**Advanced Knowledge Discovery Tools: Designed to Unlock and Leverage Prior Knowledge to drive Cost Reduction, Increase Innovation and Mitigate Risk**

Ian F Mitchell BSc CEng MIMechE(1)

1 IHS Markit , Bracknell, United Kingdom

E-mail contact of main author: Ian.Mitchell@ihsmarkit.com

**Abstract**

The nuclear industry is knowledge-intensive and includes a diverse number of stakeholders. Much of this knowledge is at risk as engineers, technicians and project professionals retire, leaving a widening skills and information gap.

This knowledge is critical in an increasingly complex environment with information from past projects often buried in decades-old, non-integrated systems enterprise. Engineers can spend 40% or more of their time searching for answers across the enterprise instead of solving problems.

The inability to access trusted industry knowledge results in increased risk and expense.

Advanced knowledge discovery technologies slash research times by as much as 75% and accelerate innovation and problem solving by giving technical professionals access to the information they need, in the context of the problems they are trying to solve. Unlike traditional knowledge management approaches, knowledge discovery tools powered by semantic search technologies are adept at uncovering answers in unstructured data and require no tagging, organization or moving of data, meaning a smaller IT footprint and faster time-to-knowledge.

This session will highlight best-in-class knowledge discovery technologies, content, and strategies to give nuclear industry organizations the ability to leverage the corpus of enterprise knowledge into the future.

1. **Search vs Research**

Not only the nuclear industry but many industries face the same challenges today around how to manage and effectively work with legacy data buried in silos most of which have been forgotten about or are not even known to exist by the present employees.

This problem is regularly seen to mean considerable time wasting as information cannot be found leading to duplicate work and even reinvention. A senior manager at NASA has said “If we want to go back to the moon again we’ll be starting from scratch because all the knowledge has disappeared. It would take as long and cost at least as much again to go back”.

This is a perfect example of the Knowledge gap or loss of critical information when employees leave their employer. In the case of one company in the UK a team of retired engineers, all over 75, has been bought back to work on a project to help with the uncovering of critical information and another facing the situation of 20% of their workforce eligible for retirement within the next 12 months.

With such scenarios occurring every day in business projects are delayed and overspend is inevitable as effort goes into discovery of the data and not the analysis (Fig1)



Figure 1 Time wasted data gathering and not on Analysis

As companies seek to access their internal information more effectively keyword searching is not adequate as far too many results are delivered with not relevant answers.

For innovation and problem solving activities the requirement is an understanding of the context of a user’s query- their design intent and need. Keyword searches, advanced Boolean searches and statistically based search methods retrieve piles of disconnected documents, not the precise results engineers and other technical professionals seek.

A common complaint is that it is easier to find information outside of the organisation than within and with Enterprise search applications designed to find documents and objects so long as you know what you are looking for and you also know where to look.

With the decommissioning programme underway in the UK it is not only now that this information has to be accessible it is through the next 70-80 years as the Post Operational Cleanout (POCO) is completed that access to this information will become critical.

Document management systems have typically been in place for many years are being pushed beyond their limits becoming slow and difficult to search. Content stores do a great job of capturing corporate knowledge but do not enable engineers and other technical professionals to search ideas and related concepts across that content and other broader data sources. Many also require additional work such as tagging or classification before they can be added to the repository. There are also volumes of scanned documents and note stored on File shares.

1. **Nouns, Verbs and Semantics**

Being able to reuse corporate knowledge in a more effective way helps technical professionals improve the quality of their problem solving, reduce research time and avoid duplication.

The most recent Advanced Research and Knowledge Discovery tools federate search to enable rapid discovery and retrieval of key intelligence based on meaning- extracting relevant solutions, concepts and ideas from design documents, voice of the customer surveys, product manuals, field reports, warranty claims, material specs, proposals and other content housed in documents in an enterprise management system, email, shared drives or other corporate data stores.

Patented Question Answering Technology enables the uncovering of answers and concepts across multiple document types and formats, location, languages both residing inside a company firewall or outside it.

