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.2

Abstract: n/a

Type	Report
Document ID:	CUBIST D5.3.2
Workpackage:	WP5
Leading partner:	SAP
Author(s):	Frithjof Dau (SAP) Kenneth McLeod (HWU) Emre Sevinc (SAS)
Dissemination level:	PU
Status:	final
Date:	27 September 2012
Version:	1.0

Versioning and contribution history

Version	Description	Contributors
0.1	Draft	Frithjof Dau (SAP)
0.2	Final draft, ready for first, non-formal review	Frithjof Dau (SAP) Kenneth McLeod (HWU) Emre Sevinc (SAS)
0.3		
0.4		
1.0	Final version after review	Frithjof Dau (SAP)

Reviewers

Name	Affiliation
Constantinos Orphinanded	SHU
Emre Sevinc	SAS

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.

This is the second of three consecutive deliverables, D5.3.2 “Standardization report, v.2”. In D5.3.1 “Standardization report, v.1” we have focused on an early analysis of the emerging project results with respect to their potential for standardization. This deliverable reports about the progress of our efforts. It 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 could also be relevant for standardization in different fields.

With respect to standardization, as stated in D5.3.1, the CUBIST consortium identified 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 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 BI-applications, thus it was initially targeted to work on SPARQL query language extensions addressing particular BI needs.

2.1 FCA-specific formats

A comprehensive overview over the existing standards for FCA-applications has been provided in D5.3.1. These standards comprise

- Formal Context Formats like the Burmeister (.cxt) and FIMI (.dat) formats,
- Preprocessing & Metadata Formats, namely the storage format of FcaBedrock, and
- FCA-related visualization formats.

As discussed, the formats for Formal Context and FCA-related visualization formats are already sufficiently covered by existing de-facto standards. Thus in the following, we focus on a possible contribution of CUBIST to Preprocessing & Metadata Formats.

2.1.1 FCAbedrock and Analytics in CUBIST

The information in the CUBIST repository is queried using SPARQL. Similar to SQL, the result of a SPARQL query is a table (possibly with empty cells). A table as such is not suited for FCA: it must first be interpreted or converted in order to transform it into a formal context.

The current approach taken in CUBIST is to split the columns of the table into two sets: columns which generate formal objects and columns which generate formal attributes. Each row in the table then induces a relationship between the generated object and attribute (see [D2011] for further details). This approach is not sufficient. Attributes of entities might come in different data formats, like numbers, strings, or dates. If such data is to be analysed, such data must be converted into formal contexts. This is usually done using an approach called "conceptual scaling". A tool capable to do so is FcaBedrock, a formal context creator for FCA.

FcaBedrock currently exists as a standalone desktop tool. It is currently redeveloped in CUBIST; its functionalities are provided as web-services such that the CUBIST prototype can utilize them.

To recap information given in D5.3.1: 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.

Ordinal attributes, continuous attributes, and dates can be discretised using user-defined ranges (e.g. 0- 10, 10-20...), or by hierarchical scaling (e.g. >0, >10, >20...). With these capabilities, FcaBedrock allows to tackle standard BI queries ("show me the numbers") as well.

2.1.2 Potential for CUBIST contributions

Using a dedicated namespace, the meta-information about analytics in CUBIST are stored in the triple store itself, using a specific class "Analytics" with five (simple) data attributes. An example for such an analytics is provided in Fig. 1. As described in the previous section, the current analytics are not expressive enough. The next step in CUBIST (which has already started) is to incorporate the facilities of FCABedrock which allow for each column of the result set of a SPARQL query to describe how it is converted into a formal context. This will lead to more expressive analytics. The meta information for this conversion directives will be based on the FCAbedrock files, but in CUBIST, of course a semantic (RDF-based) description of this information is targeted. Thus in the course of incorporating FcaBedrock into CUBIST, we will develop a small, use-case-independent ontology which will allow to describe an analytics in all needed detail.

Currently there are no standards for pre-processing metadata, as there are not many formal context creators and for the few that exist, each has their own way of dealing with metadata. Hence, the RDF 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 "FCA metadata" (files that can be used as input on FCA tools).

```
:query_gene_strength_TS07
     rdf:type :Analytics ;
     :hasName   "Genes and Levels for TS07"^^xsd:string ;
     :hasDescription  "Genes  and  Levels  of  Expressiveness  for
Theiler Stage 07"^^xsd:string ;
     :hasObjectType       hwu:Gene ;
     :hasAttributeType    hwu:Strength ;
     :hasQueryBody
          """
          PREFIX :<http://www.cubist_project.eu/HWU#>
          PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
          PREFIX owl:<http://www.w3.org/2002/07/owl#>
          PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
          PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
          select distinct ?o1 ?a1 where {
          ?x1 rdf:type :Tissue .
          ?x2 rdf:type :Gene ;       rdfs:label ?o1 .
          ?x3 rdf:type :Strength ; rdfs:label ?a1 .
          ?ta rdf:type :Textual_Annotation .
          ?x1 :has_theiler_stage     :theiler_stage_TS07 .
          ?ta :in_tissue ?x1 ;
               :has_involved_gene ?x2 ;
           :has_strength ?x3 . }"""^^xsd:string .
```

Figure 1. A CUBIST analytics

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.

