Linked Data Analytics for Business Intelligence SMEs: a Pilot Case in the Pharmaceutical Sector

Barbara Kapourani
Critical Publics
4 Flitcroft St., London, WC2H 8DH, UK
barbara@criticalpublics.com

Eleni Fotopoulou, Anastasios Zafeiropoulos
Ubitech
Thessalias 8 & Etolias 10, 15231
Chalandri, Athens, Greece
{efotopoulou, azafeiropoulos}@ubitech.eu

Dimitris Papaspyros, Spyros Mouzakitis, Sotiris Koussouris
National Technical University of Athens,
DSS Lab,
9, Iroon Polytechniou str., Zografou,
Athens, 15780, Greece
{dpap, smouzakitis}@epu.ntua.gr, skous@me.com

ABSTRACT
The adoption of linked data concepts from SMEs, meshed up with sophisticated analytics and visualization techniques within an integrated environment, called the LinDA Workbench, appears to reduce the effort for the realisation of specific tasks within the company, by almost 50% in terms of time, while at the same time, it supports the re-design of existing business processes and introduces increased efficiency and flexibility. In this paper, we briefly present the initial findings of the Business Intelligence Analytics (BIA) pilot operation of the LinDA project, which concerns the Over-The-Counter (OTC) medicines liberalisation in Europe. It aims at examining the association among the usage of OTC medicines and pharmaceutical parameters, with other healthcare, socio-economic and political ones. Focus is given on the added value emerged, through the consumption and production of linked data for analysis purposes, as well as the challenges faced for the execution of such a pilot.

Keywords
Linked data, business intelligence analytics, SMEs, Over the Counter (OTC) medicines.

1. INTRODUCTION
It seems that the information era has given its place into the analytics one, where the only way for SMEs to manipulate and extract intelligence from the huge amount of data and information out there is the investment on innovative solutions. The LinDA project [2] is an effort towards that direction. It is a co-funded European project, under the FP7 framework, aiming to support the SMEs' efforts to effectively adopt Linked Open Data (LOD) in their pursuit of competitiveness, by providing a complete set of tools for publication, consumption, analysis and visualization of linked data in an easy, user-friendly way. In this paper we are shortly presenting the LinDA concepts through a Business Intelligence Analytics (BIA) scenario. In Section 2, a short reference to existing challenges and related work is provided, while Section 3 presents the tools composing the integrated LinDA environment. Section 4 provides a step-by-step presentation of a real life scenario, which concerns the Over-The-Counter (OTC) medicines liberalization in Europe. Next, in Section 5, the SME's added-value identified is shown, while Section 6 concludes the paper, with plans for future work.

2. CHALLENGES AND RELATED WORK
Challenges for Linked Data provisioning and consumption regard mainly the renovation, compilation, maintenance and update of proper, meaningful and high quality datasets that may be easily consumed via a set of tools, as well as the need to work with heterogeneous and high volume data sources in many cases [5][6]. The challenges can be split into three distinct categories: (a) reviewing the datasets and preparing them in a proper format, (b) interpreting or extracting knowledge from the data through interlinking, inferences, as well as analytics extraction, and (c) maintaining and updating the data regularly.

Several projects have handled parts of these challenges, delivering however standalone frameworks that do not provide a holistic workflow. Among such initiatives, the LOD2[1] project aims to contribute high-quality interlinked versions of public Semantic Web data sets, promoting their use in new cross-domain applications, by developers across the globe. The DIACHRON[2] project takes on the challenges of evolution, archiving, provenance, annotation, citation, and data quality in the context of LOD and intends to automate the collection of metadata, provenance and all forms of contextual information, so that data are accessible and usable at the point of creation and remain so indefinitely. SDI4APPS[3] handles the uptake of open geographic information through innovative services based on LOD. With regards to enabling Linked Data analytics, no holistic framework exists -to the authors' knowledge- that is able to consume Linked Data towards the production of analytics and produce output data interlinked with the input data [4]. Finally, it should be noted that large effort is also given towards the design of systems for the production of Big Data analytics taking into account the collection of data in a distributed way, without providing mechanisms for evaluating or improving the quality of the available data, or providing techniques for producing Linked Data prior to their processing, by the analytics tools [5][6].
The distinguishing characteristic of LinDA is that -compared to existing tools- it provides a complete open-source package of Enterprise Linked Data tools to quickly map and publish your data in the Linked Data Format, interlink them with other public or private data, analyse them and create visualizations.

