Combining and Uniting Business Intelligence with Semantic Technologies

Acronym: CUBIST

Project No: 257403

Small or Medium-scale Focused Research Project

FP7-ICT-2009-5

Duration: 2010/10/01-2013/09/30

CUBIST Standardization Report v.1

Abstract: n/a

Type	Report
Document ID:	CUBIST D5.3.1
Workpackage:	WP5
Leading partner:	SAP
Author(s):	Johannes Meinecke (SAP) Kenneth McLeod (HWU) Cassio Melo (CRSC) Constantinos Orphanides (SHU)
Dissemination level:	PU
Status:	Final
Date:	27 October 2011
Version:	1.0

Versioning and contribution history

Version	Description	Contributors
0.1	Draft	Johannes Meinecke (SAP)
0.2	First complete version	Johannes Meinecke (SAP)
0.3	Review 1	Chris Rafferty (INN)
0.4	Review 2	Constantinos Orphanides (SHU)
1.0	Internal review comments taken into consideration	Johannes Meinecke (SAP)

1 Introduction

The CUBIST project investigates ways to bring Business Intelligence to a new level of precise, meaningful and user-friendly analytics of data by combining technologies from the fields of Business Intelligence, Semantic Technologies, and Visual Analytics. In CUBIST, the task 5.3 "Standardization" is concerned with evaluating, planning and submitting specifications that can be used as a basis for a formal standard of a European standardisation body, and is supported by the appropriate industry. This is achieved by collecting the specifications made in the other work packages and transforming them into a consistent set of documents.

As a first of three consecutive deliverables, D5.1.5 "Standardization ~~Dissemination~~ report, v.1" focuses on an early analysis of the emerging project results with respect to their potential for standardization. The deliverable is structured as follows: After this introduction in section 1, section 2 deals with three identified areas of project results and for each gives an overview of existing standards in the area as well as an assessment of new contributions from CUBIST. Section 3 is concerned with potential target standardization channels. Section 4 concludes with a summary.

2 Subjects of Standardization

CUBIST brings together different fields including Business Intelligence, Formal Concept Analytics, Visual Analytics, Semantic Technologies as well as the application domains of the use case partners. Hence, CUBIST project results that can be generalized may also be relevant for standardization in different fields.

Based on discussions in the consortium and the emerging results from the first project year, we decided to focus the investigation on three areas of project results that are subject of the following three subsections, in the order of interest. First, an important aspect in the CUBIST architecture is the exchange of data structures for the formal concept analysis with new or extended formats. Moreover, a part of the activity in CUBIST is concerned with modelling the domains of the use case partners as (light-weight) ontologies, which as shared conceptualizations of the domains naturally lend themselves for reuse beyond the project, possibly based on new standards. Finally, the work in CUBIST is also concerned with querying these ontologies for the purpose of analysis and visualization, where the area of query languages and query language extensions for CUBIST-like applications is of potential interest for standardization.

2.1 FCA-specific formats

2.1.1 Existing standards

In the area of formal context analysis, there exist already a number of formats for storing and exchanging data for different purposes. In the following, we describe formats for representing the formal contexts themselves, for representing additional pre-processing information and metadata (corresponding to the current practice in the available tools), as well as formats used to support FCA visualizations.

2.1.1.1 Formal Context Formats

The two most common FCA formats are the Burmeister (.cxt) and FIMI (.dat) formats. Burmeister originates from the FCA community and the FIMI originates from the Frequent Itemset Mining Implementation (FIMI) community [G2004].

The Burmeister is a popular format for FCA tools. A Burmeister file begins with the number of objects, followed by the number of formal attributes and then lists the objects, followed by the formal attributes. It then stores the body of the formal context as a grid, using crosses for True values and dots for False values.

