



Slowly Maturing, Semantic Web Technologies Reach Pilot Project Stage in Pharma

March 03, 2010

Newsletter: [BioInform](#)

By [Vivien Marx](#)

Semantic technology appears to be slowly finding its way into pharma's informatics toolbox, as proponents of the technology reported at a recent conference that they are moving beyond prototypes and into pilot projects that put semantic technologies to work.

At last week's Conference on Semantics in Healthcare and Life Sciences in Cambridge, Mass., scientists from pharma, vendor firms, and academia discussed opportunities for semantic technologies in life science research, as well as barriers to widespread adoption and strategies for swaying skeptical colleagues.

Several pharma researchers in attendance told *BioInform* that they are shifting from a more abstract sell of semantic technologies to producing usable and beneficial pilot projects.

Unlike previous CSHALS meetings, where the presentations were more about smaller prototype projects, this year's meeting showed "serious projects," said Ted Slater, assistant director of systems and knowledge discovery at Boehringer Ingelheim.

Despite some progress, however, many speakers noted that they are still seeing opposition among upper management in pharmaceutical firms, as well as resistance from colleagues in IT who are struggling with tightening budgets and workforce cuts and are skeptical about the return on investment for semantic tools.

In one session a show of hands made it clear that most pharma scientists face an "uphill battle" when proposing semantic web projects, yet many said that the power of the technology will make these battles worthwhile.

One pharma scientist, who did not wish to be identified, said that large firms don't change easily, which presents hurdles in development. In particular, IT departments tend to distrust semantic approaches because they appear to be a big departure from current tools.

While traditional relational databases are akin to different human languages, semantic technologies are "a schema-less way of saying things," Slater said. Expressing concepts through semantic approaches, in subject-predicate-object triples, lets scientists reason over concepts and pose complex queries, "allowing us to ask questions we couldn't ask" otherwise, said Michael McGlashen, executive director of knowledge management in worldwide licensing and external research at Merck.

"To me it seems that most pharma companies are using semantic web to some extent," Susie Stephens, director of biomedical informatics in J&J's Pharmaceutical R&D, division told *BioInform*. She noted that this interest is being driven by the trend towards translational research, which requires an awareness of many data sources and the need to integrate, mine, and analyze data from different parts of the business and different industries.

Scientists applying semantic methods can save the time needed to buy a new server, design a database schema, fill it, and deploy it for every new dataset, Slater said. Once storage of subject-predicate-object triples is set up, a new database only requires downloading an RDF file into that triple store. "You load it and you're done," Slater said. "Now your data integration problem is gone."

Because semantic web technologies are standards-based, researchers can integrate disparate resources much faster and cheaper than with standard development tools, said Lee Feigenbaum, vice president of technology and standards at semantic web consulting firm Cambridge Semantics.

Resource Description Framework, or RDF, is the language used to represent information in the semantic web. RDF statements formed in the subject-predicate-object format, as triples, can be collected and searched as machine-readable graphs with each component of a triple tagged with a Uniform Resource Identifier.

There are several semantic syntaxes available, including RDF/XML, N-Triples, and N3. Turtle, or terse RDF Triple Language, is frequently used by developers as a somewhat more user-friendly alternative to RDF, Feigenbaum said. RDF graphs can be queried with the SPARQL query language. Many tools in this space are open source, such as D2R, a way for mapping relational databases with RDF.

Jim Hendler, a computer scientist from Rensselaer Polytechnic Institute and former chief scientist of the information systems office at the US Defense Advanced Research Projects Agency, said in his talk that semantic technology is no longer "deep academic cogitation" about a future application but is "maturing" technology that can sit on top of web infrastructure, is extensible, and oriented toward data-sharing.

Hendler cited signs of "commercial excitement" regarding semantic technology, including Microsoft's acquisition of semantic search engine Powerset for \$100 million in 2008 and new semantic vendors such as Sandpiper, Intellidimension, Intellisophic, and Ontology Works that are emerging even though "this is not a friendly market for new companies."

Eric Neumann, director of the pharma consulting firm Clinical Semantics Group, highlighted that in many firms only a "small, small group" of people understand semantic approaches and enterprise IT groups remain skeptical. Hendler responded that perhaps not everyone in an IT group need be "deeply" involved in semantic projects, which, given

the increasing availability of tools, can now be set up as a "show 'em process."

Living the Ecosystem

In his talk, Tim Schultz, senior analyst at J&J Pharmaceutical R&D IT systems engineering, presented the firm's semantic data pilot called knowIT, which has a focus on translational neuroscience and enables "novel translational queries" of available data.

J&J R&D has a "huge data ecosystem" that includes deep and shallow repositories and flat files on network shares, Schultz said. Stephens noted that when her team assessed J&J scientists' needs, they found that researchers "know about their favorite one or two data sources, but they didn't know the others."

The J&J team leveraged an existing semantic media wiki, used to catalog IT infrastructure at the company, and extended it to include metadata fields that describe data sources and can be accessed via SPARQL queries.

The wiki captures metadata about the data source, not the data itself, she said. It includes five tabs that describe the data source, the business owner, technical contact, licensing information, data source interface, and content captured with keywords from the Neuroscience Information Framework, or NIF, ontology.

"If there is an RDF representation of the data available, it includes the URL for the SPARQL endpoint," she said.

Stephens and her team are adding more data that has an RDF representation. "That's what makes it easier for scientists to do a quick dig-down into the data," she said.