3. LINDA WORKBENCH
As a result of the LinDA project, the LinDA Workbench [3] is the integrated environment, consisting of a set of tools that facilitates the manipulation of linked data towards the realisation of analysis. More specifically, it consists of: (a) the LinDA Transformation

[1] The LOD2 project, http://lod2.eu/

[2] The DIACHRON project, http://www.diachron-fp7.eu/

[3] The SDI4APPS project, http://sdi4apps.eu/

engine, a lightweight transformation to linked data tool, (b) the LinDA Vocabulary repository, for increasing the semantic interoperability for your data, (c) the LinDA RDF2Any, a tool for converting RDF to conventional data structures in order to be used by legacy applications, (d) the LinDA Query Builder and Query Designer, to easily navigate and query your data, (e) the LinDA visualization, to perform smart visualizations on linked data out-of-the box and (f) the LinDA Analytics package for running analytic processes against your data. In an easy, straightforward and user friendly way, the user can follow a simple 3-steps procedure in order to transform the data to RDF, link (and query) them with public and private endpoints, and at last, to analyse and/or visualize the resulted information. The 3-steps procedure is as follows:

Turn Data into RDF: Using the LinDA Transformation engine, users can publish their data as linked data in a few, simple steps. They can simply connect to their database, select the data table they want and make mappings to popular and standardized vocabularies. LinDA assists even more by providing automatic suggestions to the mapping through its Suggest API (Oracle).

Query / Link your Data: With the LinDA Query Builder and Query Designer, users can perform simple or complex queries through an intuitive graphical environment that eliminates the need for SPARQL syntax. With simple drag and drop functionality users can perform complex SPARQL queries and filtering, including interlinking with external SPARQL endpoints.

Visualize / Analyze your Data: LinDA Visualization and Analytic engines can help enterprise users gain insight from the data that the company generates. The added-value of LinDA visualizations and analytics, in comparison with traditional tools, is that it takes advantage of the enriched metadata contained within the Linked Data format to produce more meaningful visualizations. On top of that, users can gracefully link their data with any other private or public data, therefore realizing an ecosystem of data extractions and visualizations, which can be bound together in a dynamic and unforeseen way.

4. REAL-LIFE SCENARIO

In this section, a short presentation of a real-life scenario implemented within LinDA Workbench is presented, aiming at validating the provided functionality and showing the added business value that it can be gained. For this purpose, we are considering an SME operating in the Business Intelligence (BI) sector that is making use of the LinDA Workbench for realising part of its daily business operations. The considered SME is providing consultation services in a wide range of national and multinational enterprises, organizations and governments, helping them to build their communication strategy and uphold their decision making processes. This is achieved by employing consultants that gather and assess huge amount of data in a daily basis, from different sources, in different formats and by using a variety of tools. Usually, the volume, variety, velocity, the time-sensitivity, the heterogeneity and non-interoperability of the data to be handled is a great burden for the SME, in terms of effort, time, resources and complexity.

The pilot scenario is related with the Over-The-Counter (OTC) medicines liberalization in Europe, in terms of price, retail and entry. As stated in [1], the sales of the OTC medicines are increasing during the last years, despite of the global financial crisis, while at the same time, the role of the OTC in the pharmaceutical market is quite promising and prominent, opening new businesses opportunities and potentials for growth. Many studies have been conducted for the OTC market, revising different aspects of it, focusing only in a subset of the correlated parameters that affects or are being affecting by the OTC liberalization, which basically reveals the complexity of the domain.

With the usage of the linked data approach, through the LinDA BIA pilot, we have tried to compose a complete, cumulative, conceptualization map of all the correlated parameters for the OTC market; trying to fill the gap that the rest of the studies left. The objective is to investigate the OTC liberalization effects in various parameters (e.g. healthcare expenditures, OTC revenues) by studying and analyzing the countries where the liberalization has already taken place, and by combining this information with the unique parameters of the rest European countries of interest, where the liberalization has not been yet implemented. Based on this conceptualization map, and by using the LinDA Workbench, we have tried to identify the indicators that play a significant role in the OTC liberalization, to find their correlations and to analyze their impact, aiming at getting business intelligence insights that can be proven helpful towards decision-making processes.

4.1 Scenario Implementation

Figure 1 presents the concrete steps followed for the implementation of the OTC scenario within the LinDA framework.

