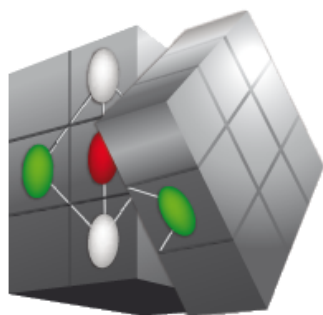


PROJECT PERIODIC REPORT



cubist

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 2011 to Sept. 30 2012

Version Version 1

Date tbc

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	3
Project Context and Objectives	3
Requirement Analysis	3
Architecture and Software Components of CUBIST	4
CUBIST Workshop and Special Journal Edition, Public Dissemination	6
Use Cases	6

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 be 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, adaptations 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 **Error! Reference source not found..**

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 C# by means of web-services. 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.

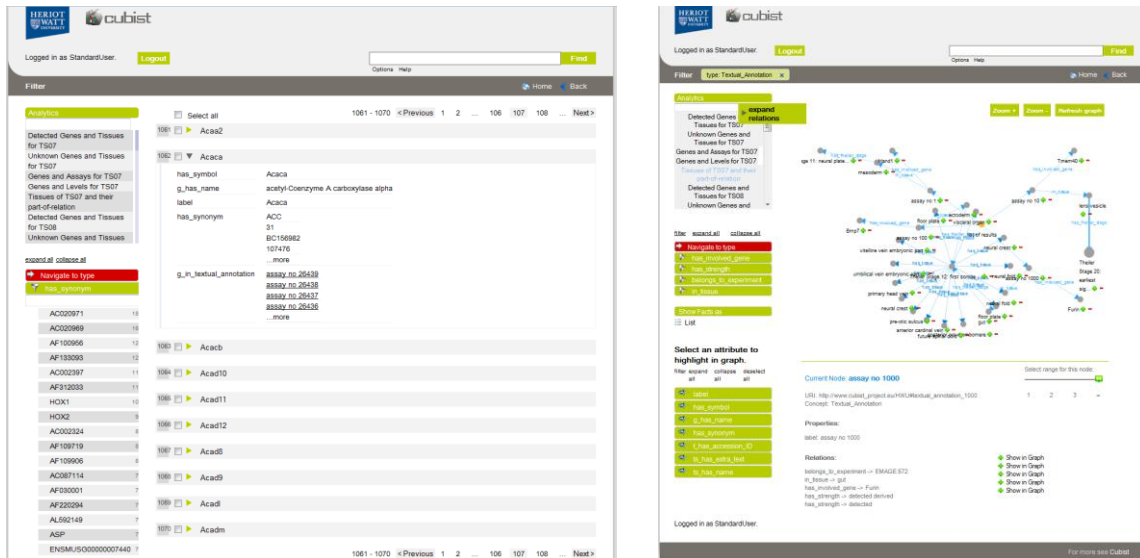


Fig 1: Faceted Search and graph exploration in NowaSearch

InClose2 has been further optimized and the time required to compute formal concepts has been reduced in half (in most scenarios). InClose2 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 InClose2 is planned by adding further approximation techniques such as fault tolerance.

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 active 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 Fig 3 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.



Fig 2: 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 third workshop is planned to be collocated with the 11th International Conference on Formal Concept Analysis, which will take place in May 2013 in Dresden, Germany. The workshop is dedicated to topics related to CUBIST, but not to participation of CUBIST members only. In fact, in both workshops conducted so far, we have received submission from outside the consortium.

A special CUBIST edition of the International Journal of Intelligent Information Technologies¹ is currently in preparation; the authors of the best publications of the CUBIST workshops have been invited to submit extended versions of their papers.

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/>).

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.

¹ <http://www.igi-global.com/ijiiit>

- 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 then have been converted into formal analytics, available in the prototype.
- For each use case, the general prototype has been customized to the respective use case.

With respect to the latter steps, Fig 3 provides for each use case a screenshots of the start screen, the list of types, the list of analytics, and one visualization.

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. Current work considers the use of techniques related to FCA, such as fault tolerance, to reduce the incompleteness and inconsistency naturally found within all life science domains. This work is the subject of a journal paper currently under review.

The ontology used within the current CUBIST prototype will evolve during year three as the spatial annotation data is added. These annotations are images that contain a series of spatial-temporal biological data. So far, a number of techniques for modelling this data in RDF have been reviewed. Whilst no mechanism is ideal, and further research on this task continues, it is now possible to encode the spatial annotations in RDF in such a way that they can be included within the CUBIST repository. This activity will be undertaken in year three.

During year two a number of assessments have been carried out with potential end users of CUBIST. One clear result of this work is the biologist's desire to see high quality visualisations of their data that enable the navigation and querying of data, alongside the analysis of results. There is a demand for more visualisation than CUBIST has attempted to date; accordingly, HWU will appraise a number of visualisations techniques in year three. When considering the subjects of this use case, it is worth noting the educational role of CUBIST. When the project started the biologists at EMAP did not have access to BI tools, and thus were unaware of the potential they offered. One clear achievement of CUBIST has been the enlightenment of these individuals, and the creation of a group of potential end users excited to see what the project delivers.

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.