Here is an example of a Burmeister .cxt file (Listing 1):

```
B

5
8

0
1
2
3
4
bruises?
gill-size-broad
gill-size-narrow
veil-type-partial
veil-type-universal
ring-number-none
ring-number-one
ring-number-two
XX.X.X..
X.XX...X
..XX.X..
XX.X..X.
..XX.X..
```

Listing 1: Example.cxt, Burmeister context file

As opposed to Burmeister, a FIMI file only consists of rows of numbers; each row represents an object and each number represents a formal attribute. The ordering of the attributes is as one would expect from the formal context, taking the first column of the context to be attribute one, and so on. The corresponding FIMI file of the Burmeister file above can be seen below (Listing 2).

```
1 2 4 6
1 3 4 8
3 4 6
1 2 4 7
3 4 6
```

Listing 2. Example.dat, FIMI context file

In the FIMI format, instead of a true/false grid we only have the featured attributes for each object. The first row, for example, is interpreted as "object 1 has the 1st, 2nd, 4th and 6th attribute".

It can also be noticed that the FIMI format is semantically 'poor' when compared to Burmeister, mostly due to the fact that this format is used in testing the efficiency of concept mining algorithms (InClose and FCbO being some of them [A2009,KOV2008]) where object names, attribute names and attribute values do not matter.

2.1.1.2 **Preprocessing & Metadata Formats**

FcaBedrock is a formal context creator for FCA, currently being developed in CUBIST.

In FcaBedrock, the user supplies the tool with appropriate metadata for conversion, such as the names of the attributes and their values, and with decisions as to what to convert and how to convert it. After reading in the original data file, these metadata are used by FcaBedrock to create a formal context file in a standard form for FCA (Burmeister and FIMI are both available as options). The metadata are stored in a separate text document called a *Bedrock* (.bed) file. This can be used for subsequent conversions and act as a record of the interpretation made of the dataset. Bedrock files can be loaded into FcaBedrock, allowing the reproduction of context files and allowing changes in the interpretation to be made. User-defined constraints applied to the data allow different analyses to be carried out. Each analysis can be documented with a Bedrock file. Multiple data files with the same attributes can be converted using the same Bedrock file.

FcaBedrock metadata supports multiple attribute types to cater for all kinds of analyses:

a) Categorical attributes: this is the typical many valued attribute (e.g. 'Color' can have multiple values such as red, green, blue, black, yellow etc)
b) Ordinal attributes: this is the same as categorical apart from the fact that the order in which attribute values appear is significant (e.g. in a 'Month' attribute January comes before February, August comes before July and so on). Ordinal attributes can be grouped using ranges (e.g. January-March, April-June, etc)
c) Continuous attributes: these are numerical attributes which can be grouped by defining ranges (e.g. 10- <20, 20- <30, >=30 etc)
d) Boolean attributes: The typical Boolean attribute with true/false values.
e) Dates: Attributes representing dates in various formats.

2.1.1.3 **FCA-related visualization formats**

The *Hasse* diagrams produced by FCA represent graphs with particular properties. Most FCA software offers data-centric formats for formal contexts and concepts. Only a few of them allow saving and loading the concept lattice. Here is a resume of popular FCA software lattice export options:

Conexp

- *CSC* format stores drawings primitives like lines thickness colour and position, labels but no structural information;
- *TXT* describing structural information only such as nodes, edges positions and labels;
- Static image file (*JPG and PNG*).

Galicia

- Static image file (*PDF, JPG and PNG*);
- *SVG* (Scalable Vector Graphs): A very popular xml-based schema for vector graphics visualization in general. No structural description support;
- *LAT.XML*: xml-based schema for lattices, describes the hierarchy of nodes and the min-support for the lattice. It is the only tool to support lattice meta-data natively.

ToscanaJ

- Static image file (*JPG and PNG*).

Alternatives for lattice visualization format can be found in graph modelling/analysis software. The most popular are:

GML (Graph Modelling Language): Offers a very simple syntax for structural description only (nodes and connections).

GraphML: The most popular format for graphs description. It features structural (nodes and connections) or visualization (colour, size, thickness) description. It is extensible enough to the addition of custom attributes/values.

