Semantic OLAP with FluentEditor and Ontorion Semantic Excel Toolchain

Dariusz Dobrowolski
Faculty of Mathematics, Physics and Computer Science
Maria Curie Sklodowska University
Lublin, Poland
e-mail: dariusz.dobrowolski@umcs.lublin.pl

Paweł Kapłański
Department of Applied Informatics in Management
Gdansk University of Technology
Gdansk, Poland
e-mail: pawel.kaplanski@pg.gda.pl
Cognitum
e-mail: p.kaplanski@cognitum.eu

Andrzej Marciniak
University of Economics and Innovation
Lublin, Poland
e-mail: andrzej.marciniak@wsei.lublin.pl

Zdzisław Łojewski
Faculty of Mathematics, Physics and Computer Science
Maria Curie Sklodowska University
Lublin, Poland
e-mail: zdzislaw.lojewski@umcs.lublin.pl

***Abstract*—Semantic technologies appear as a step on the way to creating systems capable of representing the physical world as real time computational processes. In this context, the paper presents a toolchain for an ontology based knowledge management system. It consists of the ontology editor, FluentEditor and the distributed knowledge representation system, Ontorion. FluentEditor is a comprehensive tool for editing and manipulating complex ontologies that uses Controlled Natural Language (CNL). Its main feature is the usage of Controlled English as a knowledge modelling language. Ontorion is a Distributed Knowledge Management System with Natural Language interfaces (CNL) and a built-in rules engine. The Ontorion system is equipped with plugins for connection with other software environments, for example rOntorion using an R language package to access ontologies. It is exemplified with the semantic extension of On Line Analytical Processing (OLAP) using R language.**

Keywords- Semantic OLAP, Semantic, OLAP, Semantic Web, Ontorion, FluentEditor.

I. INTRODUCTION

Business Intelligence (BI) is a technology that enables the business to make intelligent, data-driven decisions. Intelligence here is governed by the laws of statistics that are applied on loosely coupled statistical variables, however to understand the meaning of data we need to link statistical variables to the real-life entities. This improvement can be implemented nowadays with aid of semantic technologies. As a result, we obtain the "smarter" BI system that represents the physical world as real time computational processes.

The remainder of the paper is structured as follows. In Section II, we present the semantic knowledge management framework that can be built with particular solutions available on the market. In Section III, we present the idea of OLAP - a powerful BI tool and its possible implementation in the R language. Semantic OLAP, the result of our research on bridging together both semantic toolkit and OLAP, is introduced and discussed in the Section IV, followed by the conclusion, in Section V.

II. SEMANTIC KNOWLEDGE MANAGEMENT FRAMEWORK

The expectations of business and science require new, global, flexible and much more effective technology of data exchange and processing. When the whole world is braided with effective communication links, what we need is a new efficient middleware working in the existing infrastructure but possessing new possibilities. After a decade of using file exchange systems, much experience was acquired. Very simple and easy rules of metadata connection gained great popularity.

Some factors should be mentioned which are important from our point of view:

- Easy exploitation: the end clients do not have any barriers;
- Accessibility: they could be used anywhere on many platforms and media;
- Effectiveness: acceptable from the point of view of the data receiver;
- Stability: information about resources must always be reliable;
- Independence: each node is completely autonomous within the system;
- Limitation of platform co-share: data are of a very simple form and the system is not able to provide co-sharing of more complex information.

The factors mentioned above indicate a tendency to recognize the meaning of a given resource, and in a later stage to the machine "understanding" of its content (i.e.,

 ISBN: 978-1-61208-420-6

ascribing semantic qualifiers to it enabling automatic decisive processes). The systems working in this layer use many technologies, which can be divided into the following categories:

- Natural language processing;
- Artificial intelligence and teaching machines;
- Ontologies;
- Meta-information, standardization and tagging documents.

By modelling domain ontologies with Semantic Web Rule Language (SWRL) [1] rules we are able to define a knowledge scheme in any semantic knowledge base. The store for the knowledge base can be implemented in Not only SQL (NoSQL) technology (e.g. Cassandra [2], Azure Tables [3]) or in Resource Description Framework (RDF) [4] data stores (e.g. AllegroGraph, Virtuoso) [5][6][7]. A relatively simple interface to model ontologies is supported by Protégé [8] or NeON [9] editors. Although these interfaces are rather simple for experienced practitioners, they are not for common users that do not know the nuances of ontology engineering. On the other hand, Semantic Rules Representation in CNL using FluentEditor [10] is the simplest way to represent knowledge in a natural language way. Nevertheless, using natural language is unattainable for the current technology and thus for the machines that should understand this knowledge. The most appropriate solution seems to use controlled natural languages.

