEO of Zepheira) and his 2004 contact information. These statements in English summarize the following: “There is a person identified by <http://www.w3.org/People/EM/contact#me>, whose name is Eric Miller, whose e-mail address is em@w3.org, and whose title is Dr.”

Source: W3C, 2004

a thing—such as an individual musician, a song of hers, or the label she records for—within a page, rather than just the page itself. URIs don’t have to be tied to a Web location; a phone number, for example, does not need its own location. But they do have to be global and persistent to be broadly useful and reliable.

Imagine if every data element you needed had a fixed address you could point to. When you were confident of the source and its relevance, you’d just make the connection to that source and be done with it. That’s the inspiration behind URIs.

Some companies are already using URIs. For example, the British Broadcasting Corporation (BBC) links to URIs at <http://www.bbc.co.uk/music/>, a version of the structured information on Wikipedia, to enrich sites such as its music site (<http://www.bbc.co.uk/music/>).

Tom Scott, digital editor, BBC Earth, says, “The guys at DBpedia can do their thing and worry about how they are going to model their information. As long as I can point to the relevant bits in there and link to their resources, using the URIs, then I can do my thing. And we can all play together nicely, or someone else can access that data.”

Global identifiers are essential to achieve the data Web. Without them, each connection requires a separate agreement. The Linked Data Initiative, which advocates a few simple best practices, considers URIs so fundamental to data federation that they’re behind each of the initiative’s four principles:

1. Use URIs as names for things.
2. Use HTTP (Web locatable) URIs so people can look up those names.
3. When someone looks up a URI, provide useful information.
4. Include links to other URIs, so they can discover more things.¹

RDF takes the data elements identified by URIs and makes statements about the relationship of one element to another. In this vision of the Web, data aren’t in cubes or tables. They’re in graphs consisting of triples—subject-predicate-object combinations. In this universe of nouns and verbs, the verbs articulate the connections, or relationships, between nouns.² Each noun then connects as a node in a networked structure, one that scales easily because of the simplicity and uniformity of its Web-like connections. Figure 1 illustrates an RDF graph, one that depicts the relationships among former World Wide Web Consortium (W3C) Semantic Web Activity Lead Eric Miller, his former title, and his contact information. Understanding RDF graphs as uniform yet

¹ Tim Berners-Lee, “Linked Data,” May 2, 2007, <http://www.w3.org/DesignIssues/LinkedData.html>, accessed March 12, 2009.

² Mike Bergman, “Advantages and Myths of RDF,” *AI³*, April 8, 2009, <http://www.mkbergman.com/?p=483>, accessed April 28, 2009.

flexible structures that lend themselves to Web-scale aggregation is imperative to understanding the concept of Linked Data and the Semantic Web.³

Graphs, which depict the elements and their relationships, are the connecting tissue of the data Web. The larger and more intricate the graph connections among data elements, the more expressive the relationships. This expressiveness of the data and the ability to connect any element to another distinguish the RDF family of standards from relational data. It's a first step in the direction of self-describing data.⁴

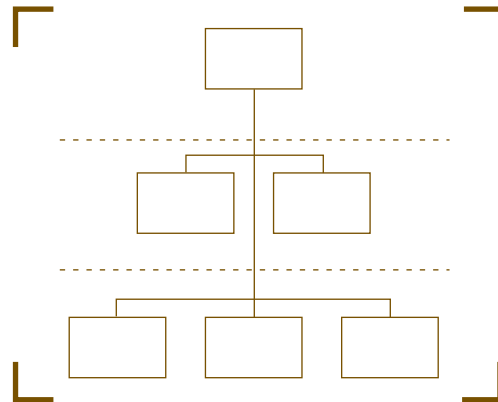
Triples translate neatly into the rows, columns, and cells of tables, including the relationship statements. Unlike relational data, triples allow machines to view more context for making connections between and among elements. A single data element gains even more context each time a new RDF graph is introduced that relates to it. Machines can infer new relationships on the basis of the new connections, and a network effect occurs when the number of connections related to that data element grows.

Through the RDF graph structure, combining data sets is a matter of linking one global identifier (say, for an individual work of music) with another (for an individual recording) via a triple. In this way, the combination exposes a logic that allows machines to bring distributed elements together, a logic both people and machines can add to or modify. The logic and the global identifiers simplify the task.

In the case of a music catalog, the ability to reach out and grab information about a single song becomes easier because of the context resulting from other similar interconnections and how they vary slightly by category. RDF provides more context about how one data element relates to another. The granularity and simplicity empowers users to connect with outside sources

more easily and often. Ontologies build on this capability and make it possible to aggregate larger data sets at the domain level.

Taxonomy



Ontology

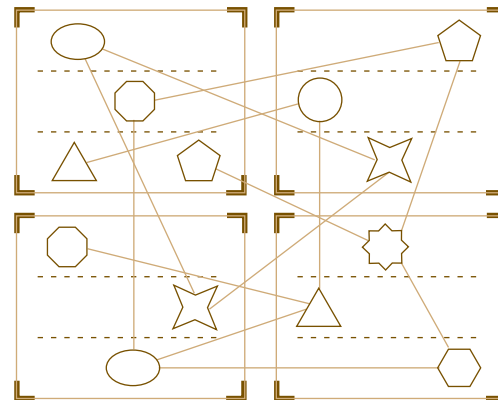


Figure 2: Taxonomies versus ontologies

Ontologies make use of taxonomies, but expand on them, adding a dimensionality taxonomies lack on their own. The expressiveness of RDF Schema and OWL derive from their use of the same flexible graph structure as RDF.

Source: PricewaterhouseCoopers, 2009

3 Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee, "The Semantic Web Revisited," *IEEE Intelligent Systems Journal*, May/June 2006, <http://www2.computer.org/portal/web/csdl/magazines/intelligent#4>.

4 Dean Allemang, "RDF as self-describing data," S is for Semantics Weblog, http://dalleman.typepad.com/my_weblog/2008/08/rdf-as-self-describing-data.html, accessed March 19, 2009.

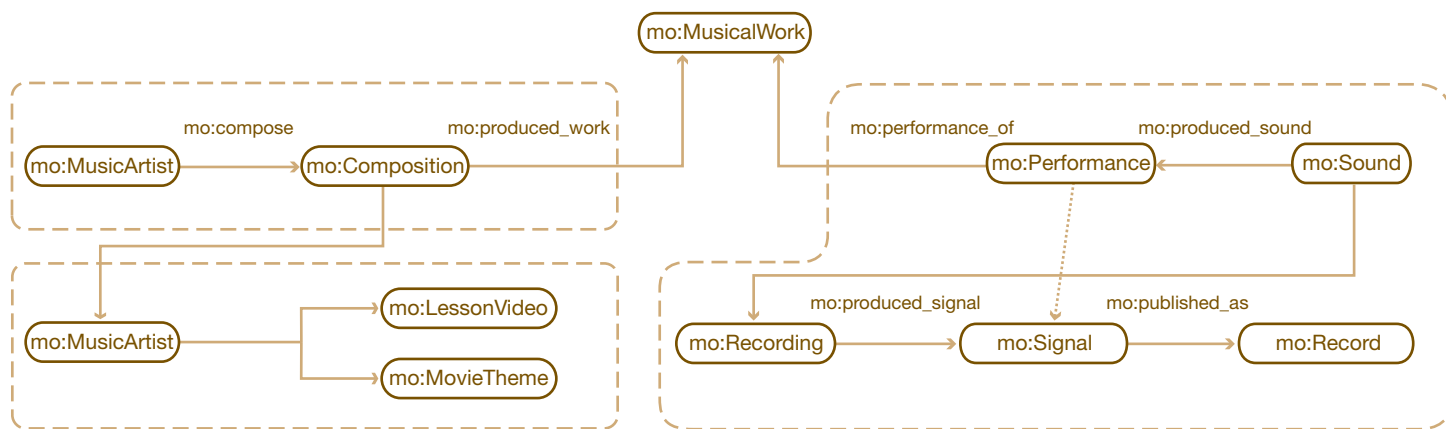


Figure 3: **Example of a simple ontology**

Ontology graphs make the linkages between elements explicit for sharing purposes. In this music industry example, a composer is associated with her performances and recordings.

Source: Music Ontology Specification, 2009

Ontologies for easier sharing and mapping

In philosophy, an ontology is a theory about the nature of existence, an abstraction about the real world that helps people understand and communicate that understanding. In computer science, an ontology describes the characteristics of data elements and the relationships among them within domains. Ontologies describe relationships in an n-dimensional manner, easily allowing information from multiple perspectives, whereas taxonomies show just hierarchical relationships, as illustrated in Figure 2.

A domain includes all the data that share a single context. The human resources department, for example, could have its own domain, and the HR ontology would contain the concepts and domain-specific language associated with HR data. From a Semantic Web perspective, this ontology would be a conceptual framework specific to HR, an overarching structure that allows computers to make more sense of the data elements belonging to the domain.

At every level from RDF to RDF Schema to OWL, Semantic Web standards contribute to the expressiveness of data descriptions, and ontologies are the most expressive. Individual RDF graphs by themselves

contain meaning and logic relevant in an ontological context.⁵ But more elaborate schemas and ontologies can be added, and it's these that provide the ability to federate entire data sets.

In an RDF environment, ontologies provide a capability that's quite useful from a business standpoint, one that extends the utility of taxonomies. Figure 3 depicts a simple example of a music ontology, one that connects an artist, a composition, a performance, a sound, a recording, a signal, and a record. Each relationship among these elements is different. Just as in RDF graphs, explicitly described relationships among elements give ontologies their power.

Let's say your agency represents musicians, and you want to develop your own ontology that contains the same kinds of data elements and relationships as shown in Figure 3. You might create your own ontology to keep better tabs on what's current in the music world, including musicians, venues, media, and so on. You also can link your ontology to someone else's and take advantage of their data in conjunction with yours.

⁵ "OWL 2 Web Ontology Language: RDF-Based Semantics," W3C Working Draft, Dec. 2, 2008, <http://www.w3.org/TR/owl2-rdf-based-semantics/>, accessed March 12, 2009.

Contrast this scenario with how data rationalization occurs in the relational data world. Each time, for each point of data integration, humans must figure out the semantics for the data element and verify through time-consuming activities that a field with a specific label—which appears to be a relevant point of integration—is actually useful, maintained, and defined to mean what the label implies. Although an ontology-based approach requires more front-end effort than a traditional data integration program, ultimately the ontological approach to data classification is more scalable, as Figure 4 shows. It's more scalable precisely because the semantics of any data being integrated is being managed in a collaborative, standard, reusable way.

With the Semantic Web, you don't have to reinvent the wheel with your own ontology, because others, such as musicontology.com and DBpedia, have already created ontologies and made them available on the Web. As long as they're public and useful, you can use those. Where your context differs from theirs, you make yours

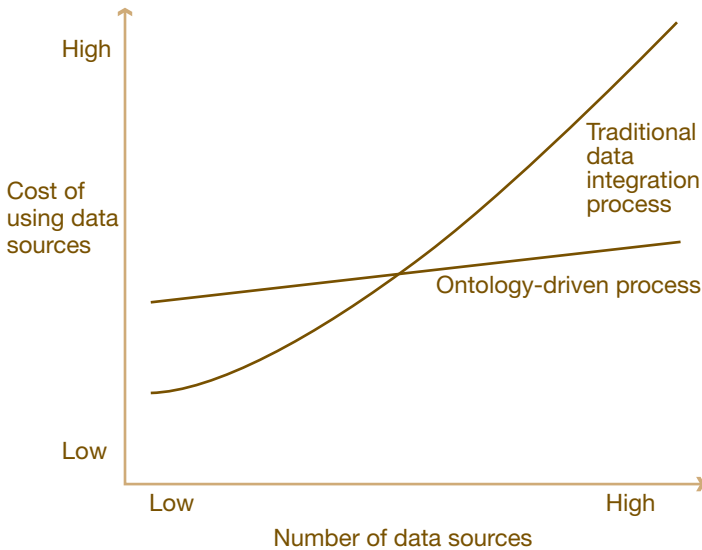


Figure 4: An ontological approach offers scalability

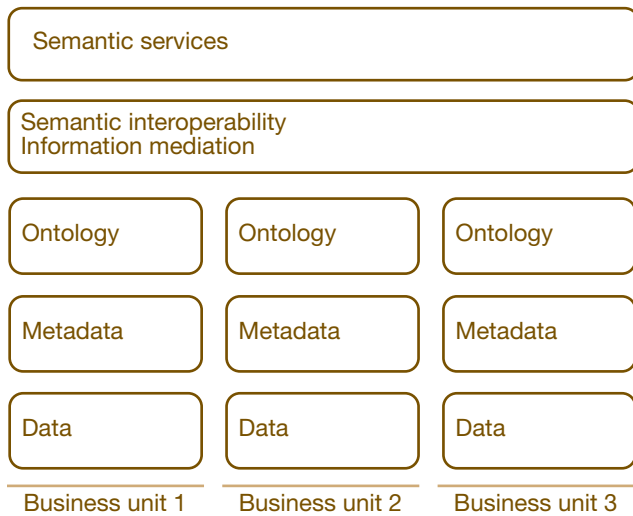
Data federation methods based on Semantic Web standards require a larger initial amount of effort. The benefit of this method becomes clear when scalability becomes a critical need. By lowering the barrier to integrate new sources, ontology-driven processes will eliminate data silos.

Source: PricewaterhouseCoopers, 2009

The explicitness and detail of ontologies make them easier to link than mere taxonomies, which are classification schemes that primarily describe part-whole relationships between terms. Not only do ontologies describe the relationships in RDF graphs, but they also can augment other metadata and less formal kinds of tags, and connect these to the rest.

specific, but where there's commonality, you use what they have created and leave it in place. Ideally, you make public the non-sensitive elements of your business-specific ontology that are consistent with your business model, so others can make use of them. All of these are linked over the Web, so you have both the benefits and the risks of these interdependencies. Once you link, you can browse and query across all the domains you're linked to.

The explicitness and detail of ontologies make them easier to link than mere taxonomies, which are classification schemes that primarily describe part-whole relationships between terms. Not only do ontologies describe the relationships in RDF graphs, but they also can augment other metadata and less formal kinds of tags, and connect these to the rest. In essence, ontologies are the organizing, sense-making complement to graphs and typical metadata, and mapping among ontologies is how domain-level data sets become interconnected over the data Web. Ontologies provide a richer, more unified base of metadata for machine reading, interoperability, and human comprehension of the data. Figure 5 shows how placing ontologies within each business unit allows local specifications



Use semantic interoperability as a platform for searching, processing, and analyzing data across multiple data sources

Semantic interoperability:
Mappings between ontologies that provide a cohesive virtual view over the distributed data stores

Ontologies:
Logical conceptual structures that organize metadata according to semantic principles

Metadata:
Description of what the data are, virtually linked to the physical data the description refers to

Data stores:
Structured, semi-structured, and unstructured distributed physical data stores from different business units or organizations

Figure 5: Ontologies feed the information mediation layer

The ontology layer provides standard logic and organization to simplify mapping between data stores. Interoperability between these stores allows search, query, and analysis of the aggregated data pool.

Source: Peter Rittgen, *Handbook of Ontologies for Business Interaction*, 2009

of meaning, and how a separate semantic interoperability mapping layer links separate business domains together.

As the conceptual characteristics of the data Web become more explicit and machine readable in ontologies, graphs, and organized metadata, they will become the means businesses use to connect to other data sources.

