PROJECT PERIODIC 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 2010** **to Sep. 30 2011**

Name, title and organisation of the scientific representative of the project's coordinator[1]:

Frithjof Dau, PhD, SAP AG

Tel: +49 351 4811 6152

Fax: +49 6227 78-51425

E-mail: frithjof.dau@sap.com

Project website[2] address: www.cubist-project.eu

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement .

[2] The home page of the website should contain the generic European flag and the FP7 logo which are available in electronic format at the Europe website (logo of the European flag: http://europa.eu/abc/symbols/emblem/index_en.htm logo of the 7th FP: http://ec.europa.eu/research/fp7/index_en.cfm?pg=logos). The area of activity of the project should also be mentioned.

3.1 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.

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 order to guide the use case partners in the creation of requirements, two workshops have been conducted. The first workshop in Feb. 2011 targeted all of the above-mentioned techniques apart from mock-ups. For mock-ups, a dedicated three-day workshop has been carried out in May 2011. An example screen of the use-case independent, generalized mockup can be found in **Fig 1**.

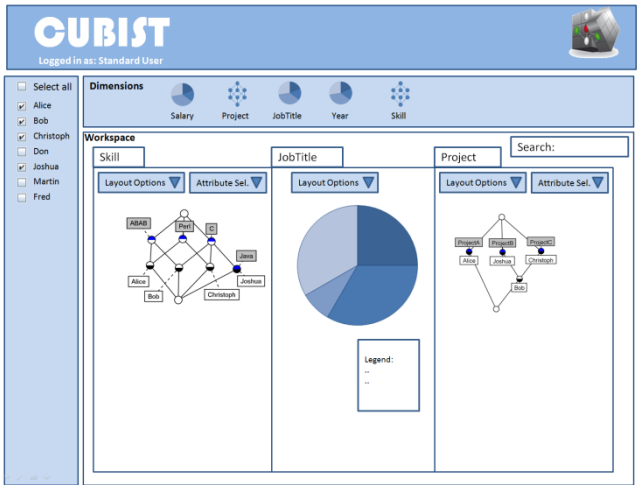

Fig 1: CUBIST Mockup

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.

Significant results are a strongly improved understanding of the use cases and their needs by the technical partners and generalized, use-case independent requirements and mock-ups.

Architecture and Software Components of CUBIST

Another main effort during the reporting period has the beginning of the design of the overall architecture for CUBIST, which includes the identification of main functional blocks and core services as well as the definition of interfaces between components. Moreover, existing software components have been adapted and extended for CUBIST, as well as new software components have been started to be developed. In the following, core software components, developed within the CUBIST consortium, 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 faceted search and graph exploration for the information stored in the semantic layer. It will be the basis for the CUBIST integrated prototype: other components will be integrated into this prototype. A screenshot, based on data of the HUW use case, is given in **Fig 2**.

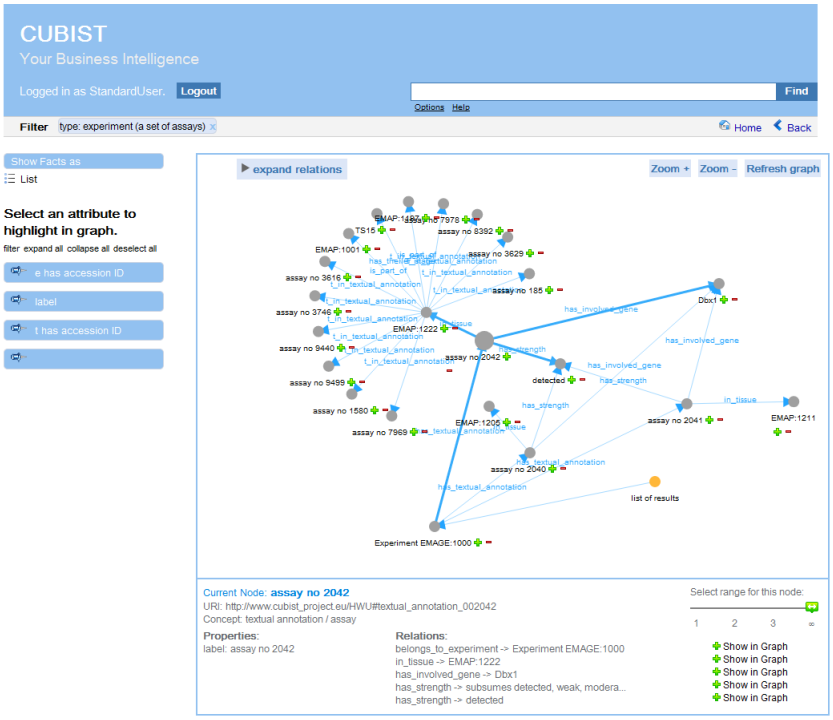

Fig 2: CUBIST prototype with graph exploration

FCABedrock and InClose2 (SHU): FcaBedrock is a desktop tool for converting CSV data to formal contexts, and In-Close2 is a fast algorithm and implementation for computing formal concepts out of formal contexts. For CUBIST, both tools are currently integrated and redeveloped in Java.. The FcaBedrock part will serve to create the BI dimensions and visualizations used in CUBIST, and In-Close2 will serve to compute the corresponding concept lattices.