A. *Ontorion SDK*

Ontorion [11] is a Distributed Knowledge Management System with Natural Language interfaces (CNL) and a built-in rules engine. It is compatible with Web Ontology Language 2 (OWL2) [12] and Semantic Web Rule Language (SWRL) and can be hosted in the Cloud or OnPremise environments. Ontorion is a family of products of server and client-side components for desktop and web allowing for the broad integration of custom software and existing corporate infrastructure. Ontorion performs real-time reasoning over the stream of data with the aid of an ontology that expresses the meaning of the given data (see Figure 1).

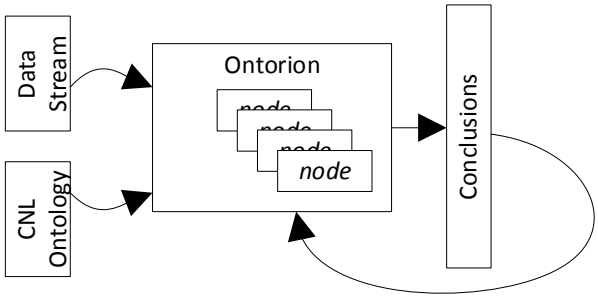


Figure 1. Ontorion [11] - Knowledge Management System

Ontorion is a set of components equipped with algorithms that allows one to build large, scalable solutions for the Semantic Web. The scalability is realized by both the NoSQL, symmetric database and the ontology Modularization algorithm. Modularization algorithm splits the problem into smaller pieces that are able to be processed in parallel by the set of computational nodes, therefore; Ontorion is a symmetric cluster of servers, able to perform reasoning on large ontologies. Every single Ontorion Node is able to make the same operations on data. It tries to get the minimal suitable ontology module (component) and perform the desirable task on it. Symmetry of the Ontorion cluster provides the ability for it to run in the "Computing Cloud" environment, where the total number of nodes can change in time depending on the user needs.

B. *FluentEditor 2014*

FluentEditor 2014, an ontology editor, is a comprehensive tool for editing and manipulating complex ontologies that uses CNL [13].

FluentEditor, shown in Figure 2, provides a more suitable alternative for human users to eXtensible Markup Language (XML)-based OWL editors. Its main feature is the usage of Controlled English as the knowledge modelling language. Supported via Predictive Editor, it prohibits one from entering any sentence that is grammatically or morphologically incorrect and actively helps the user during sentence writing.

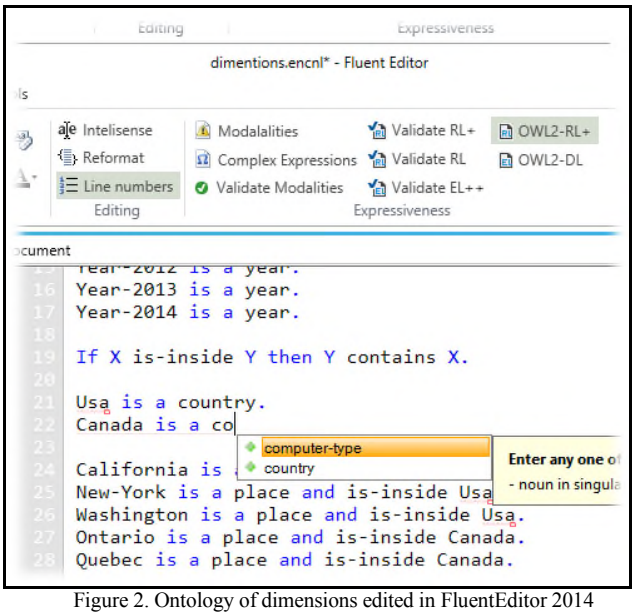


Figure 2. Ontology of dimensions edited in FluentEditor 2014

Controlled English is a subset of Standard English with restricted grammar and vocabulary in order to reduce the ambiguity and complexity inherent in full English.

