11th International PNEC Data and Information Management Conference[1]
Amsterdam, April 2007

The first EU PNEC Data and Information Management conference was well attended with over 100 delegates. In this short Technology Watch report we summarize presentations that we have not covered previously, or that will be treated in our upcoming report from the 2007 Houston PNEC.

Petris presented work performed for Saudi Aramco leveraging web services and an 'authoritative data store' for the management of large, long-lived assets.

Shell continues to strongly endorse the Flare catalog of upstream terminology – now the focus of the company's data. Allodi (Petroleum Development of Oman) described the search for the 'perfect' search engine and how emerging 'heuristics'-based search might power future, more accurate document tagging and retrieval.

Another paper from Allodi compared traditional classification systems with 'folksonomies,' presenting the results of a PDO survey of user-classified information. The subject of 'to tag or not to tag' was discussed by various speakers. While it is generally recognized that author tagging is great, it is also hard to define a suitable level of annotation detail. It is also extremely hard to get authors to fill out the requisite fields. A PDO internal straw poll showed little enthusiasm for tagging.

On compliance, it was generally agreed that the push has to come from top management and the results monitored as a part of an individual's performance review. Of course, to have effective management support presupposes that management understands the issues. To make sure this is the case, Statoil has been providing management with IM/DM[2] training for several years.

Highlights

Contents

Technology Watch subscription information

This report has been produced as part of The Data Room's Technology Watch reporting service. For more on this subscription-based service please visit the Technology Watch home page or email tw@oilit.com.

[1] *PNEC Conferences are run by Phil Crouse's Petroleum Network Education Conference organization – http://pnecconferences.com.*

[2] *IM – information management. DM – data management.*

TW0707_1 Knowledge management in Shell's E&P R&D – Sushma Bhan & Vic Hitchings, Shell

Described KM in Shell's global R&D unit – in particular how Shell is establishing 'simple, robust and sustainable KM practices'. Researchers are walking encyclopedias but information sharing does not come easily to some of them. Shell's management is working to counter this with a 'work not published is work not done' policy. Components of the KM 'value chain' include R&D portals, technical publishing (OpenText/Livelink), SiteScape knowledge management (see SiteScape/Shell case study on http://www.sitescape.com/clients_partners/pdf/Shell.pdf) and the Shell/Flare E&P Catalog. The R&D Portal includes knowledge bases and databases organized into 'information segments', owned by subject matter experts. The SiteScape-based communities of practice are managed by subject matter experts and global coordinators. Researchers present their work to peers at monthly seminars. Search capability requirements go beyond E&P and leverage Autonomy, MetaCarta, EP Catalog and Invention Machine's 'GoldFire' semantic search engine.

Q&A

How do you get researchers to use these tools?

It is not easy and has to be simple. For example, Livelink didn't seem fit for purpose. Researchers are also evaluated and publication is rewarded.

Younger people are used to social networking and expect collaboration. Older people are now seeing this which is something of a change in Western culture.

TW0707_2 Taxonomies and 'folksonomies' – Alessandro Allodi, PDO[3] (Shell)

All about taxonomies, from hierarchical (Linnaean) through linear (Dewey Decimal) to 'facets' and 'folksonomies.' A facet-based system classifies documents on a multi-branched system. See for instance http://nasataxonomy.jpl.nasa.gov/faq.htm. A folksonomy[4] is an 'open ended labeling system that allows users to classify content.' Content is classified by user-provided tags. The assumption behind folksonomies is that 'if enough people tag an object, interesting patterns will emerge.' Examples from the world wide web include www.flickr.com, and del.icio.us. Folksonomies are free from constraining hierarchies. But there are potential problems with users 'polluting' a corporate taxonomy. A PDO survey of E&P, library, web and administrative personnel found that even a single photograph of a PDM asset could generate a large number of folkonomic tags with various spellings. Moreover there was little enthusiasm for the process – about 70% expressed little or no interest in such spontaneous tagging. Allodi concluded by recognizing the dangers of folksonomists 'confusing catalog structure with personal opinion.' Choosing between traditional taxonomies and folksonomies is not a straightforward decision. The solution is likely a judicious balance of 'traditional hierarchies, facets, tags and folksonomies.'

TW0707_3 Mastering real-time data quality control – Gerhard Thonhauser, Leoben University

Drillers are not interested in analysis, they are interested in real time. This means that data quality control must be done at the source. A well may provide some 12 million data points while drilling with sensor data streaming into the office by satellite. Automating data acquisition and QC avoids costly fixes to historical data and time wasted looking for data. Thonhauser recommends using standards like WITS and WITSML, but cautions that outlier removal and filtering 'cannot be done in real time.' Raw data visualization is of limited application – real time drilling data would require 54 screens! But combining depth and time data and a navigation bar for browsing through data provides drillers with the 20 minutes or so of data that is of interest. Drillers 'are not interested in going back in time.' QC steps have to be auditable and must be done in real time for drilling.

