Pfizer Partners with IO, Franz on Semantic Proof of Concept to Build Bridges Between Data Resources

March 04, 2010

Newsletter: BioInform

By Vivien Marx

Pfizer has partnered with two semantic web technology firms to rapidly develop and deploy a proof-of-concept system for integrating drug discovery data.

Vijay Bulusu, senior manager of R&D Informatics at Pfizer, discussed the system, called Information Dissemination Extraction and Aggregation, or IDEA, at last week's Conference on Semantics in Healthcare and Life Sciences in Cambridge, Mass.

Bulusu said that the project, intended to serve as a "real-world example" of semantic technology in the pharma setting, was expected to take four to six months to accomplish, but was actually delivered in six weeks.

While Bulusu weighed the use of open-source software for the project, he determined that approach would require "a lot of investment on our part" in terms of skilled staff and resources, and chose to work instead with two semantic technology vendors: IO Informatics and Franz.

The Pfizer team applied a combination of Franz's AllegroGraph RDF database and IO's Sentient Web Query and Knowledge Explorer.

"I needed a back-end repository and I needed a tool that could host a SPARQL end-point," Bulusu said. Franz's AllegroGraph provided a system to store the "semantic triples" used in the semantic language Resource Description Framework, or RDF, and IO's system served as the front-end user interface for the platform, which was intended to serve a range of users, some of whom are "not so technology-savvy," he said.

"We were able to create an application ontology, do the RDF translation, and start testing SPARQL in a matter of days," Robert Stanley, president and CEO of IO Informatics, told BioInform this week via e-mail.

Franz and IO integrated their tools and converted Pfizer's data to RDF to enable querying with SPARQL. Bulusu and his team copied a year's worth of datasets to a server to which

IO and Franz had access, and helped them understand the data and its structure. The companies had not worked together before, he said.

IDEA is not intended to be "a system of record" but rather to function as a "referential source" that points Pfizer's research staff "in the right direction as to where data is or [highlights] what connections exist between data," Bulusu said.

Find it Fast

The system is currently being used to integrate data regarding compound purity verification and drug product stability analysis — information that is siloed in various databases across Pfizer that cannot communicate well with each other. Bulusu said that the IDEA concept would also be applicable to other areas within the company facing similar data integration issues.

As an example, Bulusu explained that one communication hurdle exists between analytical data that is held in a lab information management system and data held in Empower, which is Waters' chromatography data repository.

Prior to IDEA, Pfizer researchers looking for information about compound purity had to conduct a manual search across repositories and individual projects to find data that corresponded to information captured in the LIMS, "but there are no common identifiers," Bulusu said, which means that each repository required its own, separate query.

In the IDEA project, IO and Franz transformed the different types of data via ontologies, brought them together in AllegroGraph, and then used the IO interface and SPARQL queries to find datasets.

Previously, manual data verification took between two to six weeks, but using the semantic approach on a smaller dataset, queries took less than an hour, Bulusu said. While the semantic search did not always deliver exact matches, it gave the Pfizer researchers "real starting points" for their cross-repository search, he said.

In another example, Bulusu said that formulation scientists often have trouble identifying excipients that are a close match a potential new drug because the data repositories are almost "unlinkable" and there is a lack of standardized identifiers. IDEA delivered a "graphical view" of search results with data from different systems, all of which held "enormous value for formulation scientists," he said.

In addition, these projects had "very little infrastructure set up time," he said.

A Growing Market

Franz, based in Oakland, Calif., and founded in 1984, has broad experience in semantic technologies, but "we do not have specific domain experts in the pharmaceutical industry," Jans Aasman, CEO and president of Franz, told BioInform via e-mail. AllegroGraph is used by "a number" of Fortune 500 companies he did not name and there is "significant interest" coming from the US Department of Defense.

Aasman said he believes the semantic technology field is "growing rapidly," and moving toward practical approaches. Whereas last year the company fielded general questions about triple store technology, now "everyone just accepts it as a new technology and people this year are talking about serious applications," he said.

RDF can encode any type of data since all data types are "encoded as strings," which makes it "much, much easier to write ad hoc analysis queries against the combined databases without writing new master schemas or changing the database schemas all the time," he said.

"AllegroGraph is an exception in triple store land because we encode numeric [data] types such as time, dates, values, geospatial locations … as native types in our storage layer." That architecture makes range queries as "as efficient as relational databases," he said.

While Franz hasn't had much of a life science focus, IO specializes in healthcare and life sciences content.

IO's Stanley said that the firm's customers use Sentient to manage second-generation sequence, proteomics, metabolic, chemistry, and imaging data in both open and proprietary formats. He noted that the US Food and Drug Administration's Center for Veterinary Medicine is using the platform to integrate data from multiple studies for biomarker research. Other IO customers include pharmaceutical firms, contract research organizations, and government and academic labs.

Stanley said that the Franz tools "complement IO's tools and experience with different types of life science data and the regulatory compliance environment."

IO's components were directly integrated with AllegroGraph via APIs, he said, which did not involve anything "idiosyncratic." He said that IO has just integrated its software with the new AllegroGraph 4.0 database.

Stanley said he sees semantic projects "picking up steam" in academia, where scientists and clinicians are connecting life science data to medical data, and said that he hopes pharma can move beyond proof-of-concept projects.

He added that he sees opportunities for semantic technology vendors in the pharma market, where many firms are using "patched together open source tools or all-or-nothing pie-in-the-sky plans," which might hold back semantic technologies.

Stanley proposed that these firms instead work with "best of breed vendors who have experience solving real problems in the market."

Don't Stop the Show

Both the Franz and IO methods let users "do pretty much everything in memory, leaving the original data in their original format and location," Stanley said. This approach avoids the "nightmare" of traditional datamarts and warehouses, which require data migration and schema mapping.

Versioning, updates, provenance, and governance are practical issues that need to be addressed but "are not show-stoppers," for semantic technology, he said.

For semantic applications to succeed, researchers must "consider all the available ontologies, vocabularies, and other standards in the life sciences domain" when converting data into RDF, since that step means that the data "is already close to being interoperable," Aasman explained.

Next comes the "non-trivial step of determining the links between the various databases," he said, noting that this step is much easier if researchers create ontologies for the RDF data sources with care.

Bulusu acknowledged that semantic technologies face "a lot of resistance" from the business and IT divisions in pharma, who question the "unproven technology," but said that proof-of-concept projects like IDEA might help alleviate these concerns.

So far, he said that feedback within Pfizer has been positive for the IDEA system.

In addition to the benefits for users, Bulusu emphasized the flexibility and scalability of the approach, and its quick turnaround time for development and deployment. "You can start building a triple store and keep adding to it" with biological, chemical and assay data from many sources — an option that is not possible using traditional data warehousing technology, he said.

"Would we have done this in the same amount of time [with a data warehouse] and have been able to keep adding to that database? No."

Bulusu cautioned that semantic technologies can be slowed if a project requires copying data from source systems and also noted that repeated calls to the underlying sources might impinge on performance. In most cases, however, "you just need a pointing system to point back to the source systems," he said.

He also recommended that pharmas looking into semantic approaches pick their use cases with care.

The "let's build it and they will come [approach] will never work with these semantic technologies," he said. This is particularly true in cases where traditional technologies are a better option, such as projects that require staff to access only one data source periodically.

Bulusu added that semantic technologies are still "a long way" from established programming languages such as Java or .Net, or traditional technologies like relational databases that have been around for years, if not decades. SPARQL's two-year history has not offered much time for performance optimization, he noted.

-->