Main features:

- CNL OWL implementation: The implementation of CNL OWL - FluentEditor grammar is compatible with OWL-DL and OWL2
- OWL 2.0 full compliance: Full compliance with OWL 2.0 standard from W3C

 ISBN: 978-1-61208-420-6

- OWL API: Compatible with OWL API, which allows it to be used in cooperation with other tools
- SWRL compliance: The user can import existing ontologies from OWL files
- Dynamic referencing of external OWL ontologies: CNL documents can dynamically reference external OWLs from Web or disk.
- Predictive Edition Support: Users have enhanced support from the predictive editor
- Built-in dictionary: The built-in dictionary makes it easier to avoid misspelling errors

Some examples of other features are:

- Advanced user Interface, in order to open up semantic technologies for inexperienced users,
- In-place error resolving support - direct information about possible errors with hints on how to resolve them,
- Importing existing ontologies – users can directly import to CNL any external ontology
- Ambiguity resolution - it keeps track of ambiguities of concepts and/or instance names imported from different external ontologies.

C. *R language and Ontologies*

R language [14] is a widely used tool for statistical analysis. Combining ontologies and statistics opens an efficient way for the quantitative-qualitative analysis of data. It is possible to use both approaches conveniently in a single place by using an R language package to access ontologies (rOntorion). rOntorion R package allows direct access to ontologies created with FluentEditor and opens them for semantic processing in the R environment.

The R language plugins for FluentEditor are shown in Figure 3.

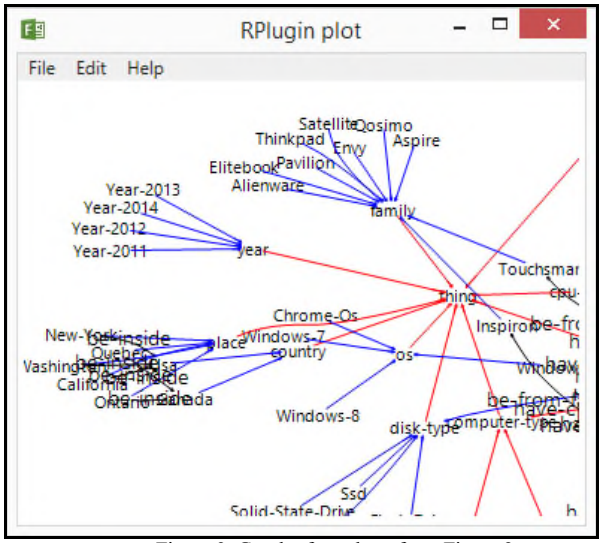


Figure 3. Graph of ontology from Figure 2

Beside development of analytical models with R and rOntorion it is also possible to build plugins for FluentEditor with R language. These plugins have direct access to the ontology within the editor host and can use any available R package. Plugins can display graphical results or textual output directly in FluentEditor.

III. INTRODUCTION TO OLAP

OLAP is a well-known method [15] used in Business Analytics to provide decision makers with Online Access to Analytical Capabilities. It is based on the concept of data-cubes, multidimensional cubes of data that if equipped with tools allow the data and problems wherein to be explored. To create a datacube, we need data that can be represented in a STAR schema. The central table contains "measures" while "dimensions" are placed in surrounding tables (see Figure 4).

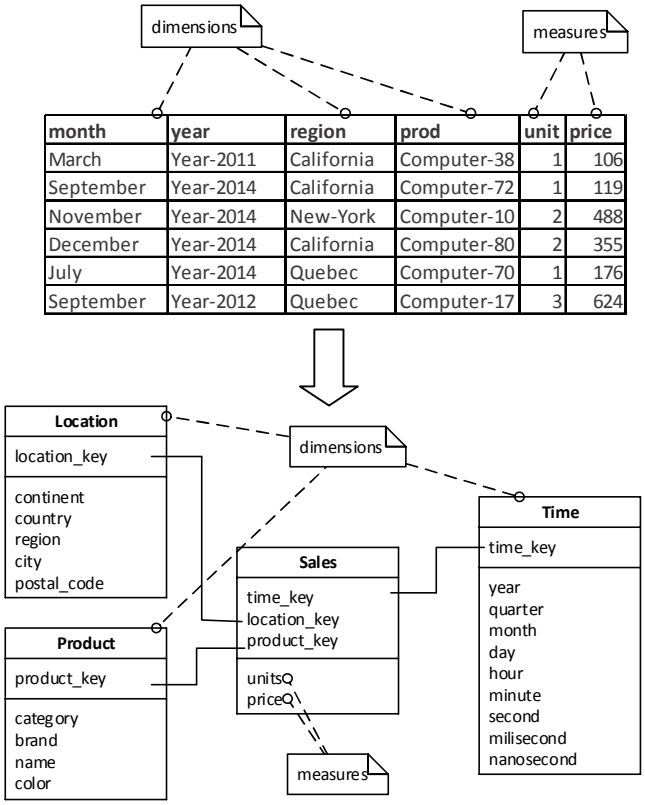


Figure 4. Transformation of a given dataset into the STAR schema (example)

