Semantic data in oil and gas

Semantic data standards already exist in the oil and gas industry which can enable you to integrate different types of data together or answer difficult questions. David Price of TopQuadrant explains how. By David Price, Director of Oil & Gas and Engineering Solutions, TopQuadrant

What if the data you use every day was freed from the shackles of a Microsoft Excel spreadsheet or an Oracle database?

What if it could be published, reached and linked as simply as browsing a website?

What if the apps on your computer, tablet or smartphone had more understanding of what your data actually means?

What if your data could grow and change and be more dynamic?

What if your data is rough and not clean but you’d like to throw it together and see how it looks anyway?

What if you have data needs like these:

Finding existing data, for example “where is the analysis of wellbore 7/4-3 performed last week”

Relating existing data, for example if you have data relating to the Morvin field and Åsgard B platform that exists in different IT systems (and Åsgard B is called “ASB” in one and “Åsg-B” in the other)

Exchanging data, for example “I need to extract the 2009 Kristin volumes, pressures and temperatures and convert to spreadsheet to load into my reporting application”

Integrating data, for example “I’d like to know last month’s production volume total for all fields in which GDF Suez E&P Norge AS is a license”

Analyzing data, for example “Over the past 12 weeks, what’s the trend in barrels of oil per day for Kristin field?”

All this is possible today, right now.

History of data

We’ll start with a little history lesson about data.

Initially, computer data was driven by the application that used it, almost always in a form only understood by that specific application. This was the case for many decades and is still true even today.

However, in the 1980s, databases like Oracle allowed for the separation of data and applications. They supported several users and made it much easier to find needles in the haystack within your data.

To make that possible, everything must be forced into the form of tables and links between tables, which are based on the values of key columns in both tables. This kind of database sits on larger servers in every large organization today.

Scaling down to your personal computer, spreadsheets like Excel are tables in the same way, except with less capability for linking and managing the data.

Spreadsheets add capability for analyzing and visualizing data that help organizations operate. The products that provide these table-based capabilities are hugely successful.

However, your data is still locked away in databases or spreadsheets that are not easily sharable. And forget about extracting from one and getting it into another without help from IT.

World Wide Web

In the early 1990s, people started looking at ways to address some of those issues.

Tim Berners-Lee was working at a European physics institute when he invented the World Wide Web. He invented and implemented his ideas and moved on to make standards of his inventions at the W3C to be shared worldwide.

One of those standards, HTML, is used in every web page you’ve ever visited. HTML supports links between pages on different computers. The Web runs on a network of computers spread around the planet and uses standard means of identifying and communicating across computers.

Your internet Service Provider allows your computer to connect into the internet while your web browser understands the communication and HTML standards while you seamlessly explore the Web by following links between pages.

Throw in a search engine to help you find where to start that exploration, and the web is the basis for an astounding capability that has transformed the world.

While the web has succeeded like few inventions in human history, there is one critical limitation. It operates on the assumption that humans are the audience for the data it presents.

HTML is fine for presenting pages, images, tables and graphs to humans, but cannot define the idea of Person, Computer, Dog, Horse or Oil Rig.

Semantic web

Over time, a vision removing that assumption has been realized.

Tim Berners-Lee and others realized in the late 1990s and early 2000s it was time to allow computers to communicate in the same way that the web has allowed computers and humans to communicate. This vision is called the Semantic Web.

The core of “semantically-aware” web data are simple statements like “Kristin field is operated by Statoil Petroleum AS”.

This thing-property-thing statement is called a ‘triple’.

Instead of thing-property-thing, you can think of node-edge-node and you’ll realize that we are talking about graphs (in the network sense). You can think of logical subject-predicate-object statements, too.

Think of a triple as the equivalent of a single cell in a spreadsheet where the column name is the property, the row identifier is the first thing and the value in the cell is the second thing.

Which ever way you consider it, a triple is very small, in fact it’s very hard to imagine how anything smaller could be the useful “atomic statement” we’d want to manage.

Triplestores

Luckily, there’s a long history of graph theory and database practice that can and has been applied to managing this kind of structure in what are called ‘triplestores’.
Production Operations

The standard underlying triples is called the resource description framework (RDF).