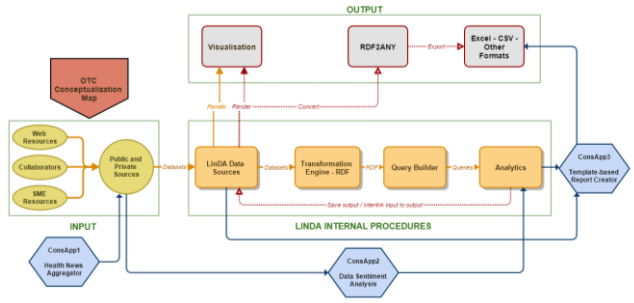

Figure 1. OTC scenario implementation steps in LinDA

As a result of an extensive scientific research and an intelligence-gathering mechanism, conducted manually by the involved actors, spanning from health advisors to business intelligence consultants, the conceptualization map of the OTC scenario has been crafted, acting as the basis of the data needed for composing the full picture of the correlated parameters for the OTC market. The identified public and private datasets of this map have been classified into five main categories: (a) healthcare indicators, (b) OTC indicators, (c) economic indicators, (d) political indicators and (e) social indicators. Overall, the map contains more than twenty unique indicators, half of which represents private datasets of the SME that have been created for the purpose of the OTC scenario. Furthermore, the indicators are spanning in various time ranges and across different European countries. Most of these datasets are in excel or csv format, while some of the public sector datasets have SPARQL endpoints to be retrieved from (e.g. Eurostat, Worldbank, Transparency.org).

Based on the conceptualization map and the acquired datasets, two major interlinking directions have been followed in the BIA pilot's scenario; in the first case (Figure 2) all the private datasets have been interlinked with the World Factbook (http://wifo5-03.informatik.uni-mannheim.de/factbook/), both by country and by year property. In the second interlinking case (Figure 3) some of the identified private and public datasets have been interlinked

together, targeted at being used towards the examination of open issues of the OTC scenario.

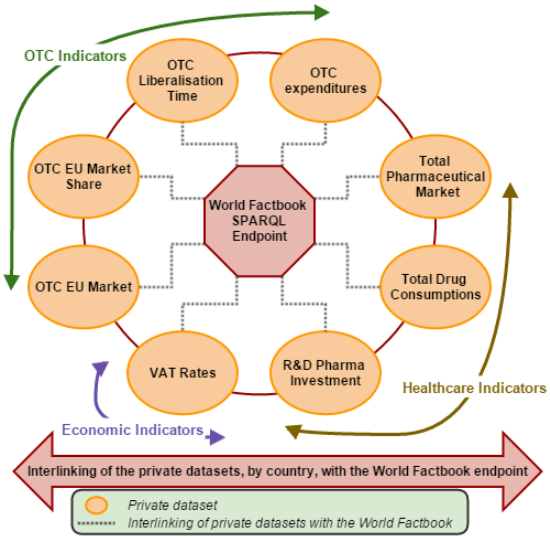

Figure 2. OTC scenario – private datasets interlinking

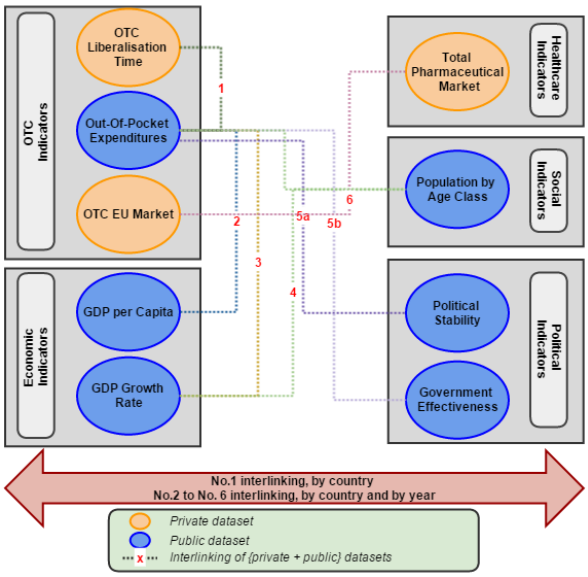

Figure 3. OTC scenario private & public datasets interlinking

After the datasets preparation and their interlinking, the SME has used the LinDA tools (e.g. Transformation Engine, Query Designer) for uploading/saving the datasets, transforming them into RDF format and formulating the interlinking questions to be used next, into the analysis phase. The LinDA Analytics tool is then being used for drilling down to the study, which is the fundamental step towards the intelligence extraction process.