To turn the data into a hypercube, we need to denormalize the STAR (by creating a single table) and what is put in each cell in the data-cube (hypercube) represents an aggregate value of measurements for a unique combination of each dimension. The aggregate is a function, e.g.: SUM, AVERAGE, MAX, MIN, COUNT, etc. See Figure 5.

 ISBN: 978-1-61208-420-6

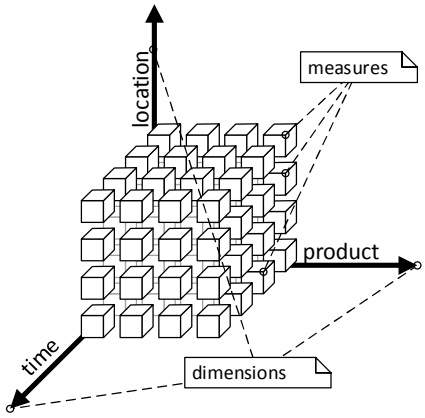


Figure 5. Extracting the data hypercube

Having the datacube we can slice and dice it (filter values), and rollup/drill-down/pivot over dimensions (see Figure 6).

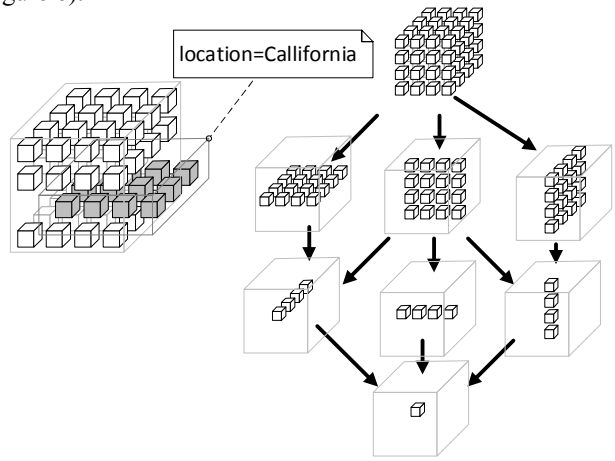


Figure 6. Slicing the data-cube over dimensions

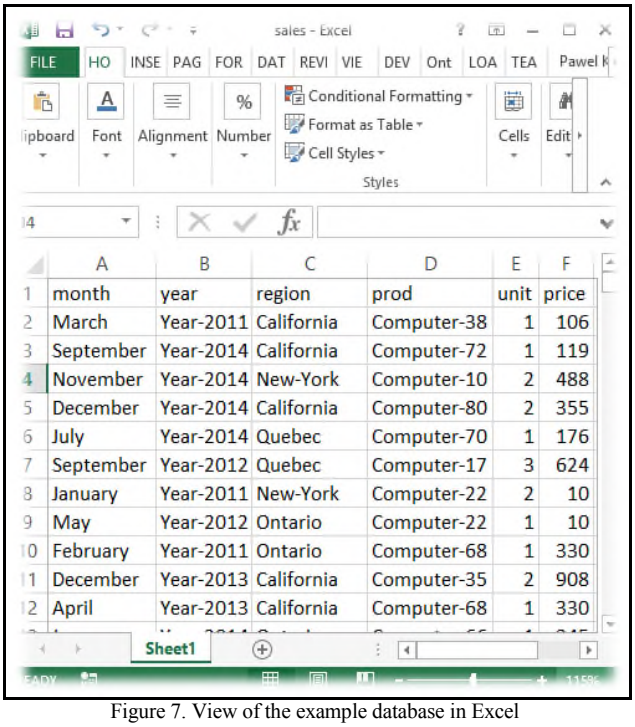
IV. THE SEMANTIC OLAP

By using the toolchain of FluentEditor and Ontorion SDK, it is possible to create something more than OLAP. We call it "Semantic OLAP", however, a solution delivered by Infotopics [16] is similar and it is called "natural query language".