GEXF (Graph Exchange XML Format): An extensible schema for graphs, it includes almost everything of GraphML, but is still in early phases of adoption by existing software. It supports structural description (nodes and edges) as well as visualization attributes, labels and graph meta-data. It allows also the modelling of graph dynamics, *i.e*, defining lifetime to nodes, edges and data.

An interesting comparison among these and other graph formats can be found at [GC2011].

2.1.2 Potential for CUBIST contributions

2.1.2.1 FCA-related data and metadata formats

In terms of formal contexts, Burmeister is a suitable formal context format for usage in CUBIST. It contains all the necessary information to conduct a semantically and conceptually meaningful analysis of the data. It is also suitable as an input format for the visualisation of the lattices.

In terms of storing all the metadata and interpretations made for each analysis, the Bedrock files are principally suitable within the CUBIST context. The structure of the Bedrock file will be modified to accommodate for the new features developed during the project time. Another aspect is that Bedrock files at the moment are stored in text formats, so as to allow front-end users to edit the document outside of the tool. For CUBIST however, a custom XML-based format of the Bedrock files would be more ideal, as they are easy to parse and more readable than text files, especially when talking about analyses involving millions of objects and thousands of attributes.

Currently there are no standards for pre-processing metadata, as there are not many formal concept creators and for the few that exist, each has their own way of dealing with metadata. Hence, the XML format developed in CUBIST could form the basis of a future standard for the purpose of new tools/software that might emerge in the future, that may want to produce "CUBIST metadata" (files that can be used as input on CUBIST tools).

2.1.2.2 FCA-related visualization formats

In the context of the CUBIST project, the visual analytics features combined with FCA adds another dimension in the complexity of formats not supplied by current standards, for instance, the history of operators and filters, calculated metrics, selections, etc. A custom XML-based lattice description format seems to be a good alternative in this case. The format should handle the structural information, visualization states, and other lattice metadata information. Drawings descriptions are not important as they are based on the actual state of the filters and selections. A formal standardization of the visualization format is judged to be only of limited value, as the information described with it (e.g. on applied filters and selections) is rather tool-specific and lattices visualization states can already be covered with existing graph formats.

From the current perspective, the mentioned XML-formats to be developed in CUBIST are judged to be of interest to the scientific FCA community. Hence, in the project, when designing these formats, the CUBIST consortium will ensure the maximum reusability beyond the boundaries of the project. In addition, the consortium will discuss these formats on

interoperability workshops similar to [CLA2010] and [ICFCA2010] to improve their reusability and to foster their propagation.

2.2 Domain Ontologies

A part of the activity in CUBIST is concerned with modelling the domains of the use case partners as (light-weight) ontologies, in order to provide a semantically unified view over heterogeneous data sources as a basis for business intelligence. Where the information and its conceptualization is of interest to the public or to a sufficiently broad industry area, it may make sense to standardize the ontologies as a basis for interoperable software solutions across the European Union.

From the three use cases, "Biomedical Atlases" (WP7) has been identified as the one with most potential in this direction. Here, the information is of broad public interest and already publically available. In the area of "Semantic Business Intelligence for Space Control Centres" (WP8), the data used in the project is either very fine-granular, numeric data or high-level message-centric (partly unstructured) data, for which the corresponding ontology is not estimated to be of sufficiently high interest for standardization. In the area of "Semantic Business Intelligence for Recruitment" (WP9), the ontologies are mainly used for the extraction of recruitment-relevant information from unstructured sources on the Web and seem too company-specific to be standardized. Hence, we will concentrate in the following on the domain of biomedical atlases and provide a brief overview of standards in the area as well as an assessment of the potential for new standards.

2.2.1 Existing standards

In situ gene expression data should be documented according to the "Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments" (MISFISHIE) standard [ISB2004]. Currently this has no semantic representation.

