# PROJECT FINAL REPORT


**Grant Agreement number: 257403**

**Project acronym: CUBIST**

**Project title: : Combining and Uniting Business Intelligence and Semantic Technologies**

**Funding Scheme: STREP**

**Date of latest version of Annex I against which the assessment will be made:** …

| | |
|---|---|
| **Periodic report:** | **1st** ☐ **2nd** ☐ **3rd** ☒ **4th** ☐ |
| **Period covered:** | **from** Oct. 1 2012 **to** Sept. 30 2013 |
| **Version** | Version 1 |
| **Date** | 31.10.2013 |

**Name, title and organisation of the scientific representative of the project's coordinator:**

**Frithjof Dau, PhD, SAP AG**

**Tel: +49 351 4811 6152**

**Fax: +49 6227 78-51425**

**E-mail: frithjof.dau@sap.com**

**Project website address: www.cubist-project.eu**

# Table of Contents

# 1. Publishable summary

## *Project Context and Objectives*

Constantly growing amounts of data, complicated and rapidly changing economic interactions, and an emerging trend of incorporating unstructured data into analytics, is bringing new challenges to Business Intelligence (BI). Contemporary solutions involve BI users dealing with increasingly complex analyses. According to a 2008 study by Information Week, the complexity of BI tools and their interfaces is becoming the biggest barrier to success for these systems. Moreover, classical BI solutions have, so far, neglected the meaning of data, which can limit the completeness of analysis and make it difficult, for example, to remove redundant data from federated sources.

Semantic Technologies, however, focus on the meaning of data and are capable of dealing with both unstructured and structured data. Having the meaning of data and a sound reasoning mechanism in place, a user can be better guided during an analysis. For example, a piece of information can be semantically explained or a new relevant fact can brought to the user's attention. In particular, we foresee a well-known semantic technique called Formal Concept Analysis (FCA) to be a key element of new hybrid BI system. Depending on relationships between different entities, FCA allows to compute meaningful, hierarchically ordered clusters in the data, which can be visualized. Thus FCA provides a means to qualitative data analysis, complementing traditional BI analysis which is of quantitative nature.

The CUBIST project develops methodologies and a platform that combines essential features of Semantic Technologies and BI. We envision a system with the following core features:

- Support for the federation of data from a variety of unstructured and structured sources.
- A data persistency layer based on a BI enabled triple store, thus CUBIST enables a user to perform BI operations over semantic data.
- Advanced mining techniques of Formal Concept Analysis (FCA). FCA guides the user in performing BI and helps the user discover facts not expressed explicitly by the warehouse model.
- Novel ways of applying visual analytics in which meaningful diagrammatic representations will be used for depicting the data, navigating through the data and for visually querying the data.

CUBIST demonstrates the resulting technology stack in the fields of market intelligence, computational biology and the field of control centre operations.

Information about CUBIST can be found on the project website: www.cubist-project.eu.

### Requirement Analysis

The first phase of the project (six months) had been mainly dedicated to the requirement analysis. This analysis has been conducted in close collaboration with the use case partners and their respective work packages. In order to guide the use case partners in the creation of requirements, two workshops have been conducted.

The following means have been used for the requirement analysis:

- *Personas* help to identify and describe different prototypical end users of the envisioned CUBIST system, including data about their profession, skills, goals, and even attitude towards CUBIST-relevant aspects of computer systems.
- *Utilization scenarios* represent typical days of the personas, described in a story-like manner. They come in two forms: An ”as-is”-scenario which describes the days in the life of a persona without a CUBIST system, and an envisioned ”to-be”-scenario which reiterates the ”as-is”-scenario under the assumption that a CUBIST system is in place.
- *Mockups* are envisioned user interfaces with an emphasis on the conceptual design of the software (and not on the design of the UI).
- *Formal requirements* finally specify (atomic) requirements in a formal manner.

In addition to the general requirement analysis, a dedicated analysis has been carried out for the analytics and visualization capabilities of CUBIST. The actual visualisation requirements have been inferred directly from the general requirements.

## Architecture and Software Components of CUBIST

In the first year of the project the design of the overall architecture for CUBIST has been started, which includes the identification of main functional blocks and core services as well as the definition of interfaces between components. Moreover, adaptions for existing software components have been started for CUBIST, as well as new software components have been started to be developed. In the second year, the components have been further adapted and extended, and they have been integrated into one overall general CUBIST prototype. Moreover, in the first quarter of 2012, dedicated effort was spent amongst the consortium to refine the architecture.

In the following, core software components, developed within the CUBIST, are described.

**OWLIM (Ontotext)** is a highly scalable triple store developed by Ontotext. It is implemented in Java, it is Sesame API compliant and supports RDFS and specific OWL profiles. OWLIM will serve as persistency layer for CUBIST.

