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.3

Abstract: n/a

Type	Report
Document ID:	CUBIST D5.3.3
Workpackage:	WP5
Leading partner:	SAP
Author(s):	Frithjof Dau (SAP)
Dissemination level:	PU
Status:	Final
Date:	27 October 2013
Version:	1.0

Versioning and contribution history

Version	Description	Contributors
0.1	Draft	Frithjof Dau (SAP)
1.0	Comments from reviewers addressed	Frithjof Dau (SAP)

Reviewers

Name	Affiliation
Simon Polovina	SHU
Katja Pfeifer	SAP

1 INTRODUCTION 4

2 SUBJECTS OF STANDARDIZATION 5

2.1 FCA-SPECIFIC FORMATS 5

2.2 DOMAIN ONTOLOGIES 9

2.3 QUERY LANGUAGES 10

3 SUMMARY 11

4 BIBLIOGRAPHY 12

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 third of three consecutive deliverables, D5.3.3 “Standardization report, v.3”. This deliverable reports about the results 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, 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 D5.3.2 we focused 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.

In D5.3.2, a standalone-tool from SHU called “FCABedrock” has been described which allows to convert tables, using an approach called “conceptual scaling”, into formal contexts. In the last year of the project, a web-service “FCAService” has been developed which provides similar conceptual scaling facilities like FCABedrock. The FCAService 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) Continuous attributes: these are numerical attributes which can be grouped by defining ranges (e.g. 10- <20, 20- <30, >=30 etc)

c) Dates: Attributes representing dates in various formats.

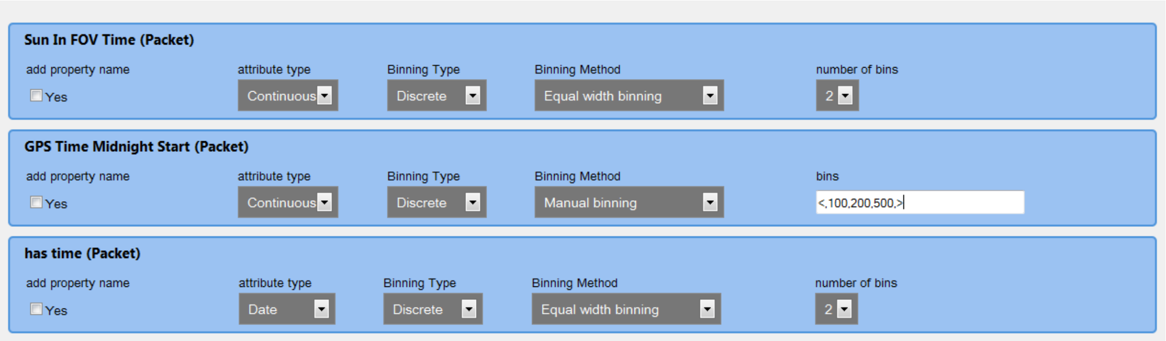
The values of those types have to be "translated" into formal attributes (in the FCA understanding). For example, continuous (and date) attributes are dealt with by producing discretized values of a continuous attribute and replacing it with the new values.. The ranges created when discretizing a continuous attribute are also called bins. In the data binning process, the original data values which fall in a given small interval (the so-called bin) are replaced by a value representative of that interval. FCAService provides several parameters to fine-tune this binning process. To name:

- **Binning Types:** CUBIST supports *discrete binning,* where disjoint formal attributes are generated (e.g. "age<20", "20 ≤ age < 30", "30 age"), and "progressive scaling", where cascading formal attributes are generated (e.g. "age<20", "age<35", "age<50").

- **Binning Methods:** CUBIST support different way how the borders of the bins are generated. There are
  - **Equal width binning:** the attribute is scaled using ranges of equal width.
  - **Equal frequency binning:** the attribute is scaled resulting in bins of equal size.
  - **Standard deviation binning**: Six bins are created using the statistical mean and a standard deviation of 10.
  - **Manual binning:** the borders of the ranges (bins) are entered manually. Moreover, one can enter two special borders "<" and ">", meaning "all values below the smallest border" and "all values above the largest border".

Depending on the method, the bins or the number of bins has to be further specified.

A screenshot of the corresponding interface in CUBIST is provided in Fig 1.