There are no existing standards for the publication of gene expression information on the Semantic Web. However, Bio2RDF [LCQP2010] – a community of researchers with the goal of making life science data available on the Semantic Web – are beginning to contemplate such a standard for microarray-based gene expression information.

The International Neuroinformatics Coordinating Facility (INCF) [I2011] is attempting to build an infrastructure that will integrate a number of existing mouse atlases, for example the Edinburgh Mouse Atlas (EMA). One of the INCF task forces engaged in this work is creating a new ontology (called PONS [I2007]) for the translation and definition of terms describing neural structures at multiple levels of granularity.

In addition to the previously described ontology, the INCF is generating an architecture specification [I2010] to allow so-called "hubs" (such as EMAGE) to communicate. The

specification includes an ontology that enables gene expression information to be shared between hubs.

2.2.2 Potential for CUBIST contributions

Following CUBIST, HWU may be in a position to help with a future semantic representation of MISFISHIE. HWU are currently monitoring Bio2RDF discussions, and have engaged in initial communications with the W3C HCLS, with a view to a possible collaboration. HWU may feed CUBIST related experience into the generation of a microarray-based representation, or perhaps participate in a more generic semantic representation of gene expression data. Dr Albert Burger is a member of the INCF's PONS taskforce. Through Dr Burger, the knowledge HWU gains, whilst developing the semantic representation of the spatial annotations, will be fed into PONS. Additionally, Dr Burger is a member of the INCF task force generating the architecture specification. Again, the experience gained whilst crafting a representation for CUBIST may be fruitfully applied to the mechanism the INCF use to share gene expression data.

2.3 Query Languages

One activity of the CUBIST project (Task 3.1) is to work on a semantic query language (extension) to improve the usage of triple stores in RDF for the purposes of business intelligence.

2.3.1 Existing standards

The major standard for querying RDF-based data is the SPARQL query language, which has been published as a W3C recommendation in the version 1.0 [PS2008]. SPARQL is based on the concept of specifying required and optional graph patterns that are matched against RDF graphs. The queries can either be performed on natively-stored RDF data, or on top of heterogeneous data sources that are exposed as RDF views with the help of a corresponding middleware. SPARQL 1.0 is not suitable for OLAP-oriented queries, as it does not support language elements for aggregating numeric values, which is an important requirement for BI use cases.

In May 2011, SPARQL 1.1 [HS2011] has been published as a W3C working draft. This version extends SPARQL with additional features. Most importantly, it supports aggregation and grouping functionality with language elements like GROUP BY, COUNT, SUM etc. Furthermore, it supports different kinds of negation, sub queries, expressions in the select clause, assignments and an expanded set of functions and operators.

In addition to these standards, RDF-related tools and software solutions implement vendor-specific extensions to support more advanced query features. A prominent feature is the

support for geo-spatial queries. In the example of OWLIM, Ontotext's triple store used in the CUBIST project, this is supported with specific function calls in RDF / SPARQL syntax, like e.g. the `omgeo:nearby` predicate [O2011]. Other vendors have their own proprietary extensions. Currently, there is no ongoing work related to geo-spatial standards that could unify the access to the vendors' functionality.

2.3.2 Potential for CUBIST contributions

At the time when the proposal was written and the project was planned, the plans of the W3C working group on SPARQL were still unclear. The original plan was hence that CUBIST would contribute to the state of the art by proposing business-intelligence-capable extensions of the SPARQL query language. The extensions that have been proposed in the W3C working draft since then now offer already a good basis for the purpose of combining and uniting business intelligence and semantic technologies, as investigated in CUBIST. Our focus in this area is hence not to drive new standards or extensions, but rather reuse the results of the W3C working group. Cf. deliverable D3.1.2 for more details on CUBIST work on query languages.

3 Standardization Channels

The CUBIST consortium members are involved in various official standardization bodies through memberships and have extensive experience in standardization processes. For example, SAP has been involved in industry standardization activities at OASIS, W3C, EPC, Auto-ID, OMG and many others bodies.