In traditional technology when a sentence is read there is no understanding of relationships between the words of the sentences and paragraphs hence, the reason for many irrelevant results that require considerable effort to sift through.

The Question Answering Technology approaches sentences, paragraphs and documents in a completely unique way analysisng the language at a number of different levels to create an understanding of the semantic relationships between the verbs and nouns in the sentences. Just for English language documents there are more than 105 patterns, rules and grammatical configurations that are used to analyse and create the index.

Using these rules and patterns also means that verb phrases, noun phrases and Subject Action Object phrases can be recognized giving the researcher a whole new opportunity for finding concepts and ideas from just one search query.

If a simple sentence is analysed the clear difference between keyword search and semantic search can be identified. (Fig 2)



Figure 2 A Simple Sentence for Analysis



Figure 3 Unconnected words

Fundamentally, keyword search is a natural extension to the analogous technology it replaced – the library card catalog – and easier to commercialize. As such, it is also the technology that many enterprise systems such as Product Life Management (PLM) chose to implement as their users find it increasingly difficult to navigate the vast amount of electronic information being amassed.

With any keyword search technology you get a data dump – and it’s up to the user to navigate that dump! The goal is to locate documents and not give answers. The keyword may be contained in the document, but you still need to read it to understand the context and decide whether it offers a solution or is not relevant or worse (Fig3).

Returning to the previous paragraph used to illustrate keyword search, it becomes clear the power of a semantic engine. Unlike keyword search engines which merely catalog words, this semantic engine indexes and programmatically “reads” the content of documents – understanding and cataloging the 105 semantic relationships within language, becoming an expert just as humans do.



Figure 4 Subject Action Object relationships as understood by the Semantic engine

With these semantic relationships (Fig4) now understood, it can then begin to intelligent assessments of the content such as Subject Action Object relationships, whole consist of parts relationships and cause produces effect relationships.

This Question Answering technology is specifically and purposefully tuned for delivering precise answers to the question engineers, scientists, and those in product development ask – and sometimes even the answers to questions they don’t know to ask.

**Identifying meaning, Creating Knowledge**



Figure The indexing process

For the Semantic engine to produce this information the meaning of each sentence that it is asked to work with has to be analysed and understood (Fig5). Some 400 document types can be read for indexing and there are connectors for the most common Document management systems, Feeds, and Databases.

One of the biggest challenges that companies face is scanned documents which are unsearchable. Being able to transfer this material into searchable content and be able to recognise such material when crawling through file shares and drives during the indexing process is critical in the nuclear industry where many builds were completed long before electronic storage was common place. Technology is incorporated and can work with third party OCR software so that no pre-work is required to select then convert and index this material.

The indexing process reads each document to analyse the words and parts of speech before looking at the sentence structure and the Grammar that has been used. From this the sentence meaning is captured, concepts extracted and a machine generated dynamic document summary (Fig6) is produced.

This functionality is performed across not only English but also French, German, Japanese, Mandarin and in the near future Russian.

All of these extracted concepts are stored in a semantic index which is accessible corporate wide, group wide or at a personal level depending on the content and the required security limitations. Whilst the index is the content that a user searches the original document remains in its storage silo and is not copied or moved to a new location.



Figure 6 Machine Produced Document Summarisation

This technology is not only applied to internal knowledge it is used to work on external content such as Patents and articles from trusted partners giving the user access to authoritative content through a single portal for a comprehensive search or limited focused search as necessary.

**Running a Search on Indexed Content**

With the extracted concepts in the form of Subject-Action-Object configuration it becomes possible for the user to pose Natural Language Questions i.e. How to separate gas from water? The semantic engine, understanding verb noun relationships, can recognize the difference if a new question is submitted i.e. How to separate water from gas?

The questions themselves are subjected to an analysis from which a search pattern is built that is applied to the Semantic index. The search pattern is also translated into other languages to enable cross language searching. Answers are displayed in three ways, Meta data, Relevancy as in a list with the most relevant result at the top and then in a series of Concept lenses breaking the subject into five areas of knowledge.