**NowaSearch Front-end and Search Service (SAP)** is a web-based research prototype for semantic information integration and search with faceted search features. It is the outcome of a former research project at SAP (Aletheia) and adapted for CUBIST. This prototype enables a factual search (based on keywords or based on semantically enriched information), faceted search and graph exploration for the information stored in the semantic layer. It serves now as the basis for the CUBIST integrated prototype: other components are integrated into this prototype. NowaSearch is adapted for CUBIST in order to meet the information access needs of a BI application.

Two screenshots, based on data of the HUW use case, are given in Fig 1. NowaSearch and Graph Exploration are currently undergoing an SAP-internal process to be published as open source.

**FCAService (SHU):** FCAService is a component developed by SHU that provides Web Services for converting results of SPARQL-queries into formal contexts. This is done via clustering the values of an attribute into so-called “bins”. For different attribute types, it allows a fine-grained tunable conversion: It provides different methods on computing or manually entering the borders of the bins, or it can be set whether bins are disjoint or whether they can include each other.

A screenshot on how the parameters in CUBIST for the binning process are set is provided in Fig 2. FCAService has been published on GitHub under the Apache 2.0 license as open source.

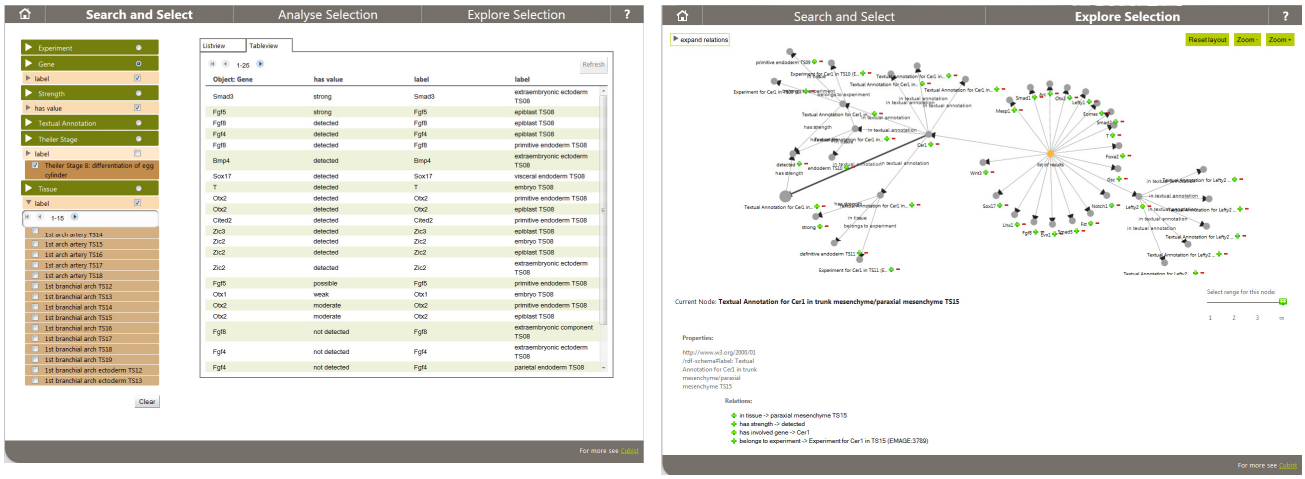


**Fig 1: Faceted Search and graph exploration in NowaSearch**

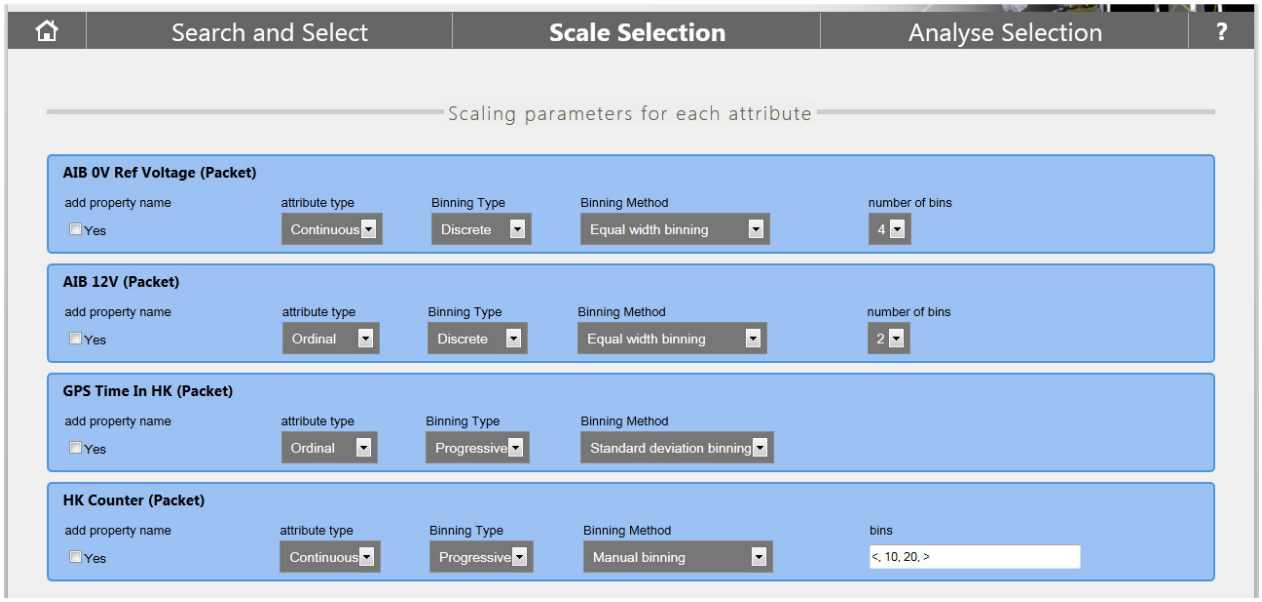


**Fig 2: Conceptual Scaling with FCAService in CUBIST**

**CUBIX (CRSA)** is a standalone, frontend based FCA visualization / analysis tool. The tool is being developed in CUBIST from scratch and continuously refined with the actively involvement of our three users groups and their use cases. A first version of the tool, which has been developed in the starting phase of CUBIST, has been based on flash and the flare visualization library. During the course of project it became evident that flash becomes an obsolete technology and will be superseded by HTML5, CUBIX has been redeveloped in HTML5, mainly using d3.js as visualization library.