Initially, a set of regression analyses are realised for examining the relationship among the OTC expenditures variation with several parameters. The first regression analysis regards the relationship between OTC health expenditures and the GDP per capita of each country. Based on the produced results, it can be claimed that increase in the GDPpc by 10 US$ leads to reduction in the percentage of OTC expenditures by 1.06%. This reduction can be attributed to the better social care that can be associated with larger GDPpc values (e.g. through an increase on the public expenditure on health) and thus less need for using OTC. The next step in the analysis regards the examination of the relationship among the OTC expenditures and the GDP growth rate. A regression analysis is realized, leading, however, to very small adjusted R-squared value and thus considering that the variation of the GDP growth rate explains a very small percentage of the variation of the OTC expenditures. Following, a set of analysis are realised, examining the relationship among OTC expenditures and governmental indicators, specifically political stability and governmental effectiveness indicators. Based on the produced results, it seems that higher political stability and higher governmental effectiveness lead to reduction in the OTC expenditures. It could be claimed that in countries with political stability and high quality of the governmental institutions, the social care is advanced and the allocated budget on healthcare is associated with effective results, leading to reduction on the need for usage of OTC medicines. However, the analysis has to be evolved in order to acquire further insights on this aspect.

Furthermore, a clustering analysis is realised upon an interlinked dataset (linking parameters such as government effectiveness, political stability, GDP per capita and OTC expenditures per country and per year). It aims to provide a set of indications regarding the existence of clusters based on the examined parameters and the investigation of the composition of such clusters. A k-means clustering algorithm is executed for the partitioning of the overall observations in three clusters. The results are depicted in Figure 4.

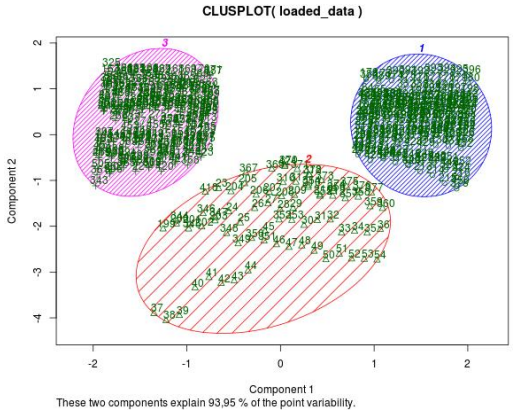

Figure 4. OTC scenario – K-means clustering results

Three really well defined clusters with no overlapping among each other are produced. The first cluster refers to countries with low political stability, low governmental effectiveness and low GDP per capita that tend to have high OTC expenditures. The second cluster refers to countries with medium-to-high political stability, high governmental effectiveness and high GDP per capita that tend to have low OTC expenditures. The third cluster refers to countries with medium political stability, low governmental effectiveness and low GDP per capita that tend to have medium to low OTC expenditures.

It should be noted that the afore-mentioned results led to initial insights based on the first phase of the BIA pilot's execution, while further analysis is envisaged to be realised the upcoming period. Furthermore, the analysis performed includes also a set of visualisations produced by the corresponding LinDA tool that provide initial views and insights with regards to the evolution of specific parameters.

5. SME'S ADDED VALUE WITH LINDA

In order to accurately evaluate the business value introduced to the SME, by the usage of the LinDA Workbench, we have introduced a simple example, where four parameters (e.g. GDP per capita, OTC expenditures, unemployment and political stability) have to be acquired, transformed, analyzed and visualized, in a weekly basis, with and without the usage of LinDA tools. The evaluation takes into consideration the execution time needed, the people engaged and the tools employed. Figure 5 presents in detail the SME's four core business phases, as well as the tasks, resources and effort that are needed when the example is implemented "as usual" and next, when it is implemented with the help of the LinDA Workbench.