In-Close2 has been further optimized and the time required to compute formal concepts has been reduced in half (in most scenarios). In-Close also supports minimum-support, which is a pre-processing technique to make the number of formal concepts in a concept lattice manageable. Further automation is planned by having In-Close automatically determining minimum-support without user intervention. Additional functionality for In-Close is planned by adding further approximation techniques such as fault tolerance.

CUBIX (CRSA) is a standalone, frontend based FCA visualization / analysis tool. CUBIX already provides novel features for visualizing large concept lattices (e.g. the transformation of lattices into trees) 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. CUBIX's workflow allows users to carry out an analysis starting from a real data set, converting it into a formal context (based on

FCAbedrock), simplifying the context to make it manageable (based on InClose), and visualizing the result as a concept lattice.

The tool is currently being developed and continuously refined with the actively involvement of our three users groups and their use cases.

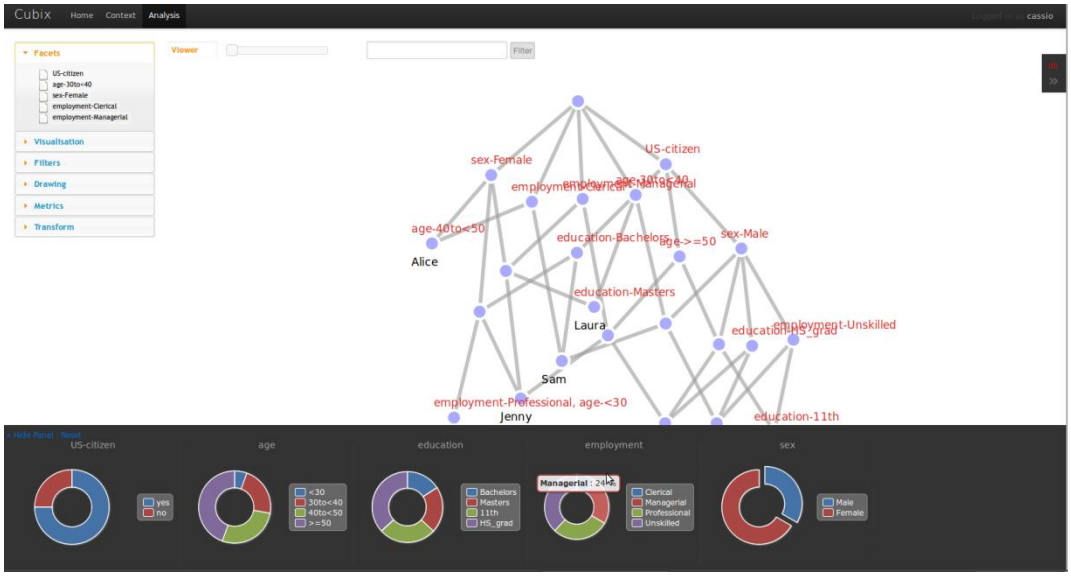

Fig 3: CUBIX prototype for VA-frontend

CUBIST Workshop

A scientific workshop for CUBIST has been conducted in July 2011. The workshop was collocated with the 19th International Conference on Conceptual Structures (ICCS), 25-29 July 2011, University of Derby, UK. The worskhop has been dedicated to topics related to CUBIST, but not to participation of CUBIST members only. he proceedings of the workshop have been published on CEUR, Vol 753 (see http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-753/).

Use Cases

In all use cases, 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.

Moreover, we can report the following advances and targets:

HWU: The potential value of the techniques being advanced (published as paper in the 1st CUBIST workshop) has been reviewed by biologists, and deemed to carry significant potential. It represents a quick way of performing the kinds of analysis they would like to carry out.

A core part of the data in the HWU use case, namely the textual annotation data, has been modelled semantically and encoded as RDF. It is acknowledged that the current version of RDF is merely a starting point for future exploration – not least because it does not contain any information relating to the spatial annotations. The long-term goal is to develop a

semantic description of biomedical images, such as those documenting the results of gene expression experiments found in this use case. In particular, this representation shall allow the spatial annotations to be converted into RDF, and thus fed into the central strand of CUBIST labour.

SAS: During the requirement analysis phase, various scenarios for utilizing the future CUBIST system have been evaluated and one particular scenario has been selected as the preferred one by the operators for the first Space Control Centres Use Case. An according data set has been described in a paper in the 1st CUBIST workshop and distributed to the partners.

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.