Q&A

Are you capturing metadata at the same time?

We flag data that has been manipulated. You should always tell the user if you change the raw data.

Where does QC happen in the data flow? Close to acquisition point? In service company?

It happens close to source – on site. It has to happen before data transfer. Certain data QC is inherent in WITSML – but the standard is not used consistently. Different companies add their own mnemonics.

[3] *Petroleum Development Oman.*

[4] *http://en.wikipedia.org/wiki/Folksonomy.*

TW0707_4 Pre-stack data and interpretation workflows – Janet Hicks, Landmark

Pre-stack data is a prerequisite for high end interpretation workflows for direct hydrocarbon detection and other sophisticated attribute mapping. But the large data volumes and complex formats involved have discouraged many from fully utilizing pre-stack data. Such perceptions are somewhat different from reality. A seismic survey over the Heidrun field in Norway was delivered to the client in OpenWorks/SeisWorks format as a 60GB dataset (including pre-stack[5]) that could be held on an Apple iPod! Pre-stack requires new data management processes to preserve the data's 'history' as it moves from pre to post-stack. Landmark advocates storing pre-stack data on disk. Apart from the speed of access, this also effectively transfers management of the data to the IT department.

Q&A

Data on disk still means a data management issue.

True but as the data is in now a generic disk file IT involvement requires less domain/format competence.

TW0707_5 Information management for large assets – David Archer, Petris

This paper was motivated by Petris' work for Saudi Aramco. 77% of oil is owned by NOCs[6], 10% by IOCs[7]. NOCs typically work with larger assets over a longer life time. Developing countries likewise have a long life 'attitude' to their assets. Both NOCs and IOCs have to cope with perhaps a terabyte of real time data streaming into their operations daily – often in complex and frequently changing formats. Managing such data requires standard nomenclatures and corporate metadata catalogues. These allow for data to be tracked throughout its lifecycle. A service-oriented architecture and an 'authoritative data store' are considered as 'enabling technologies.' The authoritative data store contains application-independent business object that are accessible through web services. Petris WINDS Enterprise offers an environment where loosely coupled systems, XML and a 'dynamic common model' provide an 'open' data management platform. A 'semantic designer' allows mapping of data object's nomenclature from a variety of data stores to the common 'corporate' data model.

Q&A

How open is the services-oriented architecture?

The Petris SOA platform is vendor neutral. It is open in that the component services are published in WSDL[8] form. Furthermore, we publish the API for data source adapters and encourage others to develop adapters for the platform.

TW0707_6 The quest for the perfect search engine – Alessandro Allodi, Shell (PDO)

Shell's Oman PDO unit produces more information in a day than a person can read in a lifetime. Shell is searching ... for the perfect search engine! The (somewhat intractable) problem involves a balance between recall (ensuring all documents are retrieved) and precision (the relevancy of what is returned). Allodi compared keyword-based and Boolean querying and ranking methods, where precision remains a problem. Tagging and metadata provides an 'excellent though expensive approach.' GIS is great. Automated search as deployed in e-commerce sites uses 'adaptive information filtering.' Google's ranking techniques may be less appropriate in the corporate environment. Enter 'information retrieval[9]' as proposed by Calvin Mooers in the 1940s. On the horizon (two to three years out) we can expect improvement as technology embeds heuristics and 'Shannon's information entropy theory[10].' The heuristic approach is already deployed in MetaCarta's GIS search engine. Heuristics will take another 2 to 3 years to mature and come out of academia. This uses 'TF-IDF' weighting (term frequency, inverse document frequency) to evaluate keyword significance in a document collection. Such mathematical measures are used to classify documents and speed retrieval.

Q&A

Can the heuristic approach determine semantic differences?

MetaCarta can do it for place names. Semantic indexing is powerful. TF-IDF methods are quick to realize.

---

[5] *Probably a subset of the field data.*

[6] *National Oil Companies.*

[7] *International Oil Companies – the majors.*

[8] *Web services description language - http://www.w3.org/TR/wsdl.*

[9] *http://en.wikipedia.org/wiki/Information_retrieval.*

[10] *http://en.wikipedia.org/wiki/Information_entropy.*

TW0707_7 Discussion – Moderators Otis Stelly (Shell), Liv Maeland (Statoil)

0703_7.1 Topic One – legacy information and data

Statoil – We have three data management projects underway. The first, called 'rescuing well data' involves moving old well data from the original media to disk. We now have the ability to manage 20,000 wells on disk. Another project concerns re-mastering seismic field data from old tapes to new media, a three year project involving 600,000 tapes. Business units decide what data is to be re-mastered. A document scanning project involves some 80,000 reports. The value to Statoil is in data and information preservation. Old information and data is better than none. We would like to perform data clean-up during these projects but it is hard to justify the value of what would be a costly extra effort.

Total – We have a project for tapes re-mastering and well data. There is an issue with very old paper logs.