Ontologies are repositories of domain-specific concepts, so business units can create them to describe their piece of the world in language that computers can interpret. Visual tools such as TopQuadrant’s TopBraid Composer make ontology development less intimidating. Ontology development is becoming a more popular business integration technique, particularly as information begins to play a larger role in the overall economy. The healthcare, media, and oil and gas industries, all of which must deal with highly distributed knowledge

sharing, are early adopters. In March 2009, Microsoft announced an ontology add-on to Word 2007. Large-scale adoption of ontologies promises to improve the visibility between business domains that is largely absent in large organizations.⁶

SPARQL: An untethered query language

SPARQL is the W3C’s recommended standard for querying data in RDF graphs and is the most recent major Semantic Web standard. SPARQL is comparable to query languages well known in the relational data world, but it can query whatever data are federated via graphs. SPARQL encounters fewer obstacles because graphs can receive and be converted into a number of different data formats. The graph structure simplifies the relationships among elements and ontologies.

⁶ For information on these and other vendors’ products, please refer to the sidebar, “A sampler of semantic technology vendors” on page 15.

These Semantic Web standards overcome some of the major technological barriers to Web-scale data federation. Before the establishment of the Semantic Web standards, de-siloing data wasn't possible on a large scale without a huge investment of time and resources. XML data are somewhat better, but they don't rise above the level of industry-specific schema or, more often, industry sector schema.

The BBC Earth's Scott contrasts SPARQL with structured query language (SQL) this way: "SQL, in some ways, makes you worry about the table structure. If you are constructing a query, you have to have some knowledge of the database schema or you can't construct a query. With the Semantic Web, you don't have to worry about that. People can just follow their noses from one resource to another, and one of the things they can get back from that are other links that take them to even more resources."

A system based on Linked Data principles provides a layer of abstraction that SPARQL rides on top of. As long as the data messages SPARQL reads are in the form of URIs within RDF graphs, tapping into many different data sources becomes feasible and less painful. Tools such as the Semantic Discovery System incorporate a graphical user interface (GUI) to let you point and click to create joins. Figure 6 contrasts a SPARQL engine's join capabilities in a distributed data environment with the traditional equivalent.

Scott points out that data federation plus SPARQL querying enables more complex queries. Examples such as the BBC site suggest how the Semantic Web might play out over the next decade. Web-scale data sets are like the equivalent of adding dozens of new columns to a spreadsheet and then filtering with the help of the new columns. A year ago, Web pioneer Tim Berners-Lee used the case of Fidelity Investments to describe this capability. He integrated some files in a semi-tabular format with elements from DBpedia and showed Fidelity the results. "OK, I'm looking at all the funds that are based in cities on the East Coast,"

Berners-Lee said to Fidelity. "Suddenly, they see this connection [between broader data sets and more specific information]," he says.⁷

In essence, the broader your data set, the more specific and relevant your query results can be. It provides the ability to filter in a new way and thus extract more relevant insights.

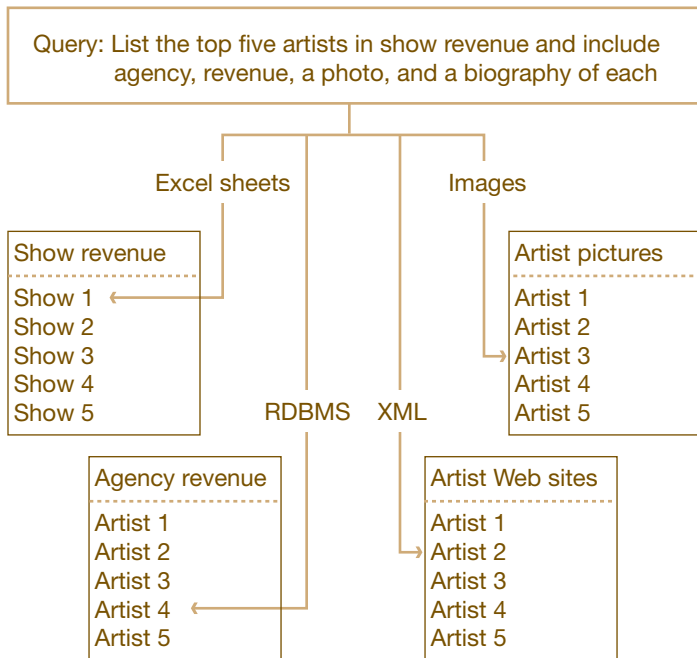
Querying distributed data puts enterprises in a realm that's different from their accustomed realms of data warehouses, traditional business intelligence (BI), and traditional knowledge management. Although it's distributed, this query capability demonstrates benefits with the internal data alone. Ontologies provide a way to stitch together BI and other data sources that reside in silos. Whereas conventional metadata development forces efforts back to the center to reconcile differences, an ontological approach can expose data (internally) in ways that empower users to link to and add data from their own departments.

Taken together, SPARQL querying, RDF graphs, URIs, and OWL distinguish the Semantic Web standards from earlier approaches to information interoperability. SPARQL became a W3C recommendation in 2008, but it made the rest much more valuable because it completes them and paves the way for innovative Web applications.

These Semantic Web standards overcome some of the major technological barriers to Web-scale data

⁷ "Sir Tim Berners-Lee Talks with Talis about the Semantic Web," Transcript of interview by Paul Miller of Talis, http://talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html.

SPARQL engines query across data sources, so joins can be a matter of point and click. Graphical tools can now enable hundreds of joins.



Other joining methods stay in silos, requiring labor-intensive integration.

Source type	Join method
Spreadsheets	-----> write a formula or record a macro
Relational database tables	-----> write SQL joins or use an SQL wizard
Image repositories	-----> query separately using a multimedia database engine, if available
XML content repositories	-----> use a query engine for joins between documents

Figure 6: SPARQL's federation advantage

SPARQL's integration capabilities derive from the use of RDF as a lingua franca for data interoperability. With that base, querying across boundaries becomes easier.

Source: Adapted from Brian Donnelly and Semantic Discovery Systems, 2009

federation. Before the establishment of the Semantic Web standards, de-siloing data wasn't possible on a large scale without a huge investment of time and resources. XML data are somewhat better, but they don't rise above the level of industry-specific schema or, more often, industry sector schema. They're still essentially trapped, and mechanisms such as XLink, a taxonomy linking standard that is designed to overcome at least part of this problem, remain underused.⁸

⁸ As early as 2002, Semantic Web standards were in broader use than Xlink. See Bob Du Charmé, "XLink: Who Cares?" XML.com, <http://www.xml.com/pub/a/2002/03/13/xlink.html>.

The challenge of creating openness and eliminating silos

What seems achievable now is a wide-open, many-to-many Web-scale advantage, not a one-off. That's significant. Semantic Web standards are designed from a 50,000-foot perspective, in Berners-Lee's terms, to enable broadly sharable data. With the help of these standards, you can query across domains. Retrieval is not limited to pages, implementations, networks, servers, or applications. Semantic differences remain an issue, but Semantic Web methods propose a way to bring near-term utility to less-than-perfect data federa-

To take true advantage of the Web, you need to be able to take advantage of its scale, and that's not possible without giving up some control. Limit your controlled environment to what you don't have to scale.

tion efforts. As shared ontologies become more numerous, semantic interoperability issues will become less of an issue.

However, cultural issues are a big reason that data environments often exist in silos. On the one hand, there's a need for compartmentalization. On the other, there's an increasingly pressing need for inter-organizational collaboration. The balance needs to shift toward collaboration, but companies aren't accustomed to sharing data at Web scale and to treating data any differently from the way it's been treated previously. That's why small projects to mine the untapped potential of Web data resources and give people a sense of that potential are important.

Enterprises need control over some data, but not all data. Many enterprises have learned that data warehousing doesn't scale to encompass all corporate data. Plus, some IT departments are consumed with reporting demands. Report generation shouldn't be an IT function—business units should be able to do it themselves. That's a data warehousing problem that needs attention, and part of the answer is not to look to the warehouse for all data needs. Limit the data warehouse to data management problems that align with its attention to detail, its connection to transaction systems, and for problems that need such heavy investments.

To take true advantage of the Web, you need to be able to take advantage of its scale, and that's not possible without giving up some control. David Weinberger of the Berkman Center for Internet & Society at Harvard University calls this the Webby way. "The Web is a permission-free zone," he says. "That's what enables it to scale." Control, he says, doesn't scale.⁹ So limit your controlled environment to what you don't have to scale.

For more information on the topics discussed in this article, contact Steve Cranford at +1 703 610 7585.

⁹ David Weinberger, *FastForward '08* featured speaker, Orlando, Florida, Feb. 8, 2008.

Adding structure to data

Devising more machine-readable data is fundamentally a question of creating and organizing links and metadata. Whether this occurs manually or through automation (realistically, it will take both), enterprises need to create information about data so that relationships between and among data elements are explicit and comprehensible from the machine's point of view.

Organizations have struggled with the metadata issue since before the Web. In recent years, some have come to appreciate the less rigid methods of developing metadata that come from the consumer Web. For example, enterprises have learned that informal user-tagging, called folksonomies and tag clouds, can also inform their metadata development.

Automated tagging by natural language processing engines (Reuters' Open Calais is one example) can provide a boost to these tagging efforts. With the help of these methods and schema development, data descriptions have moved beyond cryptic column and row headers and table joins. But the effort has yet to scale to a point where machines can do significantly more Web preprocessing. The malleability of ontologies helps with the scaling problem.

A sampler of semantic technology vendors

Semantic technology vendors number in the dozens. Most listed here focus on enterprise data integration techniques that rely at least in part on Semantic Web standards.

Cambridge Semantics: The company offers a suite of data federation products named Anzo. At the core of the suite is Anzo Collaboration Server, which normalizes the data it receives to W3C Semantic Web standards. By using extensions in the suite, the server can handle data from Oracle, DB2, and SQL Server, and its own semantic data stores. Anzo for Excel enables knowledge workers to establish a workflow that begins with siloed Excel spreadsheet files and ends with RDF knowledge bases.

Collibra: The company's three levels of Business Semantics Management tools include one each at the services (Information Enabler), governance (Platform), and data integration layers (Studio).

Metatomix: An Oracle partner, Metatomix offers services that companies can use to add to applications; frameworks or tools to build applications; and vertical products for the legal, scientific, and financial services industries. At the core of each product are semantic tools that store, create, edit, and query RDF graphs and OWL ontologies.

Microsoft: Semantic products include the open-source Ontology Add-In for Office Word 2007 to assist in the linking of documents,

and a metadata framework for Interactive Media Manager (IMM), a multimedia content management platform, which includes a metadata framework based on RDF and OWL.

OpenLink Software: The company's Virtuoso Universal Server 6.0 links heterogeneous unstructured and structured data sources at a conceptual level. Virtuoso supports the main W3C standards, including SPARQL.

Oracle: Spatial 11g, an option to Database 11g that accommodates traditional data, XML, and 3D spatial data, offers a wide range of semantic data management features, including native RDF/OWL storage capabilities. With these features, users can store, load, edit rules, and manipulate RDF/OWL data and ontologies.

Semantic Discovery Systems: The Semantic Discovery System provides a graphical user interface that adds point-and-click and drag-and-drop capabilities to SPARQL. The system's virtual RDF store makes it possible to integrate data from a large number of disparate sources.

Structured Dynamics: An OpenLink partner, Structured Dynamics offers Linked Data ontology development and mapping, legacy data conversion to RDF, Semantic Web architectural design, and open source/CMS integration.

Talis Group: The company's Talis Platform is a Semantic Web application development Software-as-a-Service platform. Developers gain access through a Web application programming

interface (API), and the platform acts as a virtual shared database that integrates individual stores, according to the company.

Thomson Reuters: This media corporation offers the Calais Initiative, a Web service that takes unstructured text, analyzes it, and returns it in RDF format.

TopQuadrant: The company's TopBraid Suite is an enterprise platform for developing and deploying semantic applications. The company also offers services and training programs to help organizations harness the knowledge distributed across their systems.

Zepheira: The company offers consulting services in enterprise data architecture for Semantic Web and Web 2.0, specializing in business rules engineering, data exchange, and project assessment.

Selected consumer services

Radar Networks: A social networking and knowledge sharing service, Twine from Radar Networks helps people track and discover information related to their interests. The idea is to enable users to share knowledge in new ways, whether through a distinctive Web interface (available now), an API (available privately), or via RDF and ontologies (in development).

AdaptiveBlue: The company's Glue service adds data linking and social networking through a browser plug-in. Glue recognizes books, music, wines, restaurants, and other topics about which consumers interact daily on the Web.

Traversing the Giant Global Graph

Tom Scott of BBC Earth describes how everyone benefits from interoperable data.

Interview conducted by Alan Morrison, Bo Parker, and Joe Mullich

In his role as digital editor, Tom Scott is responsible for the editorial, design, and technical development of BBC Earth—a project to bring more of the BBC’s natural history content online. In a previous role, he was part of the Future Media and Technology team in the BBC’s Audio and Music department. In this interview, Scott describes how the BBC is using Semantic Web technology and philosophy to improve the relevance of and access to content on the BBC Programmes and Music Web sites in a scalable way.



PwC: Why did you decide to use Semantic Web standards, and how did you get started?

TS: We had a number of people looking at how we could use the Web to support the large number of TV and radio programs that the BBC broadcasts. BBC Programmes had evolved with separate teams building a Web site for each individual program, including the Music Web site. That was two years ago.

If all you were interested in was a particular program, or a particular thing, that was fine, but it’s a very vertical form of navigation. If you went to a Radio One Web site or a particular program site, you had a coherent experience within that site, but you couldn’t traverse the information horizontally. You couldn’t say, “Find me all the programs that feature James May” or “Show me all the programs that have played this artist,” because it just wasn’t possible to link everything up when the focus was on publishing Web pages.

We concluded it’s not really about Web pages. It’s about real-world objects. It’s about things that people care

about, things that people think about. These things that people think about when browsing the BBC sites are not just broadcasts. In some situations they might be more interested in an artist or piece of music. We wanted both. The interest lies in the joins between the different domains of knowledge. That’s where the context that surrounds a song or an artist lives. It’s much more interesting to people than just the specific song played on a specific program.

There was a meeting of minds. We naturally fell into using the Semantic Web technologies as a byproduct of trying to make what we were publishing more coherent. We looked at what technologies were available, and these seemed the best suited to the task.

So that’s when we started with the programs. One of the things Tom Coates [now at Yahoo Brickhouse] figured out was that giving each program for the BBC broadcasts a fixed and permanent URL [Uniform Resource Locator], a subset of Uniform Resource Identifiers [URIs, see pages 6 and 7] that could be pointed to reliably, makes it possible to easily join stuff. So we

“At some point the data management problem reaches a whole different level; it reaches Web scale.”

started working on URLs and modeling that domain of knowledge, and then we thought about how our programs space can relate to other domains.

Some programs played music, and that means someone could view a page and go from there to an artist page that shows all the programs that have played that artist, and that might help someone find these other programs. BBC is more about programs than music. We mainly make programs, and we don't make much music. But we do have a role in introducing people to new music and we do that via our programs. Someone who listens to this particular program might also like listening to this other program because it plays the music that they like.

PwC: How does the system work?