Sun In FOV Time (Packet)

add property name: □ Yes | attribute type: Continuous | Binning Type: Discrete | Binning Method: Equal width binning | number of bins: 2

GPS Time Midnight Start (Packet)

add property name: □ Yes | attribute type: Continuous | Binning Type: Discrete | Binning Method: Manual binning | bins: <,100,200,500,>

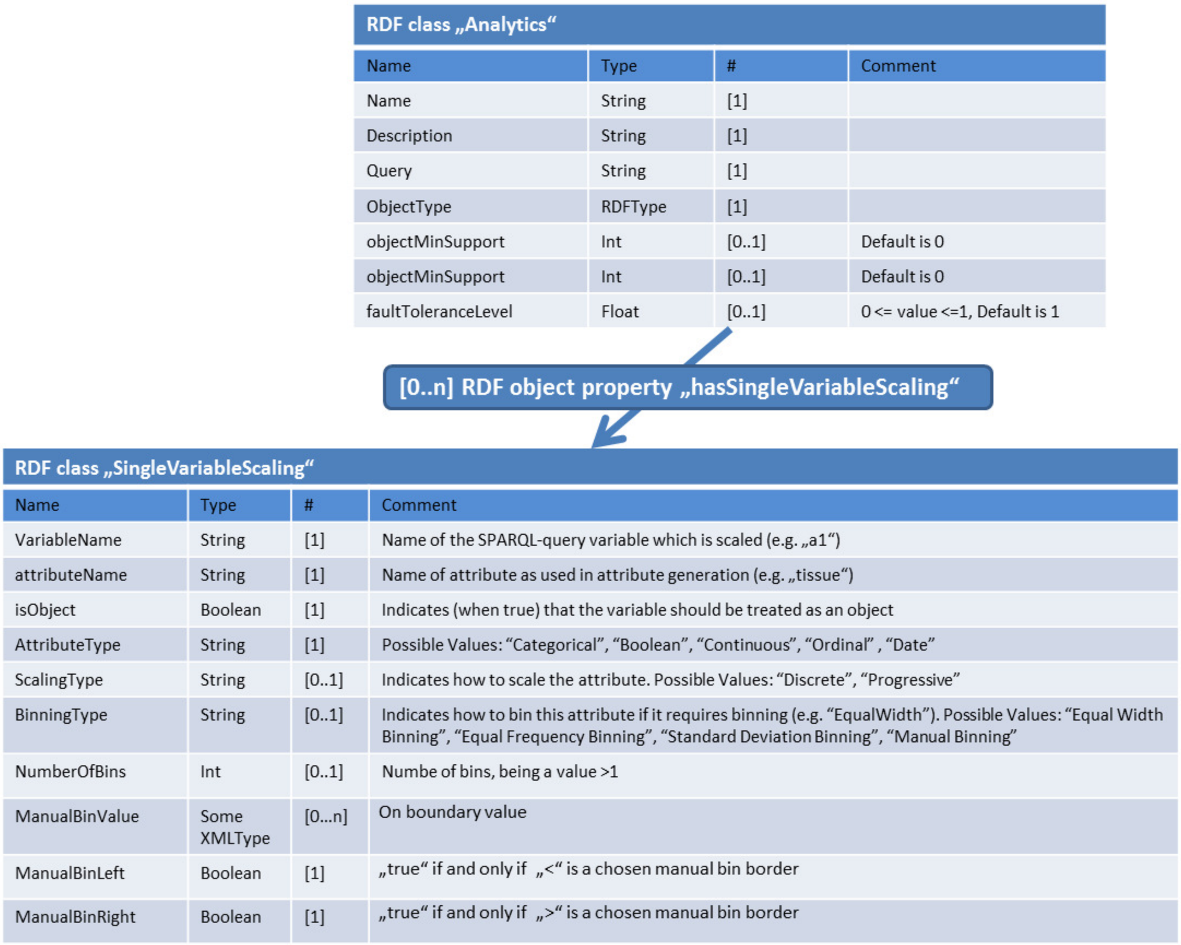
has time (Packet)

add property name: □ Yes | attribute type: Date | Binning Type: Discrete | Binning Method: Equal width binning | number of bins: 2

Fig 1: Conceptual Scaling ("Binning") in CUBIST

2.1.2 Schema for Preprocessing & Metadata Formats

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. As stated in D5.3.2., a semantic (RDF-based) description of this information was targeted, by means of a small, use-case-independent ontology. Such an ontology has been conceptually developed and is depicted in Fig 2. An instance of the “Analytics”-class contains the general information about an analytics, e.g. its name, or its corresponding SPARQL-query. The result of the query is a table, where one column generates the formal objects, and some columns generate formal attributes. Each column corresponds to a query variable in the “Select”-clause of the query. Each column which generates a formal attribute is subject of conceptual scaling. The information on how a query variable is scaled is captured by instances of the “SingleVariableScaling”-class. An instance of the “Analytics”-class should for each of its query variables which generate formal objects or attributes a corresponding instance of the “SingleVariableScaling”-class assigned, via the property “hasSingleVariableScaling”.