In D5.3.1, we have mainly identified amongst the three use cases the “Biomedical Atlases” (WP7) as the one with most potential in this direction. This use case will be addressed in the next subsections.

Though stated in D5.3.1 that in the area of “Semantic Business Intelligence for Space Control Centres” (WP8), the ontologies to be developed are not estimated to be of sufficiently high interest for standardization, SAS (Space Applications Services) has examined whether it would be possible for the project efforts to contribute to any of the space industry standards. A thorough analysis yielded that there are no particular standards for the structured data sources (such as the housekeeping telemetry data for the Columbus Payloads, e.g. SOLAR in this case). On the other hand, one of the unstructured data sources, namely the Operations Data File documents, needs to respect the SSP 50253 "Operations Data File Standards". However, no immediate connections were to be found between this standard and other standards within the context of CUBIST, such as the SPARQL standard for semantic web. Finally, it has been observed that even though there are many requirements levied upon USOCs (User Support and Operation Centers), they are not defined by formal standards of ESA.

The estimation of D5.3.1 that “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” remains valid.

2.2.1 Existing standards for the biological use case

(This section is a repetition of D5.3.1.)

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) [INCF] is attempting to build an infrastructure that will integrate a number of existing brain 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 [INCF] to allow so-called "hubs" (such as EMAGE) to communicate. The specification includes a semantic markup that enables gene expression information to be shared between hubs.

2.2.2 Potential for CUBIST contributions

An important aspect for HWU in CUBIST is semantic representation of the spatial annotations. As in the HWU use case there is a requirement to develop semantic descriptions of images. These descriptions are not simple semantic overviews using keywords; instead descriptions should include a depiction of the elements within each image, and the position of those elements. Ultimately, the goal is to enable reasoning, analysis, and comparison based on the semantic portrayal. Currently, there are no standards for spatial descriptions within the biomedical domain. However, it is not the goal of this work to create them. It would be impossible to develop such standards based on a single use case. Nevertheless, experience gained during CUBIST is being fed into the standardisation efforts of the INCF's digital atlasing initiative [INCFa]. Standards do exist for the geospatial world, e.g., [WGS84]; however, initial experiments have indicated that significant differences between the two use cases (mouse versus Earth) render those standards mute. Within CUBIST effort will focus on exploring existing use case neutral technologies (for example, Region Connection Calculus [R1991]) and research from within the biomedical world (e.g., [BG2007]). It is hoped that this activity will yield a mechanism that can be suitably amended for the HWU use case. Through Dr Burger, the knowledge HWU gains, whilst developing the semantic representation of the spatial annotations, will be fed into the INCF taskforces.

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.

2.4 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. SPARQL 1.0 does essentially provide means for retrieving information which suits some graph patterns. It does not provide essential features needed for BI--oriented queries, most importantly set-functions which allow to aggregate on sets of entities, namely functions like min, max, sum act which aggregate numeric values.

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. As of writing this deliverable, the last call for reviews has been passed (21. August 2012). SPARQL 1.1 it not an official w3c-recommendation yet, but as it is supported by all major triple stores, it has become already a de-facto-standard.

2.4.1 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. Anyhow, the upcoming of SPARQL 1.1 renders this plan obsolete. Indeed, the analytics in CUBIST utilize many of the SPARQL 1.1 new features (like set functions and subqueries), and the added BI-functionalities of SPARQL seem to be sufficient for the CUBIST use case analytics. Thus our focus in this area shifted from driving new standards or extensions to reusing the results of the W3C working group.

3 Standardization Channels

The standardization channels of the CUBIST consortium have not significantly changed since the delivery of D5.3.1. We recap the corresponding section of D5.3.1 for the sake of the reader’s convenience.

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 for potential standardization. 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. 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.

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/.

[BG2007] Bittner, T. and Goldberg, L. J., 2007. The qualitative and time-dependent character of spatial relations in biomedical ontologies. Bioinformatics, Vol. 23, Nr. 13, 1674-1682. doi: 10.1093/bioinformatics/btm155

[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/.

[D2011] Frithjof Dau: Towards Scalingless Generation of Formal Context form an Ontology in a Triple Store. In: Dau, F and Andrews, S: Proceedings of the second CUBIST workshop 2012. KULeuven press, 2012

[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.

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

[INCFa] http://incf.org/programs/atlasing

[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.

[R1992] Randall, D. A., Cohn, A. G., & Cui, Z. (1992). A spatial logic based on regions and connection. Proceedings of the 3rd international conference on knowledge representation and reasoning (pp. 165-176). San Mateo: Morgan Kaufman.

[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/.

[WGS84] W3C. (2003). WGS84 Geo positioning: an RDF vocabulary. Retrieved 09 10, 2012, from http://www.w3c.org/2003/01/geo/wgs84_pos