The next step that followed RDF conversion phase has been the user interface implementation of the Use Case prototype. This has been achieved by applying SAS web site's design characteristics (such as colour scheme, logo, etc.) to the general prototype. The next step has been implementing the Analytics part of the prototype. In order to achieve this goal, an expert operator from SAS designed a list of queries in a semi-formal language that were supposed to be run against triple store. These queries then were converted into proper SPARQL semantic web queries by closely collaborating with SAP.

The last phase of this period resulted in a semi-formal evaluation of the v1 of the Use Case prototype. An expert operator from SAS conducted a small scale evaluation and shared her findings with other CUBIST partners, showing the directions to be taken for the v2 of the prototype.

The SAS dataset 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.

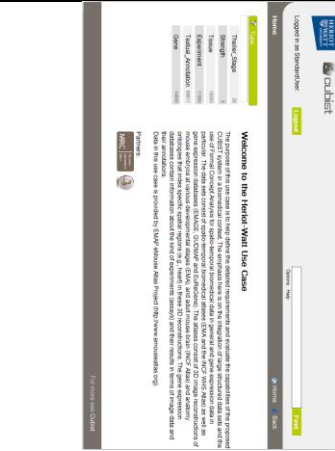
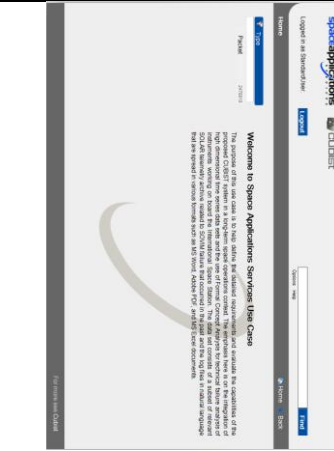
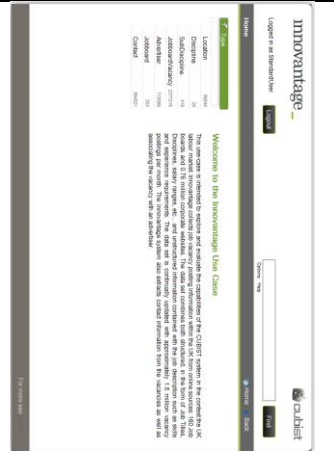
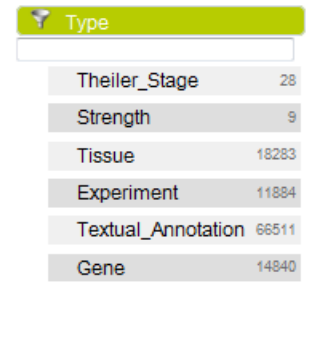
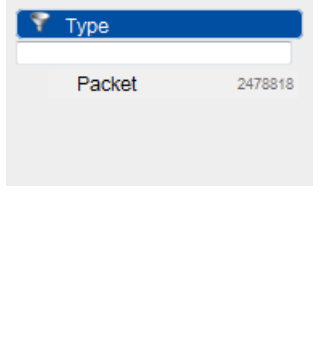
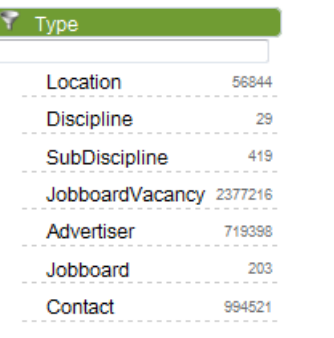
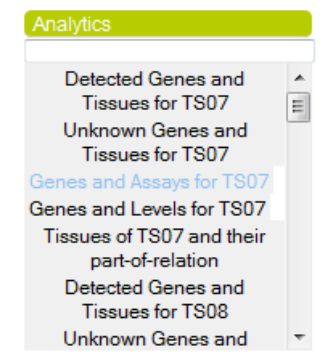
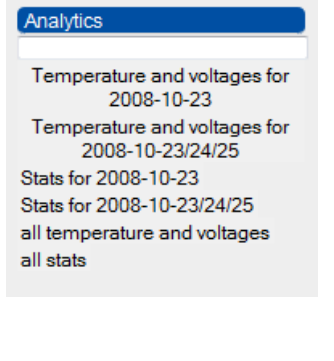

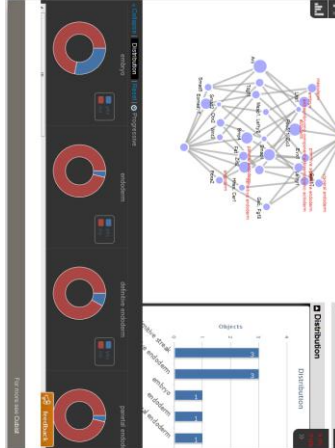
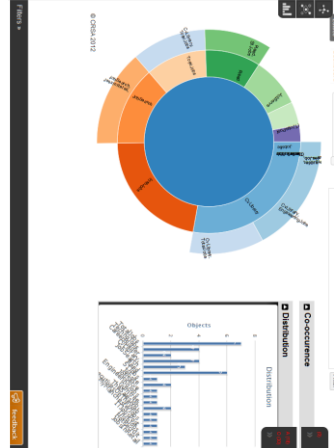

Use Cases				
	HWU	SAS	Inno	
Start Screen				
Types				
Analytics				
Visualizations				

Fig 3: Start screen, types, analytics, and one visualization for each CUBIST use case