CUBIX already provides novel features for visualizing large concept lattices (e.g. the transformation of lattices into trees, see **Error! Reference source not found.** for examples) and implements the gathered visualisation requirements. Typical uses of CUBIX include semantic data

analysis and pattern detection, anomaly detection, comparisons, information classification, and knowledge discovery. As of the end of the second year of CUBIST, CUBIX provides numerous visualizations for the lattices (Hasse diagrams, sunburst diagrams, trees, treemaps, icicles, scatter plots) as well as bar charts which show the distribution of analysed entities, as well as searching, selecting and filtering capabilities for the visualizations.

CUBIX has been published on GitHub under the Apache 2.0 license as open source.

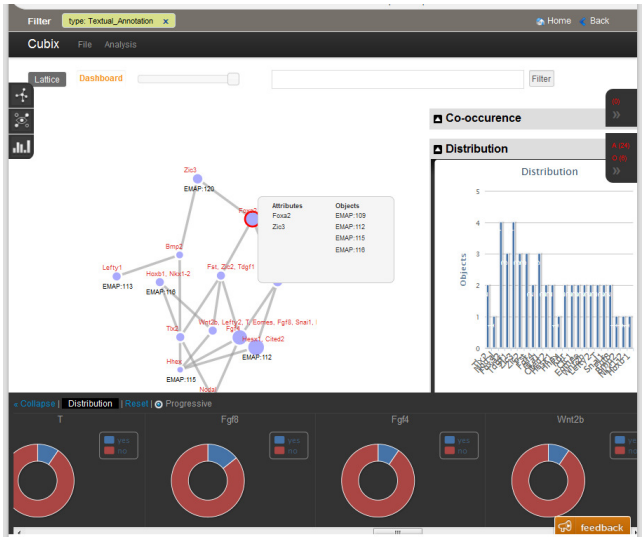


**Fig 3: CUBIX prototype for VA-frontend**

## CUBIST Workshop and Special Journal Edition, Public Dissemination

CUBIST has set up its own scientific workshop, which is annually conducted and collocated with appropriate scientific conferences. The first workshop was collocated with the 19th International Conference on Conceptual Structures (ICCS), 25-29 July 2011, University of Derby, UK. The second workshop was held in conjunction with the 10th International Conference on Formal Concept Analysis (ICFCA), 6 - 10 May 2012, Leuven, Belgium. The 3rd CUBIST workshop (CUBIST-WS 2013) was conducted in conjunction with the 11th International Conference on Formal Concept Analysis, ICFCA 2013, 21-24 May 2013, Dresden, Germany. The workshop is dedicated to topics related to CUBIST, but not to participation of CUBIST members only. In fact, in all three workshops, we have received submission from outside the consortium.

A special CUBIST edition of the International Journal of Intelligent Information Technologies[1] is currently in print; the authors of the best publications of the CUBIST workshops have been invited to submit extended versions of their papers. Moreover, a set of CUBIST related papers have been published in the International Journal of Conceptual Structures and Smart Applications (IJCSSA).