To ensure querying and results are as thorough as possible synonyms for verbs and nouns are automatically used. Both these sets can be further enhanced, where appropriate, by the user adding company Jargon for working with internal content.

Based on submitted questions or phrases the Question Answering technology, through the semantic analysis, automatically processes a set of questions on behalf of the Researcher to retrieve all that is known about the concept (Fig 7). The discovered results are then sorted into approximately 50 categories across six screens (Fig8) for ease of finding relevant concepts.

Each concept lens allows the researcher to rapidly understand most commonly discussed subject against that lens topic.



Figure 7 Display of User Interface with the three separate result areas



Figure 8 The basis of questions behind each lens in the General Facts screen

For a result to appear at the top of the list of results it must meet the criteria of an algorithm which decides on relevancy of the uncovered information. For a result to appear in a lens it must exactly match the semantic relationship between the words being posed by the question that is behind the lens.

As an example

Galvanic corrosion – is submitted as the query

Taking the first lens in the General Facts Screen is for Definitions

What is galvanic corrosion?

Results are then presented in descending order of frequency

i.e. electrochemical process (95)

This is a strong answer as when selecting the link

Galvanic corrosion is an electrochemical process in which one metal corrodes preferentially to another when dissimilar metals with different corrosion potentials (Ecorr) are in electrical contact and immersed in an electrolyte.

Or this

Most of all, galvaniccorrosion for coupled components under different atmosphere conditions (including damp- heat, pollution and marine) is a electrochemical process.

Both of these sentences are a true semantic match to the query that was submitted.

It is equally possible for many results to be returned in the Relevancy list but at the same time few results appear in the lenses. This indicates that there is a limited number of results which semantically match the submitted query yet each of the results contain the words of the query.

**What do I do with the Search Results?**

Having a page of results can be daunting, even when broken into the Research areas of meta-data, results lists and lenses. However, all three sources can be used to drill down into the areas of interest to enable focused research as quickly as possible.

Filtering by metadata enables zooming in to different document silos, corporate knowledge, patents, standards, books, Date ranges, authors, publications, publishers. (Fig 9)

 

Figure 9 Some of the filtering options using Metadata

Filtering by using the results from the lenses is another method of drilling into the content from possibly thousands of results Fig 10 to just a handful (Fig11)

 

 

 Figure 10 Submitted Query & Results



Figure 11 Typical Results from lenses

Selecting just one link presents those results giving access to the documents, highlighting of sentences, short & dynamic summaries and the option to save for later review (Fig 12)

 

Figure 12 Document links to the original papers

When a paper is selected for reading the Researcher is taken directly to the sentence presented in the user interface and highlighted for ease of finding. (Fig13) 

Figure 13 Highlighted Text landing the user in the appropriate place in the document

**Conclusion**

The Nuclear industry faces some huge challenges into the future trying to utilise years of valuable knowledge and information much of which is still stored in difficult to access repositories and in some instances not even known about.

The Semantic indexing and retrieval process presented in this paper, already being used by a number of organisations around the world, has been proven to reduce research time by up to 30% in organisations that have configured and deployed the solution in the Enterprise.

Not having to do any pre-work to content before indexing means that repositories that have been sitting unused due to lack of knowledge or the difficulties of access can be turned from information to actionable knowledge in a matter of hours with case studies of user’s experiences showing a high return on investment in a relatively short space of time.

Data gathering can be considerably reduced as multiple repositories can be accessed through a single portal simultaneously. This in turn enables the knowledge worker to dedicate the required amounts of time to the analysis of the gathered data rather than the tedious task of trying to find and gather the data often failing and duplicating previous activities or at worse giving up altogether.

**References**

**Semantic TRI**Z Mikhail Verbitsky, Ph.D., Sc.D. Invention Machine Corporation 2004 internal document

**Semantic Research Best Practices** Benoit Simonet Power to Innovate users meeting 2009

**Leveraging and Amplifying Technical Knowledge Scattered Across the Enterprise**- White paper IHS Engineering Solutions

**Question-answering system and method based on semantic labelling of text documents and user questions** Invention Machine Patent US 8666730B2 March 14