One data source is from ADNI, the Alzheimer's Disease Neuroimaging Initiative, a consortium that J&J sponsors. The data relates to MRI scans that have been analyzed with a variety of tools, Stephens explained. "There are different tools to measure the same thing, so that does lead to the data becoming complex."

The team has integrated the system into its in-house R&D informatics platform 3DX, which has analysis and visualization capabilities. "We've extended 3DX to speak SPARQL, so then you can issue a query from 3DX that goes to knowIT," to show information about data sources within 3DX.

The team used the open source tool D2R to map ADNI data captured in relational databases to RDF, which took about a month. Mapping can be straightforward, such as when using resources such as DrugBank, which includes drug names and attributes, she said. "But, when you are looking at an experimental dataset, the mapping becomes more complicated." For example, a particular data set might involve an MRI on a particular date, with a machine of a particular type, on a patient whose brain has a particular hippocampal volume, measured with a particular analytical approach.

Relational databases are "wonderful" to store data, she said. "I like the idea of using D2R to do the mapping to RDF," she said. "You're not moving the data around and I think that is a good approach."

The J&J team has entered 120 data sources into its resource thus far. If an RDF

representation of the data exists, clicking on an icon will lead a scientist to the data itself. "A big yellow button" makes it easy for scientists to upload their own resources, she said.

"We are also adding social aspects to the data," Schultz explained. Specifically, the team is integrating the resource with the semantic tool Friend of a Friend, or FOAF. Each J&J employee has a FOAF profile, he said.

Because the list of data sources will grow beyond 120, the team is working on ways to visualize the output in clusters for additional querying. Future plans include applying void, or Vocabulary of Interlinked datasets, an RDF-based schema to describe how datasets are linked to each other, to avoid the need to manually describe data sources.

No Killer App?

Some semantic web proponents have said that the technology is still waiting for its "killer app." While Hendler noted that this could help grow the technology, others at the conference said that may not be needed.

"XML didn't have a killer app," M. Scott Marshall told BioInform. He is a researcher at the University of Amsterdam and co-chairs the W3C Semantic Web for Health Care and Life Sciences Interest Group with Stephens.

"I think the killer app is an invisible app, if you will," because applying the technology means not hitting a data integration problem, said Joanne Luciano, who runs the consulting firm Predictive Medicine. Semantic technologies will help avoid manual curation, and the ongoing struggle with integrating datasets in many labs and firms, she said.

Raul Rodriguez-Esteban, another researcher at Boehringer, highlighted in his talk that integrating large datasets means applying semantic web and other technologies to make the data more like a "mixture." He said that researchers must explore ways to reflect the varying quality of data sources and varying confidence scores associated with the data.

Vijay Bulusu, senior manager of R&D informatics at Pfizer, presented a proof-of-concept project that applied semantic web approaches to the scale-up phase of drug discovery. He said that pharma scientists and development staff are dogged by incompatible databases, and that even common concepts such as compounds lack harmonization. He noted that the speed of development and deployment for the project helped sell it internally ([see related story, this issue](#)).

Stephens agreed that there is still "not a standard way to represent the data," such as compounds, and noted that the W3C Health Care and Life Sciences Interest Group has projects underway to develop a translational medicine ontology and another early-stage project to addresses collaborations between pharmaceutical firms and external partners.

W3C is looking to build alliances with several pharma consortia such as Pistoia and the Prism Forum. CSHALS attendee John Wise, who directs the Prism Forum and leads informatics at Daiichi Sankyo, welcomed the efforts and said it might be a way to relay the importance of semantic technologies to management.

Going Superduper

Martin Leach, executive director of IT for discovery and pre-clinical sciences at Merck, said that semantic web technologies can help pharma handle its challenge of multiple disconnected and often unstructured data sources. "We can't just build one big superduper data model that does everything," he said.

He noted, however, that a big component in improving pharma's informatics pipelines will be "changing the way we work" rather than buying another piece of technology.

Leach said that his department supports 6,700 scientists and over 400 software applications, and that he is seeking to help as many users as possible while also trying to drive down the cost of IT. While he approves of experiments with semantic technologies, he admitted some in his firm and other companies have flopped.

Too many proof-of-concept projects or "one-offs" without "a big bang" will not drive adoption of semantic methods, he said.

As an example, he noted that after acquiring a biologics firm, his group built a data model around that company's data with a web-based front end. However, it ended up being "an efficiency play versus something that really added true value to the data." In addition, he said that his group chose the "wrong endpoint" to "show the value of semantics."

Leach recommended that semantic web proponents should continue with small-scale projects, but also seek a "careful balance" with projects that have greater impact on more users and the business at large.

One semantic experiment at Merck, Insight X, involves "hanging data" off of high-content screening images. In a collaboration with Entagen, a Newburyport, Mass.-based computational biology firm, "we built a semantic model around that data coming off of high throughput and ultra high throughput screens" that allows scientists to drill into the data in a way that is akin to Google Earth, he said.

So far, he said, the rapidity of deployment and the way the data maps to the screening plate have met with the approval of the Merck community.

"High value" propositions exist in many areas including organizing clinical trials, scouting for new licensing opportunities, and for methods that help researchers work with structured and unstructured data from various sources — particularly as pharma increasingly externalizes research, he said.

Leach added that there might be a "saturation" of semantic tools in the -omics space, particularly for tools "gluing giant haystacks together," and warned that information might be lost if the "right kind of sifting" is not completed upfront.

Another challenge will arise from network analysis, which requires new approaches to browse, expand, and visualize large data sets. "You are not going to use a laptop to do that," Leach said.