Finally, the consortium has pushed information to the public. Most importantly are various demo videos on the CUBIST youtube channel (http://www.youtube.com/user/CUBISTFP7ICT) and information about technologies and functionalities of the prototype provided in the CUBIST external Wiki (http://wiki.cubist-project.eu/).

[1] http://www.igi-global.com/ijiit

## Use Cases

In all use cases, during the reporting period, the following efforts have been carried out:

- All use case partners participated in the requirement analysis.
- The data sources in the use cases have been sufficiently detailed described in order to start with the semantic ETL-process within CUBIST.
- For each use case the development of a business ontology, which will serve as the underlying schema for the analytic features of CUBIST, has started.
- Each use case partner has provided data sets, which have been federated into the repository.
- All use case partners have provided natural language information needs and queries, which first have been converted into formal analytics directly, but –more importantly- have been deed into the re-design of the UI, enabling "BI as a self-service".
- For each use case, the general prototype has been customized to the respective use case.

In the following, we provide more details for the respective use cases.

**HWU:** HWU lead WP7, the biological use case. Work progresses in close collaboration with the staff of the MRC Human Genetics Unit's Edinburgh Mouse Atlas Project (EMAP); EMAP provides the data utilised in this use case. The EMAP data has been divided in three: anatomy of developmental mouse, textual annotations and spatial annotations. Both the anatomy and the textual annotations have been semantically modelled and loaded into the CUBIST repository so that they feature within the current prototype. The same data set has been the focus of a collaboration between HWU and SHU, which aims to explore the suitability of Formal Concept Analysis (FCA) within the use case. Early results (published in the first CUBIST workshop) were promising. The biologists were excited by the possibility of developing an automatic, easy to use, mechanism for comparing and contrasting similar sets of information. For example, comparing the genes expressed in the left foot to those expressed in the right foot.

In the final year of WP7 four streams of activity took place. Firstly, the technical partners (working alongside HWU) developed the final CUBIST prototype, specialised it for the HWU use case, and evaluated it with use case experts from the EMAGE team. Secondly, collaboration between HWU and SHU led to a simplified lattice visualisation for use within this use case. In the third line of work, HWU focused on the development of semantic representations of biomedical images. This research involved the use of so-called spatial descriptions to describe regions of interest (gene expression) within an image. This complex activity will continue after the end of CUBIST, with increased collaboration from other parties and other use cases. The final theme of this year was the development of a semantic visualisation of gene expression information. Inspired by Cubix, HWU use sunburst and icicle diagrams to display the mouse anatomy, with each node representing a tissue in the mouse. Subsequently, the nodes are coloured in order to indicate the presence of gene expression information for the associated tissue. These visualisations have been well received by the biological users at EMAGE.

The HWU dataset currently comprises 1.367.578 triples and six types.

**SAS:** During the first year of the project, in which the requirements were formalized among other achievements, after considering various alternatives, a specific Use Case has been selected with the help of expert SAS operators to form the basis for the Use Case prototype. Following this step, the

properties and structure of the relevant data sets have been described and the initial, preliminary ontologies covering both the structured and unstructured data sources have been defined, and then shared with CUBIST partners.

The major achievement of the second year has been the implementation of the Use Case Prototype v. 1.0. To serve this goal, the initial ontology for the structured telemetry data source (Space Data Pack) has been slightly simplified, resulting in an ontology that yielded a more direct mapping between the simple tabular structure of the original data set and the triple store data (in RDF). The semantic ETL process, based on this ontology, resulted in approximately 500.000.000 RDF triples.

In the final year of WP7, the development of the final prototype took place, in which the UI was strongly redesigned. Ongoing informal evaluations and feedback from SAS help the technical partners help shaping the new UI to meet the end user needs. The final prototype was adapted to the SAS use case. Secondly, SAS took part in the final and formal CUBIST evaluation. Finally, SAS has started concrete exploitation activities, aiming at incorporating CUBIST functionalities (most importantly some CUBIST visualisations) into existing or planned products.

The SAS dataset has been extended during the last period. It currently comprises ca 500.000.000 triples and one type.

**INNO:** Similar to the other use case partners, Innovantage applied Careful analysis to the detailed documentation of the data sources to extract the semantic concepts and objects; these were then modelled in an ontology. Innovantage anticipates in future other unstructured data sources such as social media sites becoming more and more significant in the recruitment sector, in fact this is already starting to be seen with advertisements frequently being posted on LinkedIn. Innovantage scrutinized these emerging sources of data and also future semantic concepts that are not currently recognised in the data such as the skills and experience required to fulfil a vacancy.

The Innovantage dataset currently comprises ca 57.000.000 triples and seven types.