Use Case 1 in Energy - Research

Geosciences project Memory

KW Partner: IFP

1 Overview

Challenge
In a KM context to give semantic access to multiple project documents and data (software, subsurface models) to practitioners
Solution
Automated on the fly semantic annotation of documents produced by the projects
Semantic search engine for ontology based queries
Automated summary capability
Why a Semantic solution
The present search engine technology is not efficient enough on full text search
Key Business Benefits
Avoid duplication of work and ensure quality of work in future projects by access to the best available information from the organizational memory (Capitalization and knowledge sharing)
Business Partners
IFP together with its projects partners (energy industry)

Keys components

Existing Software
Documentum, Verity full text search engine
Company reference taxonomy

Research and development
Automated Semantic annotation
Ontology driven user friendly query
Semantic search engine
Semantic data integration
Natural Language Processing

Technology locks
Large domain ontologies building and maintenance
Automated on the fly annotation
Automated Summary

Figure 1 – Knowledge management system

This business case deals with the management of large collection of project documents in the field of geosciences. IFP is involved in several such projects with a number of partners in the energy industry. These projects deliver various kinds of documents, in the form of texts, geological maps, software, subsurface models, data bases, etc. (Figure 1)

Our goal is to access the documents produced by these projects in a structured manner, so that a new project can make best use of the results produced by previous projects.

We believe that semantic annotation together with semantic search engines can be of great help in this context. The usage scenario is the following:

- Before or during the project, a domain ontology is developed either from scratch or by re-using existing relevant ontologies.
- During the project, or at the end of the project, the documents produced are annotated with semantic markers from the ontology. This annotation must better take place "on the fly", that is, while producing the document or just after its finalisation. It must be automated as much as possible (e.g. 95% automated) so that the project contributors do not spend too much time in annotation.

- It should be emphasised that we intend to annotate not only textual documents, but other project documents as well i.e. software (description of input/output and internal processes), models (description of models variables, parameters, geo-localisation, hypothesis, usage mode…), data bases (structure of database, tables, fields…).
- After the project, a new user needing to retrieve information about a specific domain from past project work would query the geosciences projects semantic memory to find relevant documents. Using the domain ontology, navigating, selecting objects of interest, and making the query should guide the query.

2 Current Practices and Technologies

2.1 Current business practices

Our current practice relies on a document management system, Documentum, and the Verity full text search engine. Our project memories are built by storing the documents in the knowledge capture and sharing system, annotating them with metadata about the document types, authors, project information, confidentiality level and other coarse grain annotations.

Storing documents in the corporate project memory system is an integral part of or Quality Assurance process related to the handling of project information.

There are no ways to retrieve documents in an "intelligent" way, as queries can only match words selected by the user. In addition, interoperation of different corporate memories is not possible or very difficult because of different referencing systems.

2.2 System requirements Analysis

These Use Cases make the following problematic stand out:

1. Building and maintaining large domain ontologies in complex evolving technical domains
2. On the fly annotation of documents of various kinds, using the domain ontologies. This automation needs to be highly automated as users will not want to spend too much time on this subject. Typically, 95% of the documents should be automatically annotated, even leaving 5% of this task to users would be quite challenging for them.
3. User friendly ontology driven query. When preparing a query to the projects memory, the user will navigate among ontology elements. This navigation needs to be user friendly, as there is no chance that an average user will want to look at the internals of the ontologies. It should be possible to find the relevant concepts and relations in an easy and quick way.
4. Semantic search engine. This is a familiar problem but the semantic search should scale up well with the amount of project data and of possible queries i.e. a system like Google brings meaningful results in a few seconds, this is probably the target for a semantic search engine on large collections of documents.
5. Semantic data integration between different kinds of documents. It should be possible to retrieve a text about a map, together with the map, and present together with the links between them taken into account.

6. Defining quality assurance procedures for semantic annotation and retrieval: if we want to base our QA process on this new framework we need to demonstrate that the properties are at least better than current practice.

2.3 Review of the current systems

There exist current solutions like: Goldfire Innovator http://invention-machine.com, specialised in access to technical solutions by looking for subject-verb-object triplets in patents, or Arisem (Thales) http://www.arisem.com , based on semantic networks.