cubist
Your Business Intelligence

SEVENTH FRAMEWORK PROGRAMME

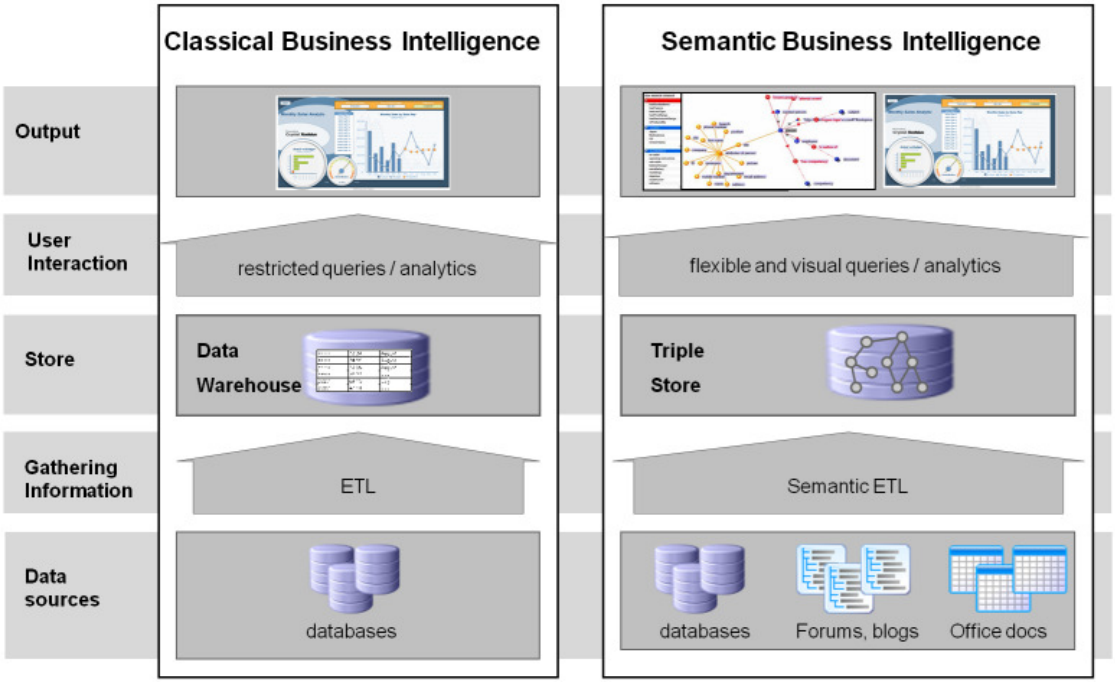
Combining and Uniting Business Intelligence with Semantic Technologies

Main Objectives

The constantly growing amounts of data and an emerging trend of incorporating unstructured data into analytics is bringing new challenges to Business Intelligence (BI). Contemporary BI solutions fall short in the following aspects: Firstly, they focus only on structured data and disregard the increasing amount of information hidden in unstructured data. Secondly, BI users are dealing with increasingly complex analyses, but the complexity of BI tools becomes the biggest barrier for their success.

CUBIST is an EU funded research project with a visionary approach that leverages BI to a new level of precise, meaningful and user-friendly analytics of data by following a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and Visual Analytics. CUBIST aims to

- support federation of data from unstructured and structured sources
- persist the federated data in an Information Warehouse; an approach based on a BI enabled triple store
- provide novel ways of applying Visual Analytics based on meaningful diagrammatic representations



Use Cases

The research results will be demonstrated in three innovative use cases:

Biomedical informatics: CUBIST will combine and explore the semantic representation of spatial temporal biomedical data from two biomedical atlases and three gene expression databases.

Control centre operations: In mission control rooms in space control centres very large volumes of data are obtained from heterogeneous sources, including structured and unstructured data. CUBIST will provide online support for taking better decisions, reveal hitherto undiscovered information and provide supportive evidence in debriefing and decision making processes related to the organisation of space control centre operations.