TS: There are logical databases running behind the scenes that hold metadata about our programs. But we also use information about artists and recordings from an outside source called MusicBrainz.com, which maintains repositories of music metadata, and we take a local copy of that. This is joined with data from the Wikipedia and from BBC proprietary content. All this is then rendered data in RDF [Resource Description Framework], JSON [JavaScript Object Notation], XML [Extensible Markup Language], and the pages on the BBC Programmes [<http://www.bbc.co.uk/programmes>] Web site.

The Web site is the API [application programming interface], if you like. You can obtain a document in an HTML [HyperText Markup Language] view format. Or, if you are looking to do something with our data, you can get it in a variety of machine-friendly views, such as JSON or RDF. The machine readability allows you to traverse the graph from, say, a BBC program into its play count, and then from there into the next data cloud. So ultimately, via SPARQL [Semantic Protocol and RDF Query Language], you could run a query that would allow you to say, “Show me all the BBC programs that play music from artists who were born in German cities with a population of more than 100,000 people.”

You probably wouldn't do that, but the point is, the constructed query was initially complex. It's not something that would be trivial and easy to think of. Because there is data that is held within the BBC but linked to data sourced from outside the BBC, you can traverse the graph to get back to that data set.

PwC: What does graph data [data in RDF format] do? What does this type of model do that the older data models have not done?

TS: I think the main difference is where data comes from—where it originates, not where it resides. If you have complete control and complete autonomy over

the data, you can just dump the whole lot into a relational database, and that's fine. As the size of the data management problem gets larger and larger, ordinary forms of data management become more complex and difficult, but you can choose to use them. At some point the data management problem reaches a whole different level; it reaches Web scale. So, for example, the hypothetical query that I came up with includes data that is outside of the BBC's control—data about where an artist was born and the size of the city they were born in. That's not information that we control, but RDF makes it possible to link to data outside the BBC. This creates a new resource and a bridge to many other resources, and someone can run a query across that graph on the Web. Graphs are about context and connections, rather than defining sets, as with relational data.

The real difference is that it is just at a higher level of abstraction. It's Tim Berners-Lee's Giant Global Graph, a term (though not the idea) I'm sure he must have used with his tongue shoved firmly into his cheek.

Originally, the Web freed you from worrying about the technical details of networks and the servers. It just became a matter of pointing to the Web page, the document. This semantic technology frees you from the limitations of a page-oriented architecture and provides an open, flexible, and structured way to access data that might be embedded in or related to a Web page. SQL [structured query language], in some ways, makes you worry about the table structure. If you are constructing a query, you have to have some knowledge of the database schema or you can't construct a query. With the Semantic Web, you don't have to worry about that. People can just follow their noses from one resource to another, and one of the things they can get back from that are other links that take them to even more resources.

PwC: Are there serendipitous connections that come about simply by working at Web scale with this approach?

TS: There's the serendipity, and there's also the fact that you can rely on other people. You don't have to have an über plan. The guys at DBpedia [a version of Wikipedia in RDF form] can do their thing and worry about how they are going to model their information. As long as I can point to the relevant bits in there and link to their resources, using the URIs, then I can do my thing. And we can all play together nicely, or someone else can access that data. Whereas, if we all have to collaborate on trying to work out some über schema to describe all the world's information, well, we are all going to be extinct by the time we manage to do that.

PwC: So, there's a basic commonality that exists between, say, DBpedia and MusicBrainz and the BBC, in the way these sources are constructed?

TS: The relationship between the BBC content, the DBpedia content, and MusicBrainz is no more than URIs. We just have links between these things, and we have an ontology that describes how this stuff maps together.

PwC: Is there a layer of semantics associated with presentation that is connected to the data itself? How did you think about that and manage the presentation rather than the structure of the data?

TS: We wanted the presentation to be good, and from there we fell into the Semantic Web. I would argue that if you structure your information in the same simple fashion as the Linked Data requires, then that creates the user experience. Linked Data is about providing resources for real world things and having documents

“This semantic technology frees you from the limitations of a page-oriented architecture and provides an open, flexible, and structured way to access data that might be embedded in or related to a Web page.”

that make assertions about those things. The first step in building a usable service is to design it around those things that matter to people. Those things that people care about. This, it turns out, is the same first step when following Linked Data principles.

I don't mean that you would expose it raw this way to an audience, but first you need to structure your information the same way and create the same links between your different entities, your different resources. Once you've done that, then you can expose that in HTML.

The alternative is to build individual Web pages where the intelligence about the structure of the data is in the HTML. You could do that to a point, but quite quickly it becomes just too complicated to create sanity across a very large data set.

If you think about music, there are things that make sense in music. They make recordings, and these are released on different media. If you pour your data into that implicit ontology, into that structure, and then expose it as HTML, it just makes sense to people. They can browse our program information and can join it to another one of the domains around other programs.

PwC: Many companies have terabytes or petabytes of data that they don't really know much about. They have to get their arms around it somehow. Is Linked Data an approach they should consider, beyond what we've already talked about?

TS: There is certainly mileage in it, because when you start getting either very large volumes or very heterogeneous data sets, then for all intents and purposes, it is impossible for any one person to try to structure that information. It just becomes too big a problem.

For one, you don't have the domain knowledge to do that job. It's intellectually too difficult. But you can say to each domain expert, model your domain of knowledge—the ontology—and publish the model in the way that both users and machine can interface with it.

Once you do that, then you need a way to manage the shared vocabulary by which you describe things, so

that when I say “chair,” you know what I mean. When you do that, then you have a way in which enterprises can join this information, without any one person being responsible for the entire model.

After this is in place, anyone else can come across that information and follow the graph to extract the data they're interested in. And that seems to me to be a sane, sensible, central way of handling it.

PwC: If we think about broad adoption of Semantic Web standards, it sounds like a lot depends on trillions of URIs being created. Some people say that we'll never get there, that it's too complicated a problem.

TS: The people who say we'll never get there are envisaging a world that is homogeneous. It's a bit like saying car ownership will never get there, because not everyone has a car. The reality is that the future is uneven, and some people will get there sooner than others. Here at the BBC, our work is around programs and music. I'm biased, but I really think that the approach has created a sane, coherent, and stable user experience for people, which is good for our audience. To provide that, we have represented our data in a way that people can now build stuff on top of. Time will tell whether people will do so.

PwC: Do you think an increased focus on data semantics is going to result in a new role within organizations, where job descriptions will include the word “ontology”? Are you being seen as an ontologist within the BBC because you are bringing that specific capability?

TS: It's more about what I used to get the job done as opposed to my job title. My job is product management, and the best way to manage and develop products in an information-rich space is to do so through domain modeling. You'll find that most of the people doing this are more interested in the outcomes than the artifacts that you produce along the way. An ontology is a useful artifact. ■

Making Semantic Web connections

Linked Data technology can change the business of enterprise data management.



Imagine you're a retailer a few years from now, assessing a site for a new store that will sell golf equipment and apparel. Before you decide, you want to develop scenarios about how the store might perform; you also want to examine the potential performance of several new stores when some existing stores are closed at the same time.

You have all the information you need. You know the site, its dimensions, and the planned inventory for the new store. Public data—including demographics, regional economic statistics and forecasts, and locations of competitors—are available. The information exists in different formats at various Internet sites, but that's not a problem because your company has

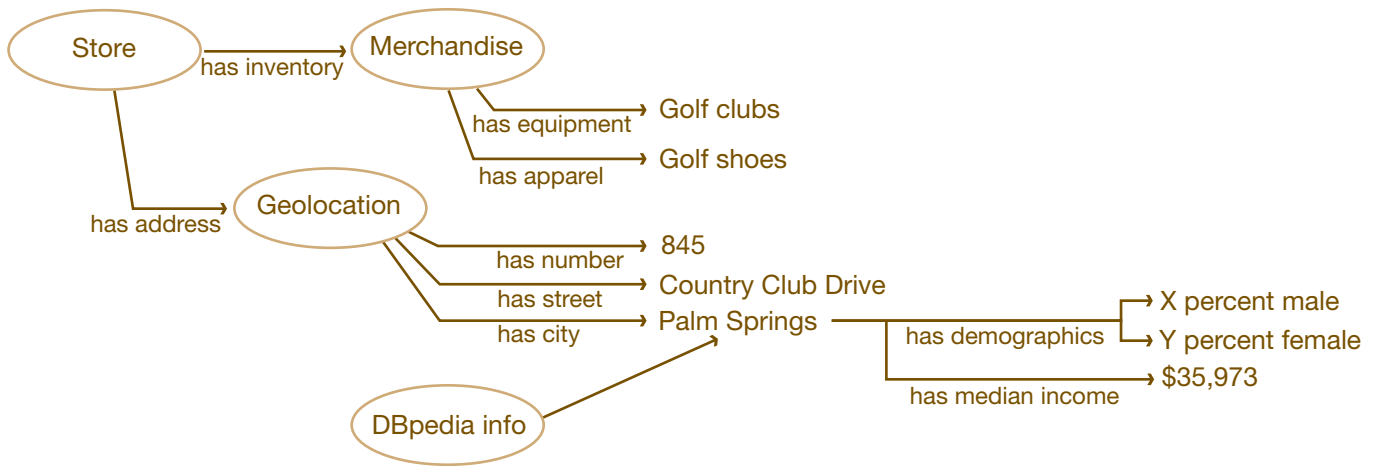


Figure 1: A sample of some retail store information in graph form

Linked Data uses a flexible graph format to connect one data element to another from disparate sources. In this example, external data are connected from DBpedia to the rest.

Source: PricewaterhouseCoopers, 2009

adopted the Linked Data federation method. This set of techniques allows data to remain in native form, but to be exposed and blended at Web scale. This method taps into a larger number of internal and external sources than otherwise would be possible and recasts them in a common format.

Based on emerging Semantic Web standards, Linked Data technologies allow you to refine your golf store scenarios by calibrating for age distribution, per capita income, and other factors by census tract or even block group—all with data extracted from disparate sources. (See Figure 1 on page 21.)

The disparate data feed into a mashup—a Web application with highly configurable data display capabilities—that updates each time you add a new store site or remove an old one. Other data in the mashup are refreshed whenever the original sources are updated. By combining various data, regardless of their format or source, you have a wide range of possibilities for greater insight and context. For example, you can use the same techniques to create information mashups as needed, not just for long-term uses such as the golf

store example. Perhaps a business analyst wants to test changes in regional product purchases against local home sales and employment data to see whether a decrease in sales is due to local economic issues or is a possible harbinger of a broader shift in tastes. You would never create a formal application for this exploration, but with the Linked Data approach, you don't need to.

The mapmaker's data approach: Web-scale federation using Linked Data

The golf store analysis just described wouldn't be easy to do using today's most common information systems. These systems offer no simple ways to efficiently and reliably link data in different formats from various sources inside and outside the enterprise. Three of the four quadrants in Figure 2 are traditionally underutilized. Enterprises have historically approached data integration as an internal engineering challenge, which complicates the task. For years, enterprises have been using the IT equivalent of watchmakers to manage their data, people who have been focused on the closeness of fit of one part with another. They should have been using mapmakers, too—people from the business units to help with what has become a huge information landscape to navigate. This difficult-to-navigate landscape is why executives complain that they don't have enough relevant information to make good decisions. In the 2009 PricewaterhouseCoopers *12th Annual Global CEO Survey*, respondents noted severe information gaps, particularly in the areas of customer needs and business risk. (See Figure 3.)

But these same executives find themselves forced to make decisions rapidly. When executives have enough relevant information, their decisions are more fruitful and enterprises gain a new capacity for agility. In the Fall 2008 and Winter 2009 *Technology Forecast* issues, we described the necessary ingredients for agile responses to a dynamic environment. It comes down to intelligent decisions about standardization versus flexibility. Agility with data is no different. Missing have been effective, universal, and scalable ways to federate data. That's what the Linked Data initiative, the current iteration of Semantic Web efforts, is all about.

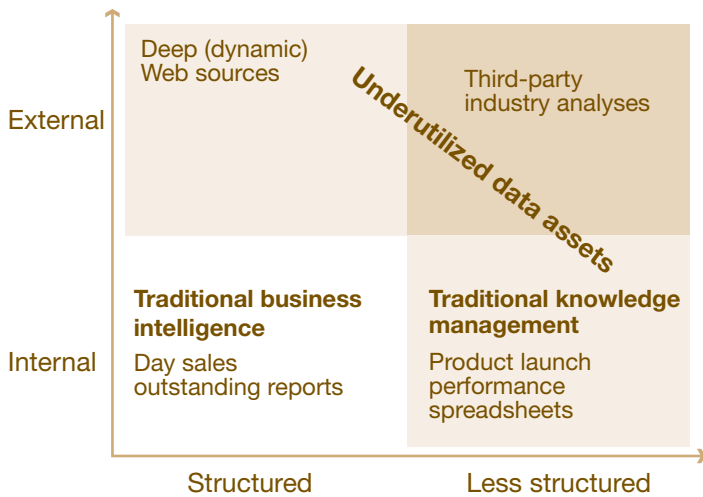
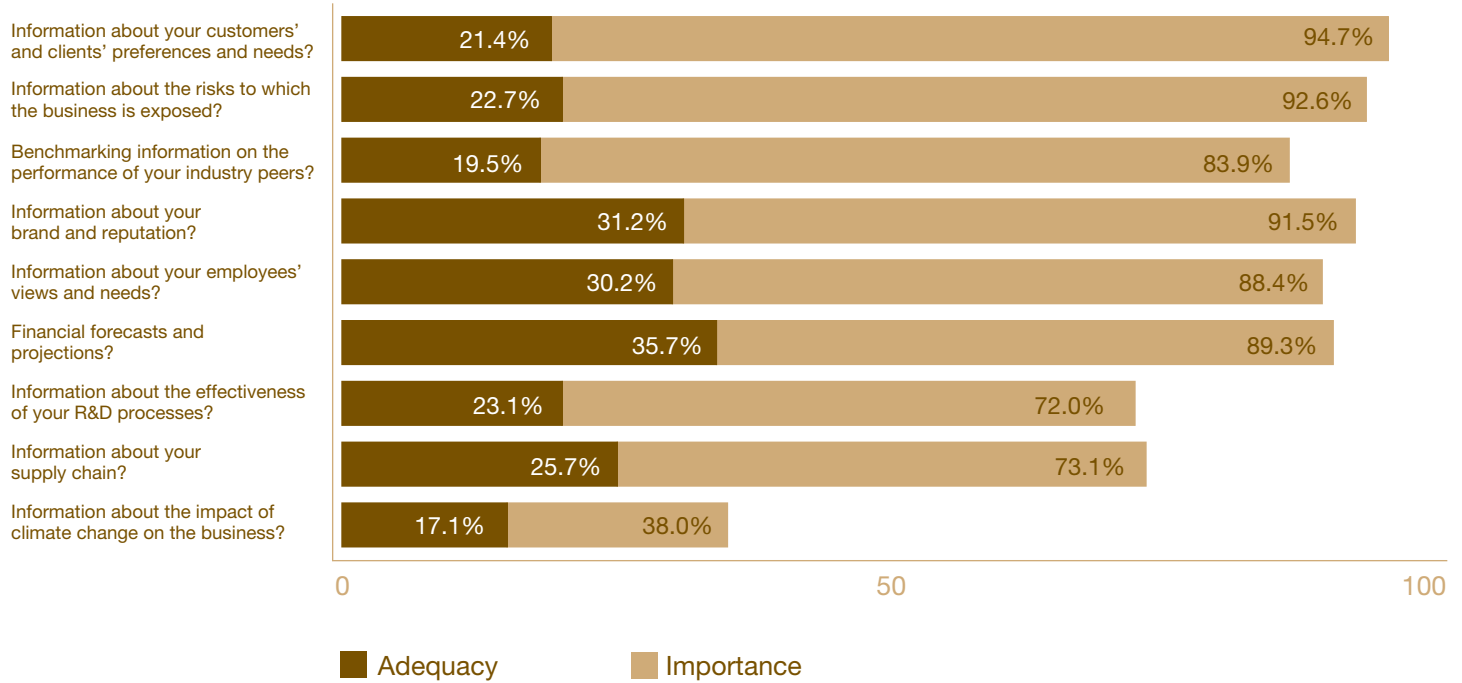


Figure 2: Enterprise information sources

Enterprises generally don't make best use of most of the data they have access to, tending instead to focus on the internal, structured data generated by core transactional systems.