RDF class „Analytics“

Name	Type	#	Comment
Name	String	[1]	
Description	String	[1]	
Query	String	[1]	
ObjectType	RDFType	[1]	
objectMinSupport	Int	[0..1]	Default is 0
objectMinSupport	Int	[0..1]	Default is 0
faultToleranceLevel	Float	[0..1]	0 <= value <=1, Default is 1

[0..n] RDF object property „hasSingleVariableScaling“

RDF class „SingleVariableScaling“

Name	Type	#	Comment
VariableName	String	[1]	Name of the SPARQL-query variable which is scaled (e.g. „a1“)
attributeName	String	[1]	Name of attribute as used in attribute generation (e.g. „tissue“)
isObject	Boolean	[1]	Indicates (when true) that the variable should be treated as an object
AttributeType	String	[1]	Possible Values: “Categorical”, “Boolean”, “Continuous”, “Ordinal” , “Date”
ScalingType	String	[0..1]	Indicates how to scale the attribute. Possible Values: “Discrete”, “Progressive”
BinningType	String	[0..1]	Indicates how to bin this attribute if it requires binning (e.g. “EqualWidth”). Possible Values: “Equal Width Binning”, “Equal Frequency Binning”, “Standard Deviation Binning”, “Manual Binning”
NumberOfBins	Int	[0..1]	Numbe of bins, being a value >1
ManualBinValue	Some XMLType	[0...n]	On boundary value
ManualBinLeft	Boolean	[1]	„true“ if and only if „<“ is a chosen manual bin border
ManualBinRight	Boolean	[1]	„true“ if and only if „>“ is a chosen manual bin border

Fig 2: An RDF Schema for FCA Scaling Information

With respect to standardization, a closer observation reveals two possible obstacles:

1) With RDF, it is only possible to model the names of classes, properties, and attributes, and it is possible to model that the property “hasSingleVariableScaling” relates instances of the “Analytics”-class to instances of the “SingleVariableScaling”-class. It is not possible to model the depicted cardinality-constraints of the attributes. For this, a heavy-weight ontology language like OWL is needed. This goes beyond the level of Semantic Technologies used in CUBIST.

2) The depicted ontology particularly models the analytics as carried out in CUBIST. In its current form, it is not suited to model other meta-information which might be needed in different FCA-applications. Devising an ontology for FCA-meta information which can be used in arbitrary FCA-applications turned out to be a major effort beyond the scope of CUBIST.

For this reason, the provided ontology can only serve as a starting point for a general description of FCA-meta-information. A paper describing FCAService and the corresponding ontology is planned for the next “Conference on Concept Lattices and Their Applications” (CLA), which will we held in Q4/2014 and which is the appropriate conference to address the FCA-community with the ontology. We hope to provide with our ontology to lay the basis for a de-facto-standard in the FCA community for modelling FCA metadata.

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 and D5.3.2, we have mainly identified amongst the three use cases the "Biomedical Atlases" (WP7) as the one with most potential in this direction. More precisely, 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. 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, knowledge gained in CUBIST is fed into appropriate channels, The most important one is 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.

Dr. Burger is a member of the INCF's taskforce on Digital Brain Atlasing. This is an on-going commitment that will persist beyond the span of CUBIST.

Through Dr Burger, knowledge of semantic spatial descriptions is constantly transferred to the INCF taskforce on Digital Brain Atlasing. Dr Burger continues to brief the INCF Digital Brain Atlasing task force on the discoveries relating to semantic spatial descriptions developed during CUBIST. He attended a Task Force meeting in Stockholm, Sweden, on August 29-31, 2013, and agreed on further collaborations with other INCF members to enhance brain atlas data integration using semantic spatial descriptions. The objective is to include this work in the INCF's Digital Atlasing Infrastructure (DAI) framework and then seek further US and/or EU funding, with the help of INCF, to deploy DAI internationally.

2.3 Query Languages

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

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. In March 2013, SPARQL 1.1 became on official w3c-recommendation, thus a standard in the semantic web field.

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 rendered this plan obsolete. Indeed, the analytics in CUBIST utilize SPARQL 1.1 new features, and the added BI-functionalities of SPARQL seem to be sufficient for the CUBIST use case analytics. Thus, as been already argued in D5.3.2, our focus in this area shifted from driving new standards or extensions to reusing the results of the W3C working group.

3 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, but no new self-contained standard has been developped. 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.

4 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