In our case, the example application that implements the Semantic OLAP approach was built on top of the following tools:

- Excel [17] – to create the database (see Figure 7)
- RStudio [18] – an open source integrated development environment (IDE) for R – to develop the software.

A piece of the code of queries is shown in Figure 8 and the result of the query from Figure 8 is displayed in Figure 9.



	A	B	C	D	E	F
1	month	year	region	prod	unit	price
2	March	Year-2011	California	Computer-38	1	106
3	September	Year-2014	California	Computer-72	1	119
4	November	Year-2014	New-York	Computer-10	2	488
5	December	Year-2014	California	Computer-80	2	355
6	July	Year-2014	Quebec	Computer-70	1	176
7	September	Year-2012	Quebec	Computer-17	3	624
8	January	Year-2011	New-York	Computer-22	2	10
9	May	Year-2012	Ontario	Computer-22	1	10
10	February	Year-2011	Ontario	Computer-68	1	330
11	December	Year-2013	California	Computer-35	2	908
12	April	Year-2013	California	Computer-68	1	330

Figure 7. View of the example database in Excel

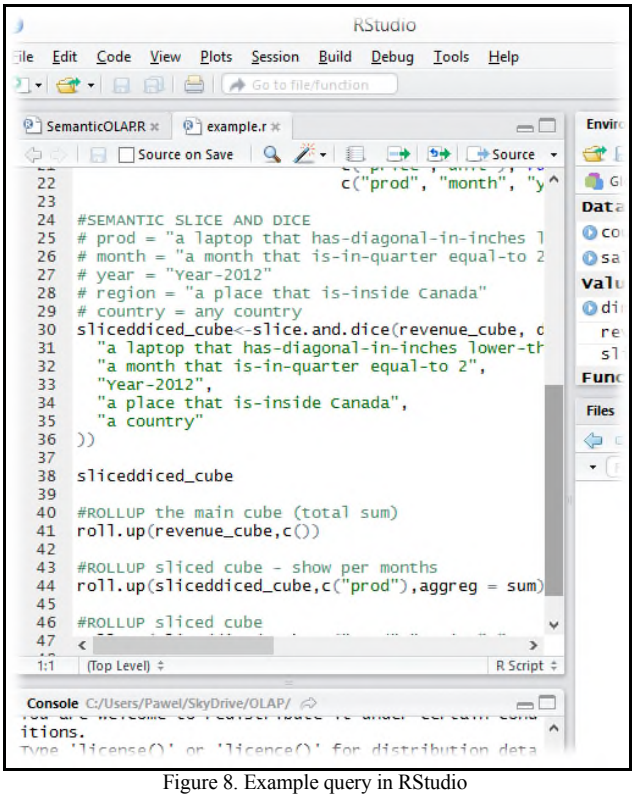


Figure 8. Example query in RStudio

```
> sliceddiced_cube
, , region = Quebec, country = Canada

             month
prod           May April June
  Computer-58   NA    NA   NA
  Computer-72   NA   357   NA
  Computer-33   NA   979   NA
  Computer-26   NA    NA  375
  Computer-32 1161   387   NA
  Computer-66  345   345   NA
  Computer-28  692    NA   NA

, , region = Ontario, country = Canada

             month
prod           May April June
  Computer-58 2640    NA   NA
  Computer-72   NA    NA  238
  Computer-33  979  4895   NA
  Computer-26   NA    NA   NA
  Computer-32   NA   774  387
  Computer-66   NA    NA 1725
  Computer-28   NA    NA   NA

, , region = Quebec, country = Usa

             month
prod           May April June
  Computer-58  NA    NA   NA
  Computer-72  NA    NA   NA
  Computer-33  NA    NA   NA
  Computer-26  NA    NA   NA
  Computer-32  NA    NA   NA
  Computer-66  NA    NA   NA
  Computer-28  NA    NA   NA
```

Figure 9. View of the query result

I. CONCLUSION

The semantic extension of OLAP is proved to be fully functional using the toolchain of domain ontology, FluentEditor and the distributed knowledge representation system, Ontorion combined with, e.g., Excel as a source of data and RStudio for OLAP. Moreover, it created the foundations for already available on the market, developed and maintained by Cognitum, a solution called Ask Data Anything (ADA) [19]. The ADA allows exploring data by using natural language directly, rather than by using CNL, therefore we classify ADA as a tool that allows to explore data with natural language.