Source: PricewaterhouseCoopers, 2009



Q: How important are the following in terms of the information that you personally use to make decisions about the long-term success and durability of your business?
 Information about the risks to which the business is exposed.
 Base: All respondents 1,124

Q: How adequate is the information that you currently receive?
 Base: All respondents where information is important or critical 427-1,064

Figure 3: CEOs and the information gap

CEOs in the PricewaterhouseCoopers 2009 survey noted very large gaps between the information they had and what they needed on the issues of customer preferences and needs and degree of risk.

Source: PricewaterhouseCoopers 12th Annual Global CEO Survey, 2009

Linked Data technologies will evolve to support more flexible, agile information architectures, a means of unlocking more value from your current information systems while pulling in information that exists outside those systems. These technologies will take advantage of the network effect and gain the interest and involvement of people who have never been directly involved before. (See “Spinning a data Web” on page 4 for descriptions of the primary standards and how they’re being used.)

The era of Linked Data hasn’t fully arrived yet, but business units in some companies and industries are moving toward it now. This article suggests a path for exploring these possibilities. CIOs also can take steps to accelerate the arrival of these methods to the enterprise. (See “A CIO’s strategy for rethinking ‘messy BI’” on page 32.)

Extending the reach of business units

Data aren't created in a vacuum. Data are created or acquired as part of the business processes that define an enterprise. And business processes are driven by the enterprise business model and business strategy, goals, and objectives. These are expressed in natural language, which can be descriptive and persuasive but also can create ambiguities. The nomenclature comprising the natural language used to describe the business, to design and execute business processes, and to define data elements is often left out of enterprise discussions of performance management and performance improvement.

In the Fall 2008 *Technology Forecast*, we described semantics as a stumbling block for communication and collaboration, particularly for large enterprises that must grapple with the different ways departments, subsidiaries, and partners define terms. It's an insidious problem, in part because the use of natural language to describe a business can fool one business unit into thinking their terminology agrees with another business unit's terminology.

In that issue, we described a CIO's effort to harmonize global operations around common processes. (See *Technology Forecast* Fall 2008, pages 26-28.) Different regions were unable to agree on standard processes. They were using the same business terms (examples might be client and price) without realizing their definitions were inconsistent. After they took the time to develop a globally consistent nomenclature that removed the implicit ambiguities in their business terms, they found it much easier to agree on globally harmonized business processes. Although the company did not create a formal ontology per se, it effectively developed a light version of an ontology of its business model. (See "Spinning a data Web" on page 9 for a detailed description of ontologies and their business relevance.) This common language became a core reference resource for defining processes and data semantics.

Standardization is venturing into more uncharted terrain today, especially as most enterprises expand from single global organizations to global business

ecosystems that have hundreds or thousands of trading partners. The sources of data are ever-growing and ever-changing. And user expectations are rising as consumer search engines contribute to user misconceptions about how easily corporate information could be aggregated and found. In short, terminology standardization must now operate at Web scale.

Enhancing business ecosystems with ontologies

In the Fall 2008 *Technology Forecast*, we argued that flexibility should be pursued when the cost of reduced efficiency is offset by value creation. Ontologies are a structured approach to exposing the choices companies must make between operational standards and operational flexibility. They become a platform for creating a shared understanding of the formal business language within the enterprise, where flexibility at a local level within the enterprise is encouraged.

The value of ontologies also extends beyond the enterprise. Few large companies today organize and control the entire vertical stack that defines their end product or service. Most operate within extended ecosystems, working with suppliers and business partners to produce value for customers. Developing a shared ontology within an ecosystem can benefit the ecosystem in two ways. First, it increases efficiencies for participants by reducing ambiguities in terminology and inter-enterprise process management. Second, it allows individual participants to explicitly define those elements of the ecosystem where they make their distinctive contributions and create value.

Frank Chum, an enterprise architect at Chevron, described a December 2008 World Wide Web Consortium (W3C) meeting¹ of oil and gas industry ecosystem participants in which they considered the value of ontologies and Semantic Web technologies. The group identified three potentially useful ontologies in the oil and gas industry:

- Upper ontologies that express concepts not specific to the industry (such as location information)

¹ <http://www.w3.org/2008/12/ogws-report.html>

- Domain ontologies that express concepts specific to the industry or heavily used in it (such as geology, reservoir characteristics, or production volumes)
- Application ontologies that express information used in a particular project or experts' experience in industry activities (such as information about geological interpretations or reservoir simulations)

The focus is on concepts and information shared widely among ecosystem participants. That's what distinguishes the Linked Data approach from what came before it. In essence, these business partners and competitors are using ontologies to enhance agility in both their own internal operations and also in the way their operations integrate with each other.

Eliminating the integration bottleneck

Ontologies sound academic. In truth, companies have been skating around ontologies for years in their metadata development efforts. Until now, enterprises have lacked a reliable set of tools and methods for creating, managing, and sharing metadata semantics in a scalable way. A high degree of business unit involvement and scalability are important so that enterprises can adjust in near real time to changes in data sources. In typical environments of multiple data silos, data warehousing methods don't scale to enhance decision making on an enterprisewide basis. Scaling can be achieved only through a process that distributes metadata management beyond the data warehouse team. Lacking a solution, enterprises have seen the proliferation of silos of disconnected internal data and a tendency to entirely ignore external information.

Traditional data integration methods have fallen short because enterprises have been left to their own devices to develop and maintain all the metadata needed to integrate silos of unconnected data. As a result, most data remain beyond the reach of enterprises, because they run out of integration time and money after accomplishing a fraction of the integration they need.

The public Web also has the problem of unconnected data on a scale that dwarfs the enterprise problem, and

organizations, including the W3C, are working to solve it. Tim Berners-Lee, director of the W3C, is focused on this task. The early experiences of the W3C and the W3C standards can provide guidance for others.

The most basic lesson is that data integration must be rethought as data linking—a decentralized, federated approach that uses ontology-mediated links to leave the data at their sources. The philosophy behind this approach embraces different information contexts, rather than insisting on one version of the truth, to get around the old-style data integration obstacles. To be meaningful, linking data from separately managed data stores requires a comparison of these different contexts with a referential context. In practice, this means comparing the meaning of metadata associated with separate data stores.

On the World Wide Web, enterprises are piloting the use of shared domain ontologies. In essence, each domain creates its own data model—an explicit collection of statements of how data elements relate to each other. These collections are called ontologies. Ontologies describe relationships between concepts. Once a business unit creates that description, it becomes a part of the shared whole. To reuse part or all of it, others can just point to what they need. Each ontology maintains its own context. By connecting metadata to shared domain ontologies, companies are learning to automate the process of comparing and establishing shared meaning.

Clarifying semantics with Linked Data techniques

By using ontologies, you can link to data you never included in the data set before. These data add more context and timely information, improving decision making. The richer data sets allow decision makers to perform more ad hoc analyses as needed, so they aren't restricted to analyzing only what they knew they needed in the past. The Semantic Web takes the page-to-page relationships of the linked document Web and augments them with linked relationships between and among individual data elements.

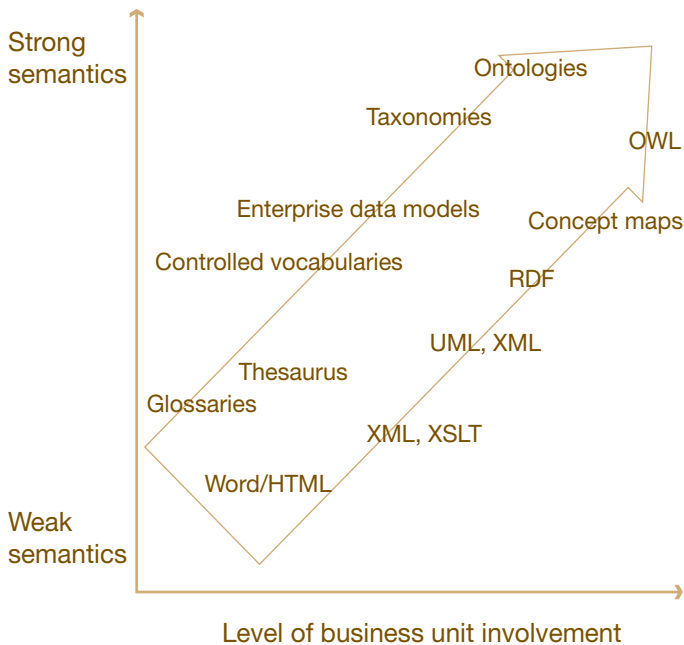


Figure 4: Web techniques and semantic clarity

Ontologies have the highest level of expressiveness, which leads to more powerful data federation capability. But using them effectively requires a strong degree of business unit involvement.

Source: Ian Davis of Talis, 2005

The overall benefit is better contextual information that leads to better understanding. Lynn Vogel, vice president and CIO of the M. D. Anderson Cancer Center, describes the goal of the center’s ontology development project as “providing not simply a bridge between clinical and research data sources, but potentially a home for both of those types of data sources. It has the ability to view data not simply as data elements, but as data elements with a context.”

Vogel offers the following example. Doctors can provide patients with an analysis of how the therapies they intend to provide worked in the last 100 cases. On the one hand, this is a clinical issue involving individual patients and hundreds of attributes. On the other hand, it’s a research issue—the analysis of patient data, which

focuses on a few attributes at a time and many different patients. The data structures that exist, he says, are suited to either one or the other, but not both.

“Our semantic tools are one way to bridge that gap,” Vogel says. “It is possible, but we’re not convinced entirely yet because we’ve been dipping our toe in this water for only the last year and a half. But the ability to provide the framework within which both of these vastly different types of data can be used is going to determine the future of how successful we are in dealing with cancer.”

Vogel describes the data integration problem in much the same way Tim Berners-Lee frames the Semantic Web challenge within the Linked Data initiative. This approach is much more granular than the document Web. The proper level of detail will take time to develop. Significant developer activity at semantic tools vendors is under way—including some that engages a broader audience—but this activity is still small and unrefined. Some applications are already emerging. For example, the BBC has a Web site that uses Semantic Web techniques to blend content about music. Many enterprises, as noted later, have completed or are conducting pilots of Semantic Web technologies and learning valuable lessons.

Distributed data on the Web and the need to aggregate and analyze it on the basis of exposed semantics has led to new data management tools, techniques, and philosophies. Conceptually, they represent the logical next step in the evolution of data management. And because Semantic Web techniques build on what’s come before, ontologies can help enterprises organize and expand the metadata they’ve already developed. (See Figure 4.) In the process, ontologies can become a vehicle for the deeper collaboration that needs to occur between business units and IT departments.

In fact, the success of Linked Data within a business context will depend on the involvement of the business units. The people in the business units are the best people to describe the domain ontology they’re responsible for. New tools are becoming available that will take the mystery out of ontology development for non-technologists.

	Traditional data integration	Linked Data approach
Data structure	Predominantly relational: focus is on sets of similar data	More flexible: focus is on relationships between things regardless of similarity
Data integration method	Extract from original source, transform to local data definitions, load on own servers	Link to source of data using data definitions in shared ontology
Data integration scalability	Each new data source expands costs exponentially	New data sources are accessible at minimal cost, and business domains share the federation cost
Contextual richness	Constrained by costs and central staff workloads	Benefits from the network effect: context gets added with new data and linkages
Information source bias	Internal	Internal and external
Business unit involvement	Report requestors	Managers of their own ontology and external data-linking activities
Standardization method	One standard, no exceptions, loss of valuable information context	Explicitly allows both standard data and contextual information

Table 1: Benefits of linked versus traditional data integration

Source: PricewaterhouseCoopers, 2009

Data's new competitive advantage

Companies have long looked for ways to facilitate the movement of information, to automate processes with minimal exceptions and reworking. The problem is that they haven't been attacking it at the right level. Traditional integration methods manage the data problem one piece at a time. It is expensive, prone to error, and doesn't scale. Metadata management gets companies partway there by exploring the definitions, but it still doesn't reach the level of shared semantics defined in the context of the extended virtual enterprise.

Linked Data offers the most value. It creates a context that allows companies to compare their semantics, to decide where to agree on semantics, and to select where to retain distinctive semantics because it creates competitive advantage. Table 1 summarizes the benefits of the Linked Data federation approach and the data integration techniques that preceded it.

The ramifications are substantial. As Federal Express and UPS learned, providing information to customers can change business models. As these organizations exposed more data, their model broadened beyond next-day delivery to providing alerts for changes in shipments. Companies across industries need to be open to leveraging, combining, and sharing information in ways that not only make their offerings more compelling, but also create more business value for customers.

In the end, data must be viewed as a key contributor to agility and distinctiveness and the means to a sustained, profitable enterprise. Organizing and managing internal data with ontologies opens the door to linking with huge resources of new data in a scalable way. The resulting context adds intelligence to decision making and better business outcomes.

For more information on the topics discussed in this article, contact Steve Cranford at +1 703 610 7585.

From folksonomies to ontologies

Uche Ogbuji of Zepheira discusses how early adopters are introducing Semantic Web to the enterprise.

Interview conducted by Alan Morrison, Bo Parker, Bud Mathaisel, and Joe Mullich

Uche Ogbuji is a partner at Zepheira, LLC, a consultancy specializing in next-generation Web technologies. Ogbuji's experience with enterprise Web data technologies reaches back to the inception of Extensible Markup Language (XML). In this interview, Ogbuji discusses how Zepheira helps companies with semantic interoperability issues, and he provides insight into the data silo problems organizations face.



PwC: What kinds of issues do Zepheira clients ask about?

UO: We're a group of 10 folks who speak a lot at conferences, and I write a lot. So we very often get inquiries from people who say, "Something that you said or a case study that you presented caused a light bulb to go on in my head, in terms of how this can help my company."

These are typically folks at a department level. They're ambitious; they are looking to take their entire company to the next level by proving something in their department, and they recognize that the beauty of something like semantic technology is that it can build from a small kernel and accrete outwards in terms of value.

These department-level people have a practical problem that often has to do with having data in a lot of silos that, if they could integrate better, they could get more value out of. Obviously that's an age-old problem, one that goes back to the primordial days of computing. But I think they see that this is an opportunity to use a very

interesting technology that has some element of being proven on the Web and that allows things to be done on a small scale.

PwC: So sometimes they're trying to organize structured data that is in silos as well as unstructured data?

UO: Right.

PwC: Speaking of structured and unstructured data, could you give us an example of how a department head might find the Semantic Web useful in that context?

UO: If you have a bunch of data in different files, some of it structured and some of it unstructured, you often have different systems developed in different areas at different times. The business rules included in them are different. They don't quite match up. You have a lot of

“There’s a large social element in building shared models, and once you have built those shared models, you have the social benefit of having people enfranchised in it.”

inefficiency, whether from the complexity of the integration process and code or the complexity of the day-to-day activities of a line-of-business person.