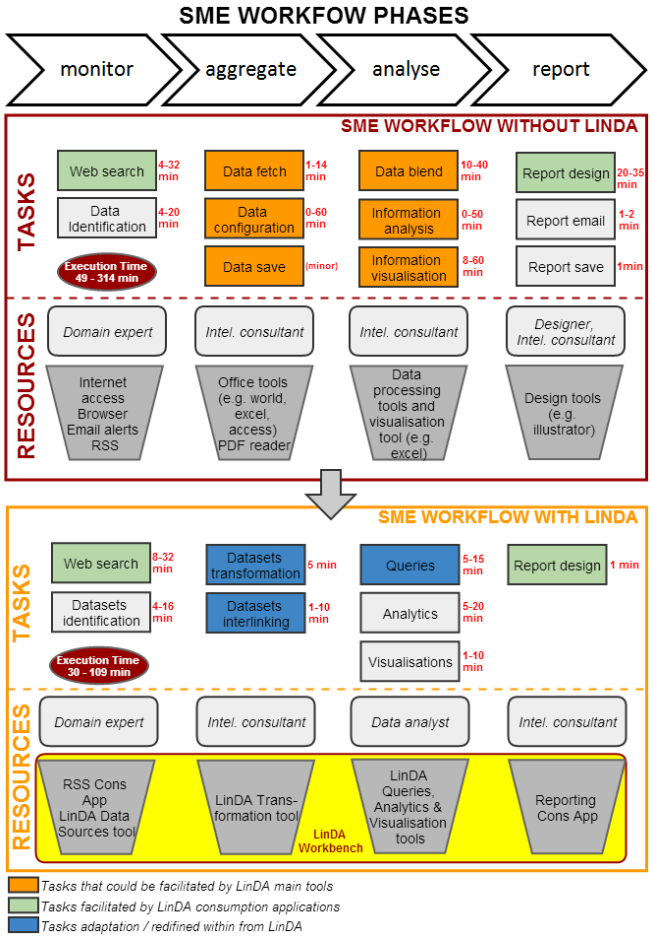

Figure 5. SME workflow, with and without LinDA

It has been noticed that the time needed for the completion of the example is reduced with the usage of LinDA of about 20min to 3,41hours per realisation of a daily analytics process. Moreover, the tasks executed are also reduced from 11 to 9. As for the involved experts, changes have been detected; with LinDA a new actor is introduced, the data analyst, while the designer that is needed in the "as usual" case is taken out of the scene. The last is quite significant, because it is closely related with the reduction of the outsourcing costs of the report's visual design. Furthermore, what is quite vital is that with LinDA all the tasks are realised within the unified environment of the LinDA Workbench. This may play the role of the "one-stop-shop", with a quite user friendly and straightforward UI, sidestepping the need of different tools and distributed, autonomous working environments, which is until now the case of the SME working status.

Overall, it is clear the added-value coming from the usage of the linked data and the LinDA Workbench. It has been calculated that the linked data concept could definitely assist and improve the access to more and instant information, while the usage of the LinDA Workbench has proven to facilitate the SME's current workflow and to reduce the effort required for specific tasks, in terms of time and resources. More importantly, LinDA has introduced new business value for the SME (e.g. analytic process which were not implemented before, due to lack of experience or time), which can boost the provided services to new levels of completeness and sophisticated analysis and visualizations.

6. CONCLUSION

In this paper we have presented the workflow followed by an SME, operating in the business intelligence sector and providing consultation services to its clients, with and without the usage of the LinDA Workbench, so as to derive insights about the usefulness of the adoption of linked data concepts. The current results, of the first pilot round, are quite promising, showcasing that the tasks are performed faster, easier and with less human resources, while new business value is also introduced. However, there are challenges that need to be further investigated, such as the need for evaluating the quality of the considered data, the need for optimally transforming business data to linked data through the usage of specific vocabularies, as well as the mentality shifting within the SMEs, towards the adoption of a more "data-scientist" oriented workflow, along with the organization of the corresponding training activities.

7. ACKNOWLEDGMENTS

This work has been co-funded by the LinDA project, a European Commission research program under Contract Number FP7-610565.

8. REFERENCES

[1] Tisman, A. 2010. The Rising Tide of OTC in Europe, *IMS Health*.

[2] The LinDA project, http://linda-project.eu/

[3] The LinDA Workbench, http://linda.epu.ntua.gr/

[4] Fotopoulou, E., et al. 2015. Exploiting Linked Data Towards the Production of Added-Value Business Analytics and Vice-versa, *DATA 2015*, Colmar, Alsace, France, July 2015.

[5] Networked and Electronic Media Initiative, Big and Open Data position paper, December 2013.

[6] UN Global Pulse, White Paper: Big Data for Development – Challenges and Opportunities, May 2012. Accessed on 30.01.2015 at http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

[7] Hu, H., Wen, Y., Chua, T. and Li, X. 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, *IEEE Access*, vol. 2, pp. 652–687, 2014.