At a glance

Partners

SAP AG
Ontotext
Sheffield Hallam University
Centrale Rechereche S.A.
Heriot-Watt University
Space Applications Services
Innovantage

Core Information

Call: FP7-ICT-2009-5
Objective: 4.3: Intelligent Information Management
Project ID: CUBIST-257403
Funding Scheme: Strep
Duration: 10/2010-09/2013
Total Cost: € 4.357.834
EC Contribution: € 3.029.193

Project Coordinator

Dr. Frithjof Dau
SAP AG
SAP Research Center Dresden
Tel: +49 (0)351 4811 6152
frithjof.dau[at]sap.com

SAP
ontotext
Sheffield Hallam University
CENTRALE RECHERCHE S.A.
spaceapplications SERVICES
Innovantage ADVANTAGE THROUGH INNOVATION
HERIOT WATT UNIVERSITY

Market and competitive intelligence: This is a job market use case which combines information from job advertisements crawled by CUBIST and an existing firmographic database. CUBIST will enable market intelligence (insights about who is recruiting, and where and when and how they recruit) and competitive intelligence to help employers track and better understand the recruitment activity of their competitors.

Research Challenges and Expected Impact

CUBIST is based on Semantic Technologies, particularly on RDF for data representation and triple stores for federating and persisting the data, and Formal Concept Analysis (FCA) for conceptually clustering the data and organizing the clusters into hierarchical relationships. These clusters will be used for the Visual Analytics. This leads to the following challenges and expected outcomes:

Semantic Extract, Transform and Load (ETL): Source data in unstructured sources might be noisy, inconsistent and incomplete. Extracted data from unstructured sources has to be brought into relationship with extracted data from structured sources. CUBIST targets semantically enriched lineage information, error detection and identity resolution within extracted data, and a semantic ETL component that provides SPARQL (the query language for triple stores) endpoints for various data sources.
Query Language: SPARQL lacks complex aggregate functionality, reporting functions and rollup/cube expressivity. In alignment with the efforts of W3C, CUBIST will extend SPARQL by needed OLAP functionalities.
Performance and scalability of the Triple Store: Current state-of-the-art implementations of triple stores are for tens of billions of triples, a magnitude lower than the data volumes for a state-of-the-art data warehouse. CUBIST will significantly multiply the number of triples the triple store can deal with.
FCA and Triple Stores: Most FCA applications are stand-alone solutions. In CUBIST, a layer within the warehouse will integrate the triple store with the FCA-based visual analytics.
Scalability of FCA: Existing FCA solutions do not scale to large amounts of data. CUBIST will investigate high-performance FCA algorithms and tools, including parallel processing algorithms for multi-core architectures.
Visual Analytics: Current FCA visualization tools have been designed without very large data sets in mind. Interlinked with best practices from known BI visualizations, CUBIST will scrutinize novel approaches for FCA-based visualisations which allow for in depicting, navigating through and visually querying the data.

To summarize, for the core fields Semantic Technologies, Business Intelligence and Visual Analytics, we expect the following impact:

Semantic Technologies: CUBIST aims to bring Semantic Technologies to a level where they can be successfully applied in industrial settings using huge data sets, comparable to established technologies such as relational databases and BI.
Business Intelligence: It is expected that incorporating unstructured data will be very important for future BI systems. CUBIST takes an essential step in this direction. From a technological perspective, CUBIST will have an impact on the architecture of future BI systems. From a business perspective, CUBIST will help overcome the barrier of complexity of BI tools and interfaces and apply BI functionalities to new business scenarios and new user groups.
Visual Analytics: From a technological perspective, using FCA for representing and navigating the large amounts of data will open up FCA to new horizons in terms of applicability in real business settings. From a user perspective, using FCA has a solid mathematical foundation and a close link to the human perception of concepts, thus FCA will drive Visual Analytics to a new level of theoretically precise and humanly comprehensible visualizations.