What we typically do on an engagement is try to capture what I would call schematic information, which is information about what relates to what. They’re deceptively simple links between entities in one silo and entities in another silo, so we’re not talking about a huge, formal, scientific, top-down modeling exercise. We’re talking about links that are almost at the level of social tagging, almost at the folksonomy level.

We’ve found that when you provide a basis for people to say that this entity, this sort of information in this silo relates to this other sort of information in this silo, then the people who are involved fill in the nooks and crannies. You don’t have to have this huge engineering effort to try to force a shared model between them.

So I think the benefit that department heads get from something like Semantic Web technology is that it’s designed to go from very slim threads and very slim connections, and then have those strengthened over time through human intervention.

There’s a large social element in building shared models, and once you have built those shared models, you have the social benefit of having people enfranchised in it. Some organizations had a situation where trying to do data governance was warfare, because of the competing initiatives. Now you have given people

the capability to do it piecemeal collaboratively, and you have less of the warfare and more of the cooperation aspect, which improves the system that they’re developing.

PwC: Can you give us an example of a company that’s done this sort of collaboration?

UO: One concrete example is the work we did with the global director for content management at Sun Microsystems. Her office is in charge of all the main sun.com Web sites, including www.sun.com, the product sites, solutions, global versions of the sites, and the company’s business-to-business [B2B] catalogs. Her department had data, some of which is Oracle database content—warehouse-type data, a lot of which is XML [Extensible Markup Language]—and some of which is spool files.

Governance was not in place to automate the pipelines between all that mess of silos. And getting out a coherent Web site was a pain. They had some real problems with price policy and traceability for, say, prices that appear on the catalog Web site. It used to be a very manual, intensive process of checking through everything. We worked with this department to put together a platform to create lightweight data models for the different aspects of product information that appeared on these Web sites, as well as to make those models visible to everyone.

“You’re always going to have difficulties and mismatches, and it will turn into a war, because people will realize the political weight of the decisions that are being made. There’s no scope for disagreement in the traditional top-down model. With the bottom-up modeling approach you still have the disagreements, but what you do is you record them.”

Everyone could see the components of a lightweight data model and the business rules in a way that’s as close as possible to stuff that a line-of-business person could understand. That helped them head off major disagreements by dealing with all inconsistencies piecemeal. It’s not perfect, but now they have a quicker time to market for reliable product announcements and reliable information updates, and that was really valuable. And on the personal and social side of things, I’ve personally been very satisfied to watch that the lady who brought us in has been promoted quite a few times since we’ve been working with her. Very often that’s the motivation of these people. They know it can be valuable, and they’re looking to do something special for their company.

PwC: What was the breakthrough that you alluded to earlier when you talked about the new ability to collaborate? While there used to be a data governance war and everybody had their own approach to the problem, what caused this ability to collaborate all of a sudden?

UO: It’s slightly different in each organization, but I think the general message is that it’s not a matter of top down. It’s modeling from the bottom up. The method is that you want to record as much agreement as you can. You also record the disagreements, but you let them go as long as they’re recorded. You don’t try to hammer them down. In traditional modeling, global consistency of the model is paramount. The semantic technology idea turns that completely on its head, and basically the idea is that global consistency would be great. Everyone would love that, but the reality is that there’s not

even global consistency in what people are carrying around in their brains, so there’s no way that that’s going to reflect into the computer.

You’re always going to have difficulties and mismatches, and, again, it will turn into a war, because people will realize the political weight of the decisions that are being made. There’s no scope for disagreement in the traditional top-down model. With the bottom-up modeling approach you still have the disagreements, but what you do is you record them.

PwC: Have you begun to understand the opportunity here for a class of business problems that have been heretofore either not solvable or too expensive to solve with traditional approaches and that define a continuum from purely Semantic Web value possibilities to purely highly structured and controlled vocabularies?

UO: You would not want a semantic technology-driven system whose end point is the XBRL [Extensible Business Reporting Language] filing to the SEC [Securities and Exchange Commission]. That would be an absolute disaster. So there is absolutely a continuum—from departments and use cases where this is appropriate, cases where it’s a hybrid, and cases where you need very, very structured, centralized control. The XBRL example is a great one. XBRL is semantic technology in itself because of the way its taxonomies use links. It doesn’t use RDF [Resource Description Framework], but it does use taxonomic links that are basically the same as RDF except for the actual tag format.

The largest companies have to file in XBRL. To meet those XBRL filing mandates, a lot of companies have centralized departments—sometimes within IT or within accounting’s own shadow IT—pull all the reports. Even ERP [enterprise resource planning] and things like it do not feed straight into the XBRL system. They have a firewall, very often, and I’m not an expert at XBRL implementations, but I’m very familiar with the space, and this is what I’ve understood. They have a firewall even between the centralized, highly controlled ERP of the enterprise and what goes into that XBRL filing, because even when you have something as highly controlled as, say, an enterprisewide ERP, it is not necessarily considered safe enough from the point of view of tight control by the party responsible for reporting.

It’s not a problem unique to semantic technology. Let’s say you had a situation where you had semantic technology on one end and you wanted information from that to go into a filing. You would still want the same sort of firewall where the auditors and the other experts could look at the semantic technology’s surface version of the truth as an input, but they would still decide what goes into the actual numbers for the filing. ■

“You would not want a semantic technology-driven system whose end point is the XBRL [Extensible Business Reporting Language] filing to the SEC [Securities and Exchange Commission]. That would be an absolute disaster. So there is absolutely a continuum—from departments and use cases where this is appropriate, cases where it’s a hybrid, and cases where you need very, very structured, centralized control.”

A CIO's strategy for rethinking "messy BI"

Take the initial steps toward your next-generation data architecture.



As CIO, you know you have an information problem. You've spent countless dollars and staff hours getting your data warehouse, financial systems, customer systems, and other transaction systems to generate meaningful reports. You've led Herculean efforts to regularize, transform, and load that data into consistent formats that business intelligence (BI), enterprise resource planning (ERP), analysis, reporting, dashboard, and content management tools can handle. Yet company executives keep asking for more detailed information to make better decisions, especially about the emerging challenges in the ever-changing markets the company is trying to navigate.

The reason for this state of affairs is not that BI and related systems are bad, but that they were designed for only a small part of the information needs businesses have today. The data structures in typical enterprise tools—such as those from IBM Cognos, Informatica, Oracle, SAP, and SAP BusinessObjects—are very good for what they do. But they weren't intended to meet an increasingly common need: to reuse the data in combination with other internal and external information. Business users seek mashup capabilities because they derive insights from such explorations and analyses that internal, purpose-driven systems were never designed to achieve. PricewaterhouseCoopers calls this “messy BI.”

People have always engaged in informal explorations—gleaning insights from spreadsheets, trade publications, and conversations with colleagues—but the rise of the Internet and local intranets has made information available from so many sources that the exploration now

possible is of a new order of richness and complexity. Call it the Google effect: People expect to be able to find rich stores of information to help test ideas, do what-if analyses, and get a sense of where their markets may be moving.

There's no way traditional information systems can handle all the sources, many of which are structured differently or not structured at all. And because the utility of any source changes over time, even if you could integrate all the data you thought were useful into your analytics systems, there would be many you didn't identify that users would want. You don't want to create a haystack just because someone might want a specific straw at some point.

There's no way traditional information systems can handle all the sources, many of which are structured differently or not structured at all.

Tom Flanagan, CIO of Amgen, a biomedical company, sums up the problem: “It is difficult to get the business to very accurately portray what its real requirements are. With the type of business intelligence that we have, almost invariably we end up having to build these data cubes, and we build them based on the requirements the business gives us. What we build oftentimes does not meet the business expectations. It may meet what they

said they wanted, but in actuality they want a very flexible type of reporting that gives them the ability to drill down to whatever layer of detail they want. So, the challenge with the historic way of providing reports is that it does not meet that flexibility that the business demands.”

A more flexible information architecture

Fortunately, the emerging concept of Linked Data points to how CIOs can extend their information architecture to support the ever-shifting mass of information sources not tidily available in enterprise information systems. (The article, “Spinning a data Web,” on page 4 explains the technologies behind Linked Data and how they can augment technologies you already have. Also see <http://linkeddata.org/> for a detailed description of Linked Data.) The Linked Data approach can help CIOs provide what their business colleagues seek by bringing in a more flexible, agile information architecture that unlocks more value from their current information systems and extends its reach to the wealth of information beyond them. (See Figure 1.)

Information systems are typically deployed on the premise that if you migrate enough data to them, you’ll get better decisions—as if software systems could replace human insight. But the premise is false, treating everything as predictable or static, known or knowable, and therefore capable of being automated. The Linked Data concept understands that this is not the case; it focuses instead on helping people to identify relevant information and to analyze it better. Humans excel at this kind of relevance processing, so why not take better advantage of their ability?

But simply using Linked Data technologies is not the path to success either. Throwing tools based on the Resource Description Framework (RDF), Web Ontology Language (OWL), and other evolving Semantic Web technologies at business users, or letting them adopt technologies helter-skelter on their own, will only create chaotic inconsistency, a manifold version of the spreadsheet problem with which many CIOs already struggle.

CIOs shouldn’t aim to create a monolithic system to provide business staff the exploratory capabilities they seek. That would be an expensive, time-consuming

investment for something whose value is difficult to quantify and whose best practices are not yet known. Instead, CIOs need to create what PwC calls an information mediation layer that lets business staff explore what-if scenarios, assess strategies and risks, and gain insight from the messy reality of the world inside and outside a company’s four walls.

As we explained in the Winter 2009 *Technology Forecast*, an information mediation layer orchestrates information from disparate sources for exploratory analysis rather than discovering an immutable “single source of truth” for archival and reporting purposes.

The CIO needs to create the framework for exploration, one that helps the analysis fit meaningfully with the

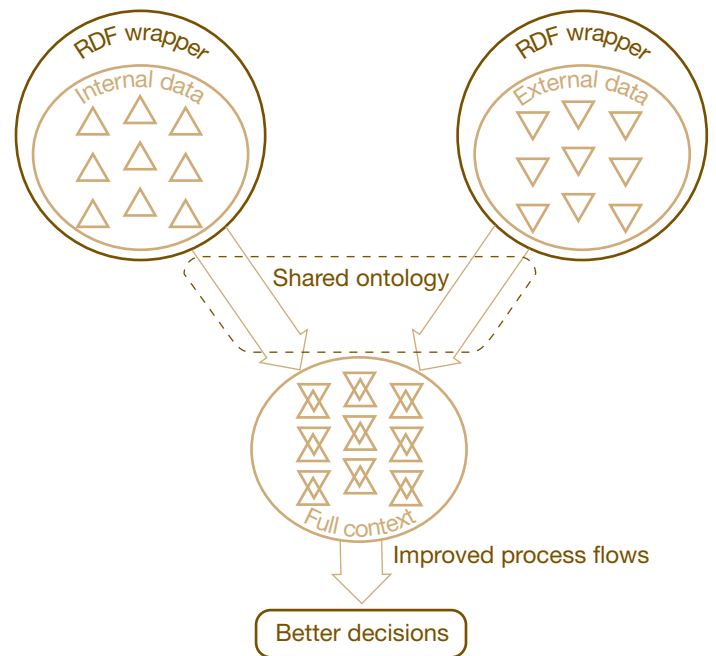


Figure 1: How internal and external data elements provide context

Your existing data warehouse can be reused by exposing warehouse data through RDF wrappers, but RDF is not sufficient. There must be a master plan—a metadata plan derived from ontologies—and to make use of external data through linking, the ontologies must have some shared elements.

Source: PricewaterhouseCoopers, 2009

enterprise's existing information sources and their often unstated assumptions—without limiting that exploration or imposing a closed worldview on it. The goal of this framework and its associated tools is to allow mapping and filtering on the fly, so you don't have to conduct expensive, time-consuming normalization activities for one-off or low-volume analyses—assuming those were even possible.

“There will always be many sources of data, and there will always be new types of data. ... We need an architecture that will accommodate this [change],” says Lynn Vogel, vice president and CIO of the M. D. Anderson Cancer Center, a hospital and medical research organization in Houston.

Some advantages of Linked Data

Unlike corporate data warehouses and other standard information systems, the Linked Data concept accepts that information has different structures depending on the purpose and context in which it was created. Linked Data tries to bridge those differences using semantics (the meaning of the information in context) and ontologies (the relationships among information sources).

Think of Linked Data as a type of database join that relies on contextual rules and pattern matching, not strict preset matches. As a user looks to mash up information from varied sources, Linked Data tools identify the semantics and ontologies to help the user fit the pieces together in the context of the exploration. The tools do not decide the connections, although the use of RDF and OWL tags can help automate the initial state for the user to review before applying human intelligence.

Many organizations already recognize the importance of standards for metadata. What many don't understand is that working to standardize metadata without an ontology is like teaching children to read without a dictionary. Using ontologies to organize the semantic rationalization of the data that flow between business partners is a process improvement over electronic data interchange (EDI) rationalization because it focuses on concepts and metadata, not individual data elements, such as columns in a relational database management system.

The ontological approach also keeps the CIO's office from being dragged into business-unit technical details and squabbling about terms. And linking your ontology to a business partner's ontology exposes the context semantics that data definitions lack.

Applying the Linked Data approach complements architectural approaches such as service-oriented architecture (SOA), inline operational analytics, and event-driven architectures that allow various functions to interact as needed to create a dynamic, flexible result that stays within the specified bounds. And it supports the inter-enterprise process flows common in today's networks of value chains, whether a traditional supply-and-delivery chain of retailing goods or an information-validation chain such as that of the pharmaceutical industry and its regulators.

Linked Data technologies, such as RDF, also have scalability and efficiency in their favor, says Jason Kolb, a technical lead at Cisco Systems who previously ran a BI company called Latigent. “By contrast, data warehousing's cost and inefficiency may be prohibitive at the large scale necessary in the near future,” he says.

Two paths for exploring Linked Data

PwC recommends that CIOs begin to rethink their information strategy with the Linked Data approach in mind. We do not recommend you embark on a big-bang initiative; that's unrealistic for an emerging technology whose best practices have yet to be learned. But we do recommend you test some of the principles of this approach as part of your larger information and data efforts. Here are some specific suggestions for ways to do this.

Depending on your own strengths and priorities, we see two possible paths for you to take with Linked Data technologies such as RDF and OWL. The paths are not mutually exclusive; you could pursue both if resources and inclinations permit.

The first path would be to extend your current data warehouse and structured data stores to account for the missing dimension of ontology- and semantics-oriented metadata. This extension will provide the necessary context to your data, allowing uses beyond

the strict purpose originally intended. This extension could be phased in over time and would unlock more value from data investments you've already made. It would ensure a consistency at the core that does matter: You want a common language all the way through the stack—you want one way of describing your resources and their relationships throughout.

The second path would be to empower your business users with exploration tools that they could use with existing internal data and with external data of their choosing. These tools would let them find the best business cases and make immediate use of the Linked Data technologies at a low cost to IT, since most of these tools are reasonably priced. Think of this as building and operating the “car”—your technology platforms and associated processes—that executes the business users’ “driving.” In essence, you would create the heads-up dashboard display that has contextual and configurable gauges for the people driving your business—unlike the fixed gauges of today’s structured systems—and let them make their own assessments and explorations. In this approach, you let data become the applications, adding the power of action and insight to data.