RDF specifies that the things and properties all have web-wide unique identifiers so they can be linked from anywhere, accessed using normal web and internet technology.

It gets into details about various ways to encode RDF graphs of triples in files.

People often use the term RDF database rather than triplestore.

It should be noted that triplestores now scale into billions of triples and commercial ones have the same sort of database management capabilities as something like you’d expect from Oracle - in fact Oracle sell a triplestore.

On top of RDF, a standard query language has been created called SPARQL - equivalent to SQL for relational database folks.

However, unlike SQL, SPARQL queries can span databases spread around the planet, within your organization and can include files sitting on your own personal computer.

Imagine querying over a Norwegian government database, some Statoil Linked Data, Wikipedia, your corporate oil platform management system and a set of spreadsheets sitting on your hard drive all at the same time - that’s what RDF and SPARQL allow.

This let’s you can handle problems like “relating Morvin field and Asgard B platform that exists in different IT systems (and Asgard B is called “ASB” in one and “Asg-B” in the other”).

Adding semantics

At this point, we’ve done little to add meaning to our data. With RDF we get graphs of triples, a little more meaning than an Oracle tables.

The semantics we’re in search of are provided by two languages are, in fact, are just more RDF triples called RDF Schema and OWL.

RDF Schema (RDFS) is a simple language for adding basic meaning to data. Let’s examine our “Kristin field is operated by Statoil Petroleum AS” statement.

Among other things, RDFS allows you to say that “Kristin” is a “Field” and “Statoil Petroleum AS” is a “Company”, which is a kind of “Organization”. It also lets you say that “field is operated by” is a relationship between a “Field” and a “Company”.

Remember that “Kristin” is actually identified by a Web-unique identifier just like a Web page.

So in reality the identifier used is something more like this:

http://factpages.npd.no/factpages/field/1854729

And the concept of Organization might be identified using a W3C standard for organizations as:

http://xmlns.com/foaf/0.1/Organization

The “type” property actually comes from RDF itself and the “subClassOf” property comes from RDF Schema, but as you can see it really is all just more data.

Since “Organization”, “Company”, “Field” and “field is operated by” are all specified as RDF data underneath, these can change easily at any time.

Since it’s all data, merging datasets based on very different sets of concepts also works perfectly well. You can quite easily query over them using SPARQL if you happen to know, for example, that the “SerialNumber” property in on set of data has the same values as the “SupplierIdentifier” property in another set of data.

The Web Ontology Language (OWL) is a more powerful, logic-based language that is an extension to RDF Schema.

As an example, OWL includes the ability to say that a property is transitive.

The following figure shows how we can know that “Wellbore 6507/11-X-4 AH” is part of “Asgard” without saying so explicitly.

OWL also allows one to define classes (i.e. sets) where the members are required to have specific properties, or where anything with specific properties are implicitly members of the set.

There are software tools that make explicit all the implicit data through a mechanism called “inference,” which is nothing more than making more data based the data you specify plus logic statements you’ve made about your data.

The details of OWL are too much for a short journal article. Let’s leave it to say that it is a powerful language, but that you can choose to use some or all of it depending on the complexity of your data.

Revelations

So, what have we revealed about how to free data and give it meaning?

First, that there are very simple, yet powerful, standards that can be applied to the problem. Being standards-based means no vendor lock-in.

Second, those standards ride on top of the Internet and web data based on that infrastructure is easily made available nearly everywhere on the planet and on many kinds of devices. Globally unique names can be given for everything, so at least we always know what we’re talking about – even if we might disagree on what it “is”.

Finally, multiple open source and commercial software tools and database are available from community efforts, and large and small software houses.

These tools scale from desktop files to billions of data elements and the range covers everything in between.

This lets you answer questions like “I’d like to know last month’s production volume total for all fields in which GDF Suez E&P Norge AS is a licensee”.

Of course, other technologies can enable such questions to be answered. However, none are as flexible and extensible or have their basis entirely in Web technology that already exists on every server, computer and smart phone on the planet.

My suggestion is to find out for yourself. If you get stuck or confused, take some training, read some blogs or watch some videos on the topic.

As always, you’re welcome ask me or any of my colleagues at TopQuadrant to explain what’s possible too.