



## The Company

**Snamprogetti** is a company operating under the ENI Group, the Italian integrated energy company, one of the largest in the world. ENI operates in the oil and natural gas, electricity generation, engineering and construction sectors, and in the petrochemical business. **Snamprogetti** is a technology-oriented global contractor.

The Company operates as an international main contractor for the design and implementation of large-scale plants for the production, treatment and transportation of hydrocarbons; the monetization of natural gas (liquefaction and conversion); the conversion and upgrading of conventional and unconventional crude oils; chemicals; they are also involved in infrastructures and the protection of the environment.

To achieve its objectives, Snamprogetti uses technology as one of its key leverages against competition.

## The need

**SnamProgetti** needed to implement a **DataDiscovery System** to support the users to extract specific documents from a huge repository of files.

The system must operate against massive volumes, to keep track of ongoing activities and to create "dossiers" (collections of documents ) about specific topics .

## The solution

The DataDiscovery system has a modular architecture built on **VoloFrame's** flexible middleware. This is necessary in order to process large amounts of data across the network: the system must analyse million of files and tens of terabytes of data, such operations imply long execution times. Therefore, the system is designed as a set of asynchronous, independent modules that can be configured and scheduled for an unsupervised batch execution.

Each of the available modules is tailored to perform a specific task: Loader module scans the file systems and builds a list of files to be processed; Parser module processes the files found by the Loader module, analyses their content and, if applicable, stores them in the DataDiscovery store; a set of modules are available to import and index documents into the conceptual search engine. An additional module (DossierExtractor) is available to automatically produce a dossier of documents.

Each of these modules relies on the common database to keep track of ongoing activity and to record each step of the process. Database queries are available to monitor the state of the process and to resolve any conflicts. The following file formats are used in SnamProgetti: Text, HTML, Word, RTF, Power Point, Excel, Adobe Acrobat, Outlook PST, ZIP archives.

The DataDiscovery system has two different processes to extract relevant documents:

- a. The "dossier" mode can be used to automatically produce a selection of documents based on a list of Boolean queries.
- b. The "**natural language search**" mode can be used interactively to scout the knowledge base for documents that may be relevant to a specific topic.

The system integrates the powerful **conceptSearching** Information Retrieval engine that enables users to query the repository in natural language.

The results of the query can be ordered by relevance to the original query. For each result, a link to the local copy of the document and a link to the original location are given. Furthermore, the conceptual engine can produce a list of documents that are similar to the selected document and/or a list of topics (i.e.: related query terms) that can be used to further enhance the search results.