Both approaches start from a common base: establishing a basic business ontology that expresses the relationships among the business’s key processes and entities. The ontology provides the common framework by which the various data sources—internal and external—can be “joined” in the exploratory analysis, ensuring that they are mapped to and filtered against common concepts no matter where they originated. The same ontology development could be extended outside your walls through partnerships with others in your industry, as Chevron is beginning to do in the oil and gas business. (See Figure 2 on page 37 and the interview with Frank Chum of Chevron on page 46.)

Conceiving a Linked Data strategy

Because of the emerging nature of the Linked Data approach that PwC forecasts will be crucial to an organization’s ability to deploy an information mediation layer, a CIO should approach the effort as directional and exploratory, not as a project to complete. The CIO

is in the best position to evangelize this concept, having both the knowledge of the core information systems already in place and the relationships with business users to understand their information needs—and to connect those to the possibilities of the Linked Data approach.

The explorations previously described would provide valuable insight into where the Linked Data approach truly helps solve “messy BI” issues and what technologies work best for areas deemed valuable. Thus, the CIO can adjust course and priorities without fear of being seen to under-deliver, thanks to the explicitly exploratory nature of any Linked Data effort.

Because Linked Data thinking is still evolving, the CIO should expect to bring in support for several areas, whether through consultancies, training, or staff members tasked to educate themselves. These areas include enterprise architecture models; RDF and OWL structures; taxonomies, semantics, and ontologies; scenario building for strategic thinking around enterprise domain subsets; and master data management (MDM).

The CIO must be prepared for the discovery that, despite their promise, the Linked Data technologies don’t deliver as hoped. “My guess is that semantic technology is the next stage of the process, but it’s too soon to tell,” says M. D. Anderson’s Vogel. Even so, the exploration of this approach should—at a minimum—create a better understanding of the organization’s “messy BI” problem and how it can be lessened. The exploratory effort burnishes the CIO’s reputation as a visionary and a strategic leader.

Creating the information framework

Identifying the benefits of an approach to handle the “messy BI” gap is itself a significant first step. Organizations either don’t know they have a problem, leaving them at risk, or they use inappropriate technologies to solve it, wasting time and money.

A CIO’s middle name is “information,” making the CIO the obvious person to lead the organization’s thinking about ontology, semantics, and metadata—the core values of information that make it more valu-

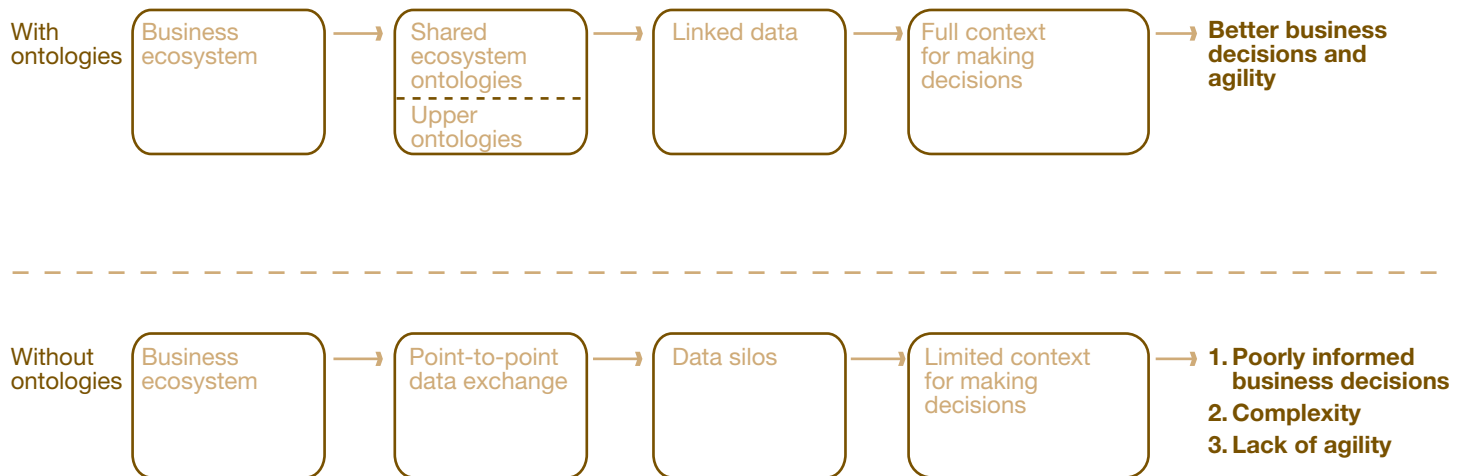


Figure 2: The business ecosystem value of shared ontologies

Silos prevent access to contextual information, and ontologies are a way to prevent siloing.

Source: PricewaterhouseCoopers, 2009

able to everyone than the typical structured data. The CIO should lead the enterprise’s information thinking, because the technology systems IT created and manages exist to deal with information. Losing sight of that shortchanges the business and relegates the CIO to little more than an infrastructure manager.

Therefore, the CIO should lead the development of the business ontology. The CIO should help key parts of the business—those with the highest business value—build their subsets. The CIO and the line-of-business managers will then have the key ontological domains in place that begin to create the metadata to apply both to new data and retroactively to existing data where it matters. Once in place, they can lead to harmonized operating models within the organization. And that leads to agility and better decision making. (See Figure 3 on page 38.)

For example, thinking about the ontologies of supplier and customer can create a better context for taking advantage of transaction data in mashups that combine with messy data to explore everything from potential product alternatives to unmet customer demands. In this way, you still can use the database-structured information at the base of your information stack without having to transform it for those flexible explorations.

To successfully apply semantics and ontologies to your existing structured data, you need that data to be consistent. All too often, an enterprise’s information management systems are inconsistent; they have differing data definitions (often handled through mapping at the information-movement stage) and, worse, different contexts or no context for those definitions (which results in different meanings). The classic cases are the definition of a customer and of a sale, but inconsistencies exist in all kinds of data elements.

Thus, it's crucial to focus on MDM approaches to rationalize the structured data and their context (metadata) in your existing information systems. The more inconsistent your internal systems are, the more difficult it will be to map that data semantically or ontologically to external sources. Thus, an MDM effort is imperative not only to reduce the cost and increase the effectiveness of your internal systems, but also to enable you to work with external sources using Linked Data approaches.

Thinking through your business ontology and semantics to create the right framework to support Linked Data explorations should help you think through your organization's overall information architecture, identifying which information has a contextual source of authority, which has a temporal source, and which has a single, master source. Knowing these sources of authority helps establish where the framework needs to be rigid and where it does not, as well as in what way it should be rigid or not.

For example, a customer might be contextual—an internal customer, an original equipment manufacturer (OEM), or an individual consumer—and thus the ontology allows multiple mappings to this concept that the user can choose from for a current exploration. But a part number is allowed to be only one thing.

Another key facet of ontologies and semantics is that they are not necessarily strictly hierarchical.

Relationships can occur at and across different levels, and some information is naturally arranged more in a tagged cloud than in a hierarchical form. This adds more flexibility to the Linked Data exploration for a real-world analysis, but it can be difficult for IT staff to think beyond hierarchical data arrangements. Take care not to force everything into a hierarchy.

Supporting business users' explorations

The benefit of information mediation is most immediate at the business unit level, where business analysts and other strategic thinkers are primed to explore ideas. "With properly linked data, people can piece together the puzzle themselves," notes Uche Ogbuji, a partner at Zepheira, which provides semantics-oriented analysis tools and training. In the past, he adds, assembling the puzzle pieces was viewed as an engineering challenge, leading to a resource-intensive effort to structure data for traditional analysis techniques.

Providing a baseline ontology for the business unit and helping its analysts "join" that ontology with those available outside the company through beta tools will let you test the value of this approach, test your ontology in real-world contexts, and create buy-in from key business users to drive further investment. Plus, ontology development, as the BBC Earth's Tom Scott says, is a

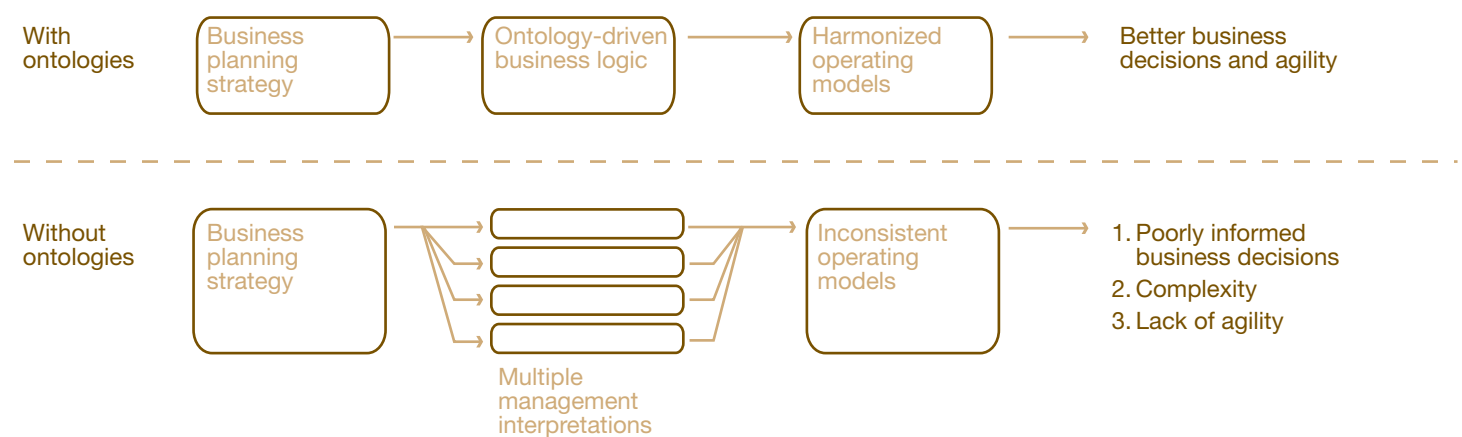


Figure 3: The business logic value of ontologies

Ontologies can work well to bridge the gap between strategy and operations by informing and harmonizing operating models.

Source: PricewaterhouseCoopers, 2009

A good rule of thumb is that semantic reach and control reach both telescope inversely to distance: The further information sources are from your core data, the less precise they will be and the more freedom users should have to manipulate their meaning to impose a precise context for their particular exploration.

“contact sport”—it is best when there’s lots of feedback, experimentation, and information exchange, so IT should not do it alone.

The CIO needs to loosen controls over the information and tools used in this exploration. You’re not building a system but testing an approach. Yes, you must retain control over your organization’s core information, which is typically the data used in transactions and archive systems. But the nondestructive nature of the Linked Data approach means you can expose that data more freely. In other words, users should be able to explore and transform data on the fly (since they’re not changing the source data and its metadata—just the copy, at most—in their virtual explorations).

A good rule of thumb is that semantic reach and control reach both telescope inversely to distance: The further information sources are from your core data, the less precise they will be and the more freedom users should have to manipulate their meaning to impose a precise context for their particular exploration.

Strategic application of information

The beauty of the Linked Data approach’s intelligent information linking is that data normalization is a transient state, like a database join, that leaves the original data untouched—eliminating the huge data-rationalization effort that usually destroys metadata along the way. This fact also makes it easier to bring external data into an analysis—from the Web, information brokers, and your value networks.

The key for an analysis is to map the metadata from the various sources, something that having a core ontology simplifies and that semantic tools help deliver for each exploration. Think of this Linked Data mapping as an information mashup. This linking is generally about providing context for the information being explored, and it’s the context that provides the specificity that makes the analysis useful. (See the interview with the BBC Earth’s Tom Scott on page 16 for an example use of mashups.)

Large, heterogeneous data sets can seem impossible to structure. However, creating domain-specific ontologies is feasible, and if they are shared, you can follow them to other domains and reuse what’s been created in those domains. Plus, the Linked Data approach means that “collaboration between departments can happen [because] full agreement doesn’t have to be enforced. Semantic technology makes it possible to agree to disagree, and the schemas can reflect the degree of agreement and disagreement,” Ogbuji says. These attributes show the powerful advantage of the Linked Data approach.

As CIO, you would be foolish to not put these approaches on your agenda. These approaches can help your organization perform better in ways that will improve the business—improvements based on what you ultimately are supposed to lead: the strategic application of information. Placing information mediation through Linked Data on your agenda puts you squarely in the strategic role your company increasingly expects you to play, rather than focusing on bits and bytes best left to hands-on technology-operations subordinates.

For more information on the topics discussed in this article, contact Steve Cranford at +1 703 610 7585.

How the Semantic Web might improve cancer treatment

M. D. Anderson's Lynn Vogel explores new techniques for combining clinical and research data.

Interview conducted by Alan Morrison, Bo Parker, and Joe Mullich

Lynn Vogel is vice president and CIO of The University of Texas M. D. Anderson Cancer Center. In addition, he holds a faculty appointment at The University of Texas in Bioinformatics and Computational Biology. In this interview, Vogel describes M. D. Anderson's semantic technology research and development and the hospital's approach to data integration.



PwC: Could you give us a sense of the IT organization you manage and the semantic technology projects you've been working on?

LV: M. D. Anderson has a little more than \$3 billion a year in revenue, about 17,000 employees, and a fairly substantial investment in IT. We have a little more than 700 people in our IT division. We do a significant amount of software development. For example, we have been developing our own electronic medical record capability, which is something fewer than a half dozen healthcare organizations in the country have even tried, let alone been successful with.

We tend, I think, to be on the high end of the scale both in terms of investment and in terms of pushing the envelope with technologies. For example, our electronic medical record is probably the single most complete model based on service-oriented architecture [SOA] that there is in healthcare, particularly in the clinical space. We have built it entirely on a SOA framework and have demonstrated, certainly to our satisfaction,

that SOA is more than simply a reasonable framework. SOA probably is the framework that we need across the industry during the next three to five years if we're going to keep pace with all the new data sources that are impacting healthcare.

In the semantic technology area, we have a couple of faculty who have done a lot of work on semantic data environments. It's turning out to be a very tricky business. When you have a semantic data environment up and running, what do you do with it and how does it integrate with other things that you do? It is still such a new development that what you would call the practical uses, the use cases around it, are still a challenge to figure out. What will be the actual impact on the daily business of research and clinical care?

We have an environment here we call S3DB, which stands for Simple Sloppy Semantic Database. Our faculty have published research papers on it that describe what we do, but it's still very, very much on the cutting edge of the research process about data

“One of the expectations, particularly around semantic technology, is that it enables us to provide not simply a bridge between clinical and research data sources, but potentially a home for both of those types of data sources. It has the ability to view data not simply as data elements, but as data elements with a context.”

structures. And although we think there’s enormous potential in moving this along, it’s still very, very uncertain as to where the impact is actually going to be.

PwC: When you’re working with S3DB, what kinds of sources are you trying to integrate? And what’s the immediate objective once you have the integration in place?

LV: The big challenge in cancer care at this point—and it really focuses on personalized medicine—is how to bring together the data that’s generated out of basic research processes and the data that’s generated out of the clinical care process. And there are a number of interesting issues about that. You talk to most CIOs in healthcare, and they say, “Oh, we’ve got to get our physicians to enter orders electronically.” Well, we’re starting to do that more and more, but that’s not our big issue. Our big issue is that a patient comes in, sees the doctor, and the doctor says, “I’m sorry to tell you that you have cancer of this particular type.” And if I were the patient, I’d say, “Doctor, I want you to tell me: Of the last 100 patients who had this diagnosis with my set of characteristics—with my clinical values, lab values, whatever else—who were put on the therapy that you are prescribing for me, what has been their outcome?”