The modern approach to BI called BigData, is currently understood to face the problem of *"(...) growing number of insights that are being produced by big data through automated forms of analysis (...) What happens to the thousands of insights that are being generated automatically by all of those nifty machine learning algorithms? How do they find their way to a person at the right time?"* [20]. Semantic OLAP as well as its successor called ADA proves that the problem can be solved with support of semantic technologies.

REFERENCES

[1] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. [Online]. Available: http://www.w3.org/-Submission/SWRL/ [retrieved: 1 june, 2015]

[2] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010. [Online]. Available: http://-dblp.uni-trier.de/db/journals/sigops/-sigops44.html#LakshmanM10

[3] S. Krishnan, *Programming Windows Azure*. " O'Reilly Media, Inc.", 2010.

[4] W3C. Rdf 1.1 primer. [Online]. Available: http://-www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/ [retrieved: 1 june, 2015]

[5] Franz Inc. (2010) AllegroGraph RDFStore Web 3.0's Database. [Online]. Available: http://www.franz.com/-agraph/allegrograph/

[6] O. Erling and I. Mikhailov, "RDF support in the virtuoso DBMS networked knowledge - networked media," ser. Studies in Computational Intelligence, T. Pellegrini, S. Auer, K. Tochtermann, and S. Schaffert, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, vol. 221, ch. 2, pp. 7–24. [Online]. Available: http://dx.doi.org/-10.1007/978-3-642-02184-8_2

[7] D. Dobrowolski and M. Lesnik, "Social graphs in acquiring knowledge," *Zeszyty Naukowe Uniwersytetu Szczecinskiego. Ekonomiczne Problemy Uslug*, no. 87, pp. 34–41, 2012.

[8] M. Musen, N. Noy, C. Nyulas, M. O'Connor, T. Redmond, S. Tu, T. Tudorache, J. Vendetti, and S. S. of Medicine, "Protege," 2010, [http://protege.stanford.edu]. [Online]. Available: http://protege.stanford.edu

[9] P. Haase, H. Lewen, and R. Studer, "The neon ontology engineering toolkit."

[10] Cognitum, "Fluent Editor 2014 - Ontology Editor." [Online]. Available: http://www.cognitum.eu/semantics/-FluentEditor/ [retrieved: 1 june, 2015]

[11] ——. Ontorion Semantic Knowledge Management Framework. [Online]. Available: http://www.cognitum.eu/-semantics/ontorion/ [retrieved: 1 june, 2015]

[12] P. Hitzler, M. Krotzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer," World Wide Web Consortium, W3C Recommendation, October 2009. [Online]. Available: http://www.w3.org/TR/owl2-primer/

[13] A. Wroblewska, P. Kaplanski, P. Zarzycki, and I. Lugowska, "Semantic rules representation in controlled natural language in fluenteditor," in *Human System Interaction (HSI), 2013 The 6th International Conference on*. IEEE, 2013, pp. 90–96.

[14] R. Gentleman and R. Ihaka. R lanugage. [Online]. Available: http://www.r-project.org/ [retrieved: 1 june, 2015]

[15] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," *ACM Sigmod record*, vol. 26, no. 1, pp. 65–74, 1997.

[16] Infotopics. Natural query language. [Online]. Available: http://www.infotopics.nl/infotopics-tableau-blog/entry/-project-stel-een-vraag-aan-tableau [retrieved: 1 june, 2015]

[17] D. Z. Meyer and L. M. Avery, "Excel as a qualitative data analysis tool," *Field Methods*, vol. 21, no. 1, pp. 91–112, 2009.

[18] RStudio, "Rstudio ide - a powerful and productive user interface for r." [Online]. Available: https://-www.rstudio.com/ [retrieved: 1 june, 2015]

[19] Cognitum. (2015) Ask data anything. [Online]. Available: http://techblog.cognitum.eu/2015/05/ask-data-anything.html [retrieved: 1 june, 2015]

[20] D. Woods. (2015) Why big data needs natural language generation to work. Forbes. [Online]. Available: http://-www.forbes.com/sites/danwoods/2015/07/09/why-big-data-needs-natural-language-generation-to-work/ [retrieved: 1 june, 2015]

 ISBN: 978-1-61208-420-6