Given the described analysis of CUBIST project outcomes, the subject that is most interesting for industry purposes is the query language with which analytics can be performed on RDF data stored triple stores based on RDF. As argued, due to the improved capabilities in SPARQL 1.1, we currently see no need for additional CUBIST contributions in this field.

The other analyzed project results have been identified as less interesting for industry-level standardization. Instead, the CUBIST consortium aims to contribute formats and ontologies engineered for reusability beyond the project through community-specific channels. This includes interoperability workshops, like [CLA2010] and [ICFCA2010], as well as specific bodies where CUBIST consortium members are already participating, like the PONS [I2007] and INCF [I2010] taskforces.

4 Summary

This deliverable reports on an analysis of current and future CUBIST project outcomes with respect to the potential for submitting them to an appropriate standardization body. We identified three areas of interest: the formats of exchange of data structures for the formal concept analysis, the domain models of the use cases (in particular, the Biomedical Atlases use case), and the language for querying RDF data for business intelligence and visualization. For the first two areas, we identified appropriate channels for discussing and propagating the project results, which will have to be designed in a way that fosters maximum reusability outside the project. In the third area of investigation, the query language, which would have been of potential interest for an industry-level standard, the extension of SPARQL that took place at the W3C in parallel to the early project rendered the early standardization ideas obsolete.

In the follow-up deliverable D.5.3.2, we will report on the progress in the identified areas and give an updated analysis based on the formats and specifications that have been completed by that time.

5 Bibliography

[A2009] Andrews, S.: In-Close, a Fast Algorithm for Computing Formal Concepts. In: Rudolph, S., Dau, F., Kuznetsov, S.O. (eds.) ICCS 2009. CEUR WS, vol. 483 (2009), http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-483/.

[CLA2010] Computational Logic Group – Universidad de Sevilla (2010): "Workshop: FCA Software Interoperability" on CLA (Concept Lattices and Application) 2010, http://www.glc.us.es/cla2010/.

[G2004] Bart Goethals (2004): Frequent Itemset Mining Implementations Repository website, http://fimi.ua.ac.be/.

[GC2011] Gephi Consortium (2011): Gephi supported graph formats, http://gephi.org/users/supported-graph-formats/.

[HS2011] Steve Harris, Andy Seaborne (2011): SPARQL 1.1 Query Language, W3C Working Draft, http://www.w3.org/TR/2011/WD-sparql11-query-20110512/

[I2007] International Neuroinformatics Coordinating Facility (2007): Program on Ontologies of Neural Structures, http://incf.org/programs/pons.

[I2010] International Neuroinformatics Coordinating Facility (2010): Digital Brain Atlasing, http://incf.org/programs/atlasing.

[I2011] International Neuroinformatics Coordinating Facility (2011): INCF Web site, http://www.incf.org.

[ICFCA2010] LARIM, Université du Québec en Outaouais (2010): "FCA Software Woskhop": ICFCA (Int. Conf. on Formal Concept Analysis) 2010, See http://w3.uqo.ca/icfca10/WorkShop.html.

[ISB2004] Institute for Systems Biology (2004): Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE), http://scgap.systemsbiology.net/standards/misfishie/.

[KOV2008] Krajca P., Outrata J., Vychodil V.: Parallel Recursive Algorithm for FCA. In: Belohlavek R., Kuznetsov S. O. (Eds.): *Proc. CLA 2008, CEUR WS*, **433**(2008), 71–82. ISBN 978–80–244–2111–7.

[LCQP2010] Universite Laval, Carleton University, Queensland University of Technology, Protech Solutions, Inc (2010): Bio2RDF.org, http://bio2rdf.org.

[O2011] Ontotext (2011): Geo-spatial indexing in OWLIM, http://www.ontotext.com/owlim/geo-spatial.

[PS2008] Eric Prud'hommeaux, Andy Seaborne (2008): SPARQL Query Language for RDF, W3C Recommendation, http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.