On the one hand, that is a clinical question. I am a patient. I want to know what my chances are. At the end of the day in cancer, it’s about survival. So that’s the first question. On the other hand, it’s also a research question, because the clinician—to have an idea about the prognosis and to be able to respond to this patient—needs to know about the data that has been collected around this particular kind of patient, this particular kind of disease, and this particular kind of therapy. So one of the expectations, particularly around semantic technology, is that it enables us to provide not simply a bridge between clinical and research data sources, but potentially a home for both of those types of data sources. It has the ability to view data not simply as data elements, but as data elements with a context.

PwC: The data you’re talking about, is it external and internal data, structured and unstructured?

LV: Yes, it could be anything. Unstructured data is obviously a bigger problem than anything else. But even with structured data, integrating data from disparate sources is a big challenge. I might have gene expression data from a series of biomarker studies. I could have patient data in terms of diagnosis, lab values, and so on. Those are very different types of data.

“Either you optimize for the clinician, looking for one patient and that patient’s many values, or you optimize for the researcher, looking at very few values, but many, many patients.”

When you look at the structure of IT in healthcare today, it’s largely patient focused, on discrete values. I want to find out what Mrs. Smith’s hemoglobin level is. That’s a very discrete question, and it’s a very clear, simple question with a very discrete answer. In that process, the clinician is looking at one patient but is trying to assimilate many, many, many attributes of that patient. That is, he or she is looking at lab values, pictures from radiology, meds, et cetera, and working toward an assessment of an individual patient.

The research question turns that exactly on its head, just the reverse of the clinical question. The researcher is interested in looking at a very few attributes, but across many, many patients. Unfortunately, there isn’t a database technology on the market today that can reconcile those issues. Either you optimize for the clinician, looking for one patient and that patient’s many values, or you optimize for the researcher, looking at very few values, but many, many patients.

And so that kind of challenge is what confronts us. From a data management standpoint, you use the data that you get from gene expression studies to match patterns of data with association studies, which is really not what you’re doing on the clinical side. Now, having said that, our semantic tools are one way to bridge that gap. It is possible that semantic technologies will provide the framework within which both of these vastly different types of data can be used. I think this is going to determine the future of how successful we are in dealing with cancer. We’re not convinced entirely yet, but there are positive indications we’ve written about in our publications.

PwC: So is this part of the new emphasis on evidence-based medicine, or is it something else?

LV: Evidence-based medicine historically has focused on the data that says, if I have a patient with a particular problem, a given therapy will work. Basically the question is: Is this patient’s cellular structure and the kind of genetic expression that shows up in this patient’s cell—is that amenable to targeting with a particular therapy? So evidence-based medicine really covers the whole gamut of medicine. What we’re trying to figure out is at a much more granular level.

We’re looking for the relationship between the development of a cancerous condition and a particular gene expression, and then a particular therapy that will deal with that condition or deal with that diagnosis under the conditions of a particular gene expression.

PwC: Is the bigger problem here that the data has probably been collected somewhere in some context, but there are no standards for what you use to describe that data, and you need to stitch together data sets from many, many different studies and rationalize the nomenclature? Or is it a different problem?

LV: One of the biggest problems with genetic studies today is that people don’t follow highly standardized procedures, and the replication of a particular study is a real challenge, because it turns out that the control processes that guide what you’re doing sometimes omit things.

For example, we had a faculty group here a year or so ago that tried to look at a fairly famous study that was published about gene expressions in association with a particular disease presentation. And when they asked for the data set, because now you’re required to publish the data set as well as the conclusions, it turns out that it was an Excel spreadsheet, and when they had actually done the analysis, they had included the heading rows as well as the data, so it wasn’t quite what they represented.

So that's just kind of sloppy work. That doesn't even get to one of the big challenges in healthcare, which is vocabulary and terminology. You know, there could be 127 ways to describe high blood pressure. So if I described it one way, and you described it another way, and I put my data in my database, and you put your data in your database, and we combine our databases and do a search, we'd have to know all 127 ways that we've described it to arrive at any kind of a conclusion from the data, and that is very, very difficult.

PwC: As you look out three to five years, and presuming we continue to find more and more value from semantic technologies, where will it have the biggest impact in healthcare IT?

LV: I think one of the big challenges in healthcare IT is that IT investments, particularly in the clinical side of healthcare, are by and large driven by a small number of commercial vendors, and they sell exclusively into the acute care market, which is fine. I mean, it's a reasonable market to sell into, but they don't have a clue about the challenges of research.

If you look at what's happening in medicine today, medicine in general is more and more based on things like genomics, which is coming from the research side of the house. But when you talk to healthcare IT

“You have two directions you can go. You can try to cram it all into one big place, which is the model we had in 1992. Or, you can say there will always be repositories of data, and there will always be new types of data. We need an architecture that will accommodate this reality.”

vendors or look at their products, you discover that they have built their products on technologies and architectures that are now 15 to 20 years old.

PwC: A meta-model of health that's care focused.

LV: It is care focused, and in most cases, it's built on a single physical data repository model. It says, “Let's take all the clinical data we have from every place we have it and dump it into one of these clinical data repositories.”

Well, vendors have discovered a couple of things. One is that even the task of integrating images into that database is very, very difficult. In fact, in most vendor architectures, the image archive is separate from the data archive. And, frankly, it's fairly cumbersome to move back and forth. So that says that the architecture we had 15 years ago wasn't really built to accommodate the integration of imaging data. All you need to do is to step into the genomics world, and you realize that the imaging integration challenges only scratch the surface. You have no idea what you're running into.

PwC: Is the British exercise in developing a National Health Service electronic medical record addressing these sorts of data integration issues?

LV: Not to my knowledge. I mean, everybody now is working with images to some extent. National Health Service is trying to build its models around commercial vendor products, and those commercial products are only in the acute care space. And, they're built on the closed data models, which in 1992 were terrific. We were excited.

But that's really why we opted to go off on our own. We felt very strongly that there will always be two things: many sources of data, and new kinds of data sources to incorporate. And you have two directions you can go. You can try to cram it all into one big place, which is the model we had in 1992. Or, you can say there will always be repositories of data, and there will always be new

“As we continue to move forward, semantic technologies will have a role to play, just because of the challenges that data creates within the contexts that need to be understood and represented in a data structure.”

types of data. We need an architecture that will accommodate this reality, and, frankly, that architecture is a services architecture.

PwC: Do semantics play just a temporary role while the data architecture is being figured out? So that eventually the new data becomes standard fare and the role of semantics disappears? Or is the critical role of semantic technology enduring, because there never will be an all-encompassing data architecture?

LV: I think, quite honestly, the answer is still out there. I don't know what the answer is. As we continue to move forward, semantic technologies will have a role to play, just because of the challenges that data creates within the contexts that need to be understood and represented in a data structure. And semantic technology is one possibility for capturing, maintaining, and supporting those contexts. My guess is it's the next stage of the process, but it's really too soon to tell.

Oracle now supports semantic representation, which basically means we have moved past rows and columns and tables and elements, to RDF [Resource Description Framework] triples. That's good stuff, but we're not clear yet, even with the time we've spent on it, where all this fits into our game. The technology's very experimental, and there's a lot of controversy, quite frankly. There are people who focus on this who have totally opposite views of whether it's actually useful or not, and part of the reason for those opposite views is that we don't really understand yet what it does. We kind of know what it is, but to understand what it does is the next test.

PwC: And do you think the community aspect—working collaboratively with the broader community on medical ontologies, terminology, and controlled vocabularies—is likely to play a role, or do you think that the M. D. Andersons of the world are going to have to figure this out for themselves?

LV: That's one of the things that worries me about the federal stimulus plan and its funding for electronic medical records. It's too strongly influenced by the vendor community. It's not that the vendors are bad guys; they're not. They're actually smart, and they offer good products by and large. They've all had their share of fabulous successes and dismal failures, and it just goes to the point that it's not the technology that's the issue, it's what you do with it that makes the difference.

PwC: But at this time they have no incentive to create a data architecture for electronic medical records that works in the way you desire, that is capable of being flexible and open to new sources of data.

LV: That is correct.

PwC: What about the outlook for interoperability on the research side?

LV: For all the talk of vendors who say they can talk to all of their own implementations, the answer is no, they can't. Interoperability is a big buzzword. It's been around for a long time. You know technology doesn't solve the organizational issues.

When you have 85 percent of the physicians in this country practicing in two- and three-person practices, that's a different issue from let's make everybody interoperable. The physicians today have no incentives to make the information technology and process change investments that are required for interoperability. I'm a physician, you come to see me, I give you a diagnosis and a treatment, and if I can't figure it out, I will send you to a specialist and let him figure it out, and, hopefully, he'll get back to me, because all he wants to do is specialist stuff, and then I'll continue on with you. But within that framework, there are not a lot of incentives to share clinical data generated from these patient interactions.

PwC: On the research side of the question, we've been reading about the bioinformatics grid and wondering if that sort of approach would have a benefit on the clinical side.

LV: I think it does. There are all kinds of discussions about the grid technology, and the National Cancer Institute has pushed its bioinformatics grid, the caBIG initiative. I think there has been significant underestimation of the effort required to make that work. People would like to think that in the future all of this stuff will be taken care of automatically. There's a new report just out from the National Research Council on Computational Technology for Effective Health Care. It's a fascinating discussion of what the future of medicine might look like, and it has an enormous number of assumptions about new technologies that will be developed.

All this stuff will take years to develop and figure out how to use it effectively. It's just very hard work. We can talk about semantic technology and have an interesting discussion. What's difficult is to figure out how you're going to use it to make life better for people, and that's still unclear.

“At the end of the day we do have to deliver for our patients.”

PwC: Siloed, structured, and unstructured data are part of the reality you've had for years.

LV: That's correct. And we'd like to eliminate that problem. You know, we have tons of unstructured data all over the place. We have a whole initiative here at M. D. Anderson, which we call our Structured and Clinical Documentation Initiative, which is addressing the question of how can we collect data in a structured way that then makes it reusable to improve the science? And people have developed a lot of ways, workarounds, if you will—all the way from natural language processing to scanning textual documents—because we have a ton of data that, for all practical purposes, will never be terribly useful to support science. Our commitment now is to change that initial process of data collection so that the data is, in fact, reusable down the road.

PwC: And there's an element of behavioral change here.

LV: It's also the fact that, in many cases, if you structure data up front, it will take you a bit longer to collect it. You could argue that once you've collected it, you have this fabulous treasure trove of structured data that can advance the science of what we do. But there's an overhead for the individual clinicians who are collecting the data. They're already under enormous pressure regarding the amount of time they spend with their patients. If you say, “Oh, by the way, for every patient that you see now, we're adding 15 minutes so you can structure your data so that we all will be smarter,” that's a pretty hard sell.

PwC: A reality check, that's for sure.

LV: Well, you can read the literature on it, and it is absolutely fascinating, but at the end of the day we have to deliver for our patients. That's really what the game is about. We don't mind going off on some rabbit trails that hold some potential but we're not clear how much. On the other hand, we have to be realistic, and we don't have all the money in the world. ■

Semantic technologies at the ecosystem level

Frank Chum of Chevron talks about the need for shared ontologies in the oil and gas industry.

Interview conducted by Alan Morrison and Bo Parker

Frank Chum is an enterprise architect at Chevron whose career as a computer scientist spans several decades. During the 1980s, he worked at Coopers & Lybrand and Texaco, where he focused on artificial intelligence. In December 2008, he co-chaired a World Wide Web Consortium (W3C) workshop that Chevron hosted about the Semantic Web in the oil and gas industry.



In this interview, Chum discusses the role of semantics in knowledge-intensive industries, Chevron's major steps to take advantage of Semantic Web techniques, and how the oil and gas industry hopes to emulate the healthcare industry's efforts in ontology development.

PwC: How will the Semantic Web address some of the business issues confronting Chevron?

FC: We spend a lot of time trying to find information. The area we call information management deals with unstructured as well as structured information. To help our geoscientists and engineers find the information they need to do their jobs, we need to add more accuracy, more meaning into their searches. The goal is to let them find not just what they're looking for, but also find things that they might not know existed.

PwC: At the end of the day, is this fundamentally about the business of looking for and extracting oil?

FC: Correct.

PwC: So there is a business decision at some point about what to do in a particular situation, given the information presented—whether to drill, whether to buy or lease or go into a joint venture with somebody else. Is that ultimately the funnel that this all points to?

FC: Yes. Actually, years ago as an artificial intelligence [AI] specialist with Texaco, I built a system for analogical reasoning. We modeled the important basins of the world that share certain characteristics. With that information, we were able to compare fields or sites that have similar characteristics and that probably would have the same type of oil production performance. That comparison involved the notion of inferencing—doing case-based reasoning, analogical reasoning.

Right now, analogical reasoning is very doable, and the benefits for the oil and gas industry compare with those

for the healthcare and life sciences industries. They have similar issues from the vantage point of drug discovery, protein mapping, and the like, so we're looking at those industries to try to model ourselves after them.

PwC: The W3C [World Wide Web Consortium] has a working group in that area. Do you hope to collaborate with some other folks in that group on a shared problem?

FC: Yes. When we joined the W3C and were working with them in Semantic Web areas, Roger Cutler, who's here at Chevron, joined the Semantic Web Health Care and Life Sciences Interest Group. He didn't have any connection to the industry-specific subjects they were talking about—he joined to learn from them how they were using the Semantic Web in a practical, industry setting. And so that's part of the reason why we had the oil and gas workshop—because we think we need a critical mass to do in the oil and gas industry what healthcare has done and to advance through that kind of industry collaboration.

PwC: What are you seeing in the Semantic Web Health Care and Life Sciences Interest Group that you specifically want to emulate?

FC: You wouldn't think that pharmaceutical companies would share a lot of information, but in fact the opposite is true, because the sheer amount of investment needed to develop a drug is in the billions of dollars in research. It's the same thing with oil companies; the amount of money invested is in the billions. Building an offshore platform can be a multibillion-dollar venture if it is in a hostile environment like deep water or the arctic, and these extremely expensive undertakings are often

joint ventures. So we need to be able to share lots of information not only with our joint venture partners, but also with the design and engineering companies that designed platforms as well as with the people who are going to manufacture the platforms, fabricate them, and put them in place. That's a lot of information, and without standardization, it would be difficult to share.

PwC: Does the standardization effort start with nomenclature? It seems like that would be really important for any set of business ecosystem partners.

FC: Yes. ISO 15926 is one initiative. There are many potentially confusing terms in drilling and production that benefit from having a common nomenclature. We also have standards such as PRODML—production markup language—and many other standards associated with exchanging data.

PwC: You've been involved in the oil and gas industry for quite a while. If you think about the history of that industry, how does the Semantic Web represent building on top of what's come before? Or is it throwing a lot of stuff away and starting over?

FC: I think it's a different approach. I think the Semantic Web actually provides maturity to AI, in a sense. To quote Patrick Winston of MIT [Massachusetts Institute of Technology] on AI, he said, "You need to consider AI not as a whole system, but actually as a little piece of a system that would make the system work more or better." Consider AI as raisins in a loaf of bread to make the loaf of bread more flavorful. People thought of AI as an AI system, entirely a big thing, rather than

that nugget that helps you to enhance the performance. I think the Semantic Web is the same thing, because you're looking at the Web as a platform, right, and data semantics are that nugget that make the Web more meaningful because a machine can understand information and process it without human intervention, and, more importantly, make the connections between the information that's available on the Web—the power of Linked Data.