Shell – The question arises as to what to keep. It may be cheaper to keep it all because time and effort is needed to decide. It is probably easier and cheaper to scan everything.

Statoil – We do recognize that a lot has no value.

Shell – In our US unit we need to prioritize. We work on an area of interest, phased approach. 'Market forces' determine what is important and should be prioritized.

Nexen – How do you know what has value? How can you prioritize? No manager will take these decisions! We need to track usage metrics to prioritize these projects.

Wintershall – Priorities may be hard to evaluate. In Germany there are many old wells with leakage issues. We have to keep legacy items because of these problems from old wells. This is a serious issue with around 12,000 legacy wells, sometimes we don't even know who was the operator.

Nexen – In Canada, we have 100,000 old wells and the company has responsibility even if the well has been sold on.

Shell Canada – If you don't index boxes – people can't see what they are, prioritization doesn't work if you don't know what's in the box.

PPDM – Over 90% of documents archived are never retrieved. But lawyers say that you can't get rid of anything! The answer lies in a good retention policy.

Shell – The reserves 'issue' cost Shell around $700 million. That buys you a lot of records management!

ConocoPhillips – Our record retention policy has been signed by the CEO. There has been a major effort. Information will be destroyed according to the retention policy.

Shell – If all our boxes were lined up they would go all around the globe! Document boxes are buried in the desert. We even sold gas stations that we no longer owned because of lost documents.

0703_7.2 Topic Two – assuring compliance

Statoil – Getting users to use standards is a different ball game that requires training, auditing, etc. Vendors need to be pressured into using data management standards. EDM solutions are implementing data models, so vendors should adopt them too! I want my data delivered with standards and quality.

Hess – In Hess data management is a strategy that comes from the top. And it is worth every penny!

Thonhauser – Data auditing is a necessity. Service companies will not do it.

Statoil – It should be included in the contract – it has to be a win/win.

Centrica – Iron Mountain knows what we store – only data for fields where we have equity. In the UK we are obligated to keep physical data.

Shell – Data management follows the same course as asset management. We push on compliance, it goes into people's performance contract.

Shell – in the Microsoft 'wilderness,' the E&P catalog from Flare is what makes our life easy!

Shell – Leadership from the top is a prerequisite. Some data comes in late. With IM and IT people in the team you can work data through 'decision gates.' Compliance is mandatory in Shell.

TW0707_8 Geo-locating unstructured information – Kay Sutter, MetaCarta

The presentation described MetaCarta's geographical lexicon-based search engine that relates place names, oilfield names and other geo located object to their position on a map[11].

Q&A

Are the MetaCarta lexicons related to ontology, the semantic web?

We partner in ontology development but ours is structured around geography.

MetaCarta has a 'flat' world view – what about global 3D or 4D data location?

Yes, we do 'bucket' on surface geography. But once you have located for instance well information, you can layer more specific data sets over the surface location.

What is required to achieve this in terms of data management and quality control?

This is a simple process that starts with deciding who wants to see what data. Security can be an issue – you need to ensure that only entitled people have access to sensitive information.

Are there any international standards for new geographic/lexicon data?

There are no appropriate international standards[12]. It is easier to deploy standards within a company.

How do you improve geo-tagging?

Our geo-tagging technology gets smarter as more names are located in a document. It also recognizes the pertinence of a place name's position in the document.

TW0707_9 Structuring company compliance in data management – Eldar Bjørge, Statoil

Once you have your data strategy in place, how do you assure compliance? Statoil has data owners who are responsible for maintaining data according to corporate standards. 'Easy access' to corporate naming conventions for stratigraphy, log curves, horizon names etc. simplifies the data manager's job. Application vendors have little focus on embedding data management in their products. Statoil has fixed this with a data management 'toolbox' containing standards as above and also best practices and workflows for data loading, quality assurance and QC. Training of both staff and management in data management makes sure that all sing from the same hymn sheet. Data 'self-help' checklists measure compliance and offer advice on everyday data issues. Data managers' careers are as closely managed as G&G and it is possible to rotate through disciplines. Of Statoil's 1,000 G&G staff, some 30 are data controllers and 15 are located in the central data management group.

Q&A

You actually are training management?

We have been for the last five years.

TW0707_10 Technology Watch subscription information

This report has been produced as part of The Data Room's Technology Watch reporting service. For more on this subscription-based service please visit the Technology Watch home page or email tw@oilit.com.

The Data Room
7 rue des Verrieres
F-92310 Sevres France

Tel (USA)	281 968 0752
Tel (UK)	020 7193 1489
Tel (France)	+33 1 4623 9596
Fax	+33 1 4623 0652

info@oilit.com

[11] *MetaCarta's technology has been presented in several previous Technology Watch reports. For more information use the search feature on www.oilit.com.*

[12] *Actually there may be. See the Open Geospatial Organization's Gazetteer Service - Application Profile of the Web Feature Service Implementation Specification at http://portal.opengeospatial.org/files/?artifact_id=15529.*