PwC: Is there a sense that you're trying to do something now that you would not have tried to do before?

FC: Four things are going on here. First, the Semantic Web lets you be more expressive in the business logic, to add more contextual meaning. Second, it lets you be more flexible, so that you don't have to have everything fully specified before you start building. Then, third, it allows you to do inferencing, so that you can perform discovery on the basis of rules and axioms. Fourth, it improves the interoperability of systems, which allows you to share across the spectrum of the business ecosystem. With all of these, the Semantic Web becomes a very significant piece of technology so that we can probably solve some of the problems we couldn't solve before. One could consider these enhanced capabilities [from Semantic Web technology] as a "souped up" BI [business intelligence].

PwC: You mentioned the standardized metadata that's been in development for a long time, the ISO 15926 standard. Are you making use of initiatives that predate your interest in the Semantic Web within an ontology context and mapping one ontology to another to provide that linkage?

FC: Yes, definitely. In my use case [see <http://www.w3.org/2008/12/ogws-report.htm>], I spell out what we call ontology-based information integration. Using ontologies to structure data actually increases flexibility, because you don't have to have everything to

begin with. You can model only what you need to for starters, enhance the ontology later, and then merge them together.

PwC: It's difficult enough to get people to think as abstractly as metadata. And then going from metadata to ontologies is another conceptual challenge. Have you found it necessary to start with a training process, where you teach people about ontologies?

FC: This is a good question, because I'm involved in the master data management initiative at Chevron, too. We want to have shared definitions of concepts so that there is no ambiguity. And people need to agree with that shared definition. So in your own department you want to be in consensus with what the enterprise definition is for, let's say, people or contractors or whatever.

It's part of another project, what we call our conceptual information model. It looks at everything going on in Chevron, and we have developed 18 or 19 information classes. And then within these classes there are some 200 high-level categories that probably can describe all of Chevron's business activities. So that is ongoing, but the semantic part is what actually provides the mapping of how one of these concepts or terms relates to another.

PwC: How does the conceptual information model relate to the master data management effort?

FC: It was a precursor to the master data management initiative, all part of what we call our enterprise information architecture. The key concern is shareability. Some of these concepts need to be shared among different departments, so we need to harmonize the conceptual information model across departments. That is the other approach. But in an ontology, we aren't attempting to develop an all-inclusive, comprehensive information model of Chevron. We have more of a pragmatic approach in ontology building. We focus on building out what is needed for a specific solution, and we rely on

the flexibility of ontologies that let us merge and stitch linked ontologies together for information integration.

PwC: Can you give us an early example where you had a limited scope and a specific problem, and you used the ontology approach to generate positive results and a deliverable benefit?

FC: In the case study and in the workshop, we looked at our UNIX file systems, which hold much of our technical data. The original idea was to think of it as merging—not just the UNIX file system, but also the Windows environment—so when you do a search, you’ll be searching both at the same time.

We started with the UNIX part. We scraped the directory structure, and then we were able to extract metadata from the directory that was from the file path, because of the way the users name the path.

PwC: A folder system.

FC: Right. Folder systems or what we call file plan in the sense that they contain certain metadata. And together with the file path, we know who created that, and so we have a kind of people ontology. And then we have the project they’re working on and metadata information about when the file was created, when it was modified, and who did it. So we gathered all this information and put in place a system that described what a file was created for, who worked on it, for what project, at what time. We were then able to put in queries and ask who is working on certain projects at a certain time. This information was not in a database system, for example, but is implicit in the file metadata.

PwC: So are you looking at specific approaches to understanding the unstructured content itself, the information inside the files?

FC: Looking inside the files at the unstructured content is something we’ve talked about doing but we haven’t gotten there yet. There are an awful lot of different kinds of files in these repositories, and many of them are binary files that don’t contain easily recognizable text.

“Consider AI as raisins in a loaf of bread to make the loaf of bread more flavorful.”

Finding widely applicable ways of getting information out of the contents may be a considerable challenge. That’s why we started where we did. We do, however, also have a lot of spreadsheets, and we’re looking at ways to link information in the spreadsheets to ontologies that link and organize the information.

PwC: It seems like you’d be able to take the spreadsheets and, in conjunction with a more structured source, make sense of those sheets.

FC: That’s not easy, however, because spreadsheets can be very diverse. One way is to build ontologies from those spreadsheets. With the appropriate tool, we could import them, figure out an ontology for them, and then externalize that ontology to link one spreadsheet to the others. Once we’re able to take these spreadsheets and externalize the metadata and so on, then we’d be able to integrate them directly into workflows.

PwC: This sounds like one of those situations where you have a tool that’s helping you with, say, spreadsheets, and it’s not going to be 100 percent correct every time. It’s going to be a bit messy. So how do you approach that? Is there a feedback loop or a Web 2.0 quality to it, something that makes it a self-correcting process?

FC: We haven’t gotten that far yet, but I assume that is our next step. The University of Texas has implemented a decision support system in spreadsheets. We also run decision support systems on spreadsheets, but The University of Texas implemented it in an ontology-based semantic technology, and it’s very innovative. But we are getting there. We’re taking baby steps and are always looking at how we can use this technology.

“Two years ago, they would call it the O word.”

PwC: How far is this effort from the actual end users, such as an engineer or an earth scientist? If you were to say “ontology” to someone in that role at Chevron, do their eyes roll up into the back of their heads? Are they familiar with this effort, or is this a back-room effort?

FC: Well, I would say it was that way two years ago. They would call it the O word. “You’re bringing out the O word again.” Everyone said that. We have since made people aware of what this is. A major goal of the W3C workshop was to build awareness not just for Chevron but throughout the industry. That’s part of the objective: to get not only Chevron comfortable with the Semantic Web, but also BP, Total, Shell, and so forth. Within the Chevron community, there is more and more interest in it, especially on the information architecture side of things. People are interested in how ontologies can help and what an ontological approach brings that traditional approaches don’t, such as EII [enterprise information integration]. When we say data integration, they respond, “Aren’t we already doing it with EII? Why do we need this?” And so we are having this dialogue.

PwC: And what is your answer in that situation? What do you say when somebody says, “Aren’t we already doing EII?”

FC: In EII, you get only what is already there, but working with the Semantic Web, we call it the open world reasoning instead of the closed world. In databases, in EII, you’re connecting the data that you have. But with the Semantic Web, you’re connecting to much more information. Instead of saying, “If it’s not in the database, it’s false,” we only say that we don’t know the answer. Linking to new information is far easier. We are better able to integrate it, leading to a better business decision.

PwC: We talked to Tom Scott at BBC Earth [see page 16]. His primary focus is on leveraging information about music and artists that is already on the Semantic Web. He wouldn’t describe himself as an IT person. He’s more of a product manager. Is that something you see in some of the people you work with?

FC: Definitely. For example, some of the people we work with are geoscientists. Among them there’s already a big effort called GEON [Geosciences Network]. They are building ontologies for the different earth structures and trying to use them within our IT environment.

PwC: It sounds like in your case you have the best of both worlds. You have the central function thinking about leveraging semantic technologies, and then you have people in the earth sciences domain who are living in that domain of knowledge all the time.

FC: Yes, the best thing about them is that they know their pain point—what they want done—and they are constantly thinking about what can help them to solve the problem. In another sense, the earth science SMEs [subject matter experts] know that they need to be able to describe the world in ways that can be shared with other people and be understandable by machines. So they have this need, and they look into this, and then they call us and say, “How can we work with you on this?”

PwC: Do you work with them mostly graphically, using friend-of-the-friend bubble charts and things like that? Is that how you typically work with these domain folks?

FC: The tools that support Semantic Web initiatives are getting more and more sophisticated. We have SMEs who want to get a copy of the ontology modeling tool. They want a copy of that so they can work with it.

PwC: These are business users who want to get their own ontology modeling tool? How do they even know about it?

FC: Well, we [IT] do our modeling with it. We showed them an ontology and validated it with them, and then they said, “Whoa, I haven’t—this is good, good stuff,” so they wanted to be involved with it too and to use it.

In the Chevron world, there are a lot of engineers. They are part of the Energy Technology Company [ETC], and we are part of ITC, the Information Technology Company. ETC has some more people who have domain knowledge and also want to experiment with the new tools. As soon as we show them, they want it. Before, they were looking at another knowledge modeling tool, but the ontology tool is really capable of making inferences, so they want that, and now we are getting more and more licenses and using it.

PwC: Do you sense some danger that we could have a lot of enthusiasm here and end up with a lot of non-compatible ontologies? Are we going to enter a period where there will need to be some sort of master data model, a master ontology model effort?

FC: We already defined some standards to address that. We have a URI [Uniform Resource Identifier] standard for how you name ontologies, and it’s referenceable so that you can go into that URI and retrieve the ontology. We tried to make that shareable, and we are also starting a community type of space.

PwC: Is it discoverable somehow? If some employees somewhere in Saudi Arabia decide they need to get started with an ontology, would there be an easy way for them to find other ontologies that already exist?

FC: We’re standardizing on a [Microsoft] SharePoint platform, so we have a SharePoint site on information discovery that has Semantic Web or entity extraction for these unstructured texts and different analytics.

We have publicized that through communications, and we have people posting their work there. We try to make use of the collective intelligence kind of notion, like Wikipedia—have people come to it and have a discussion.

PwC: So you’re taking the Semantic Web out of this innovation research program within Chevron and moving it into the delivery side of the organization?

FC: We have a number of projects within the innovation, strategic research, proof of concept, pilot, to technology delivery continuum.

PwC: Is there a specific part of your business ecosystem where you are deploying it first?

FC: Yes. We are partnering with USC and have formed an organization called CiSoft [Center for Interactive Smart Oilfield Technologies, a joint venture of the University of Southern California and the Chevron Center of Excellence for Research and Academic Training]. We have created an application called Integrated Asset Management that uses Semantic Web technology to help with tasks associated with reservoir management. The end users don’t see anything that they would recognize as the Semantic Web, but under the covers it is enabling the integration of information about these assets.

PwC: You’re pretty confident that your initial Semantic Web applications are going to be in production and successful and part of the fabric of Chevron?

FC: Just because something performs well in a proof of concept, or pilot, doesn’t mean that it’s going to do well in production, right? We’re looking at scalability. That’s one of the big questions. When you’re dealing with billions of RDF triples, you wonder if it is going to give you the response time you need. We’re learning how to address this issue. ■

Acknowledgments

Advisory

Sponsor & Technology Leader
Paul Horowitz

US Thought Leadership

Partner-in-Charge
Tom Craren

Center for Technology and Innovation

Managing Editor
Bo Parker

Editors
Vinod Baya, Alan Morrison

Contributors
Galen Gruman, Larry Marion, Joe Mullich, Bill Roberts, Chrisie Wendin

Editorial Advisers
Larry Best, Brian Butte, Glen Hobbs,
Jim Kinsman, Bud Mathaisel, Justin McPherson,
Jonathan Reichental, Terry Retter, Deepak Sahi, Joe Tagliaferro

Copyedit

Lea Anne Bantsari

Transcription

Paula Burns, Dawn Regan

Graphic Design

Art Director
Howard Allen

Designers
Bruce Leininger, Diana Lira

Illustrator
Don Bernhardt

Photographer
Marina Waltz

Online

Director, Online Marketing
Jack Teuber

Designer and Producer
Joe Breen

Review

Paula Adler, Richard Beaumont, Mike Bergman,
Frank Chum, Steve Cranford, Roger Cutler,
Sean Middleton, Uche Ogbuji, Tom Scott

Marketing

Bob Kramer

Special thanks to

Dil Aneja, Jim Fisher, Larry Prager, Gerard Verweij,
Charlotte Yates

Industry perspectives

During the preparation of this publication, we benefited greatly from interviews and conversations with the following executives and technologists:

David Choy, senior consultant, and Patricia Anderson, senior marketing manager, EMC

Frank Chum, enterprise architect, Chevron

Tom Davenport, president's distinguished professor of information technology and management, Babson College

Joey Fitts and Bruno Aziza, business intelligence market strategy and execution, Microsoft

Tom Flanagan, chief information officer, Amgen

Chris Harding, director, Semantic Interoperability Working Group, Open Group

Scott Jarus, chief executive officer, Cognition Technologies

Jason Kolb, technical lead, Cisco Systems

Pankaj Malviya, chief executive officer, Longjump

Uche Ogbuji, partner, Zepheira

Mike Psenka, chief executive officer, eThORITY

Phillip Russom, analyst, TDWI

Tom Scott, digital editor, BBC Earth

Lynn Vogel, vice president and chief information officer, The University of Texas M. D. Anderson Cancer Center

pwc.com/us

To have a deeper conversation
about how this subject may affect
your business, please contact:

Paul Horowitz
Principal, Technology Leader
PricewaterhouseCoopers
+1 646 471 2401
paul.j.horowitz@us.pwc.com

This publication is printed on Coronado Stipple Cover made from 30% recycled fiber and Endeavor Velvet Book made from 50% recycled fiber, a Forest Stewardship Council (FSC) certified stock using 25% post-consumer waste.



Recycled paper

Subtext

Data federation	A form of scalable, virtual integration in which the actual data remain where they are, rather than being moved from their sources.
Linked Data	A means of exposing, sharing, and connecting individual data elements with the help of fixed addresses or global identifiers called Uniform Resource Identifiers (URIs).
Resource Description Framework (RDF)	A World Wide Web Consortium (W3C) data model that allows relationships between data elements to be described in graph form, a form that makes large-scale federation of disparate data sources possible.
Semantic Protocol and RDF Query Language (SPARQL)	The W3C's recommended standard for querying Web data in RDF graphs. In an RDF-based environment, graphical tools with SPARQL engines can join and query hundreds of sources through a point-and-click interface.
Semantic Web	An evolution of the World Wide Web in which data descriptions are explicit, making it possible to federate, query, browse, and gather information from disparate internal and external sources. The result is more complete and relevant information.
Ontology	A description of the characteristics of data elements and the relationships among them within domains. Ontologies describe relationships in an n-dimensional manner, illuminating relationships of multiple kinds among elements, whereas taxonomies show just hierarchical relationships.

Comments or requests? Please visit www.pwc.com/techforecast OR send e-mail to: techforecasteditors@us.pwc.com

PricewaterhouseCoopers (www.pwc.com) provides industry-focused assurance, tax and advisory services to build public trust and enhance value for its clients and their stakeholders. More than 155,000 people in 153 countries across our network share their thinking, experience and solutions to develop fresh perspectives and practical advice.

"PricewaterhouseCoopers" refers to PricewaterhouseCoopers LLP or, as the context requires, the PricewaterhouseCoopers global network or other member firms of the network, each of which is a separate and independent legal entity.

© 2009 PricewaterhouseCoopers LLP. All rights reserved.

The content of this document is provided "as is" and for general guidance on matters of interest only. The opinions expressed by people quoted in the document do not necessarily represent the opinions of PricewaterhouseCoopers. Although we believe that the information contained in this document has been obtained from reliable sources, PricewaterhouseCoopers is not responsible for any errors or omissions contained herein or for the results obtained from the use of this information. PricewaterhouseCoopers is not herein engaged in rendering legal, accounting, tax, or other professional advice or services.