Exploiting Linked Data Towards the Production of Added-Value Business Analytics and Vice-versa

Eleni Fotopoulou[1], Panagiotis Hasapis[1], Anastasios Zafeiropoulos[1], Dimitris Papaspyros[2],
Spiros Mouzakitis[2] and Norma Zanetti[3]
[1]Ubitech Ltd., Thessalias 8 & Etolias 10, Chalandri, Athens, Greece
[2]School of Electrical and Computer Engineering, National Technical University of Athens,
Heroon Polytechneiou 9, Zografou, Athens, Greece
[3]Hyperborea Srl., Via Giuntini 25, Navacchio (Pi), Italy

Keywords: Linked Data, Open Data, Business Analytics, Data Mining, Ontology.

Abstract: The majority of enterprises are in the process of recognizing that business data analytics have the potential to transform their daily operations and make them extremely effective at addressing business challenges, identifying new market trends and embracing new ways to engage customers. Such analytics are in most cases related with the processing of data coming from various data sources that include structured and unstructured data. In order to get insight through the analysis results, appropriate input has to be provided that in many cases has to combine data from cross-sectorial and heterogeneous public or private data sources. Thus, there is inherent a need for applying novel techniques in order to harvest complex and heterogeneous datasets, turn them into insights and make decisions. In this paper, we present an approach for the production of added-value business analytics through the consumption of interlinked versions of data and the exploitation of linked data principles. Such interlinked data constitute valuable input for the initiation of an analytics extraction process and can lead to the realization of analysis that was not envisaged in the past. In addition to the production of analytics based on the consumption of linked data, the proposed approach supports the interlinking of the produced results with the associated input data, increasing in this way the value of the produced data and making them discoverable for further use in the future. The designed business analytics and data mining component is described in detail, along with an indicative application scenario combining data from the governmental, societal and health sectors.

1 INTRODUCTION

During the last years, there is evident an increasing trend in the produced volumes of business data, both structured and unstructured. This trend is evolving, since, as stated in a study from eBay (eBay report, 2013), the volume of business data worldwide across all companies is estimated to double every 1.2 years. This continuous increase in the produced data across various business sectors is also referred as "datafication" of the world (Kreissl, 2013).

However, "datafication" cannot be automatically associated with new insights and advances in our understanding of business data and the production of knowledge (Lausch, 2015). In order to gain insight, available data has to be appropriately represented -in many cases upon transformation and harmonization processes- and be processed both statistically and analytically (Lausch, 2015) towards the production of advanced analytics. It should be noted that, according to a report from Gartner, more than 30 percent of analytics projects is envisaged to deliver insights based on a combination of structured and unstructured data by 2015 (Gartner report, 2013).

The need for the development and adoption of novel analytics tools is already identified from the interested business stakeholders. Indicatively, it can be noted that advanced analytics is the fastest-growing segment of the business intelligence and analytics software market (Gartner report, 2014). According to the Intel's annual IT business review (Intel report, 2015), the use of business intelligence and analytics tools increased Intel's revenue by USD 351 million, while, based on a survey from the International Data Corporation (IDC report, 2014), the business analytics software market is expected to grow at a 9.4% Compound Annual Growth Rate (CAGR) over the next five years.

Fotopoulou E., Hasapis P., Zafeiropoulos A., Papaspyros D., Mouzakitis S. and Zanetti N..
Exploiting Linked Data Towards the Production of Added-Value Business Analytics and Vice-versa.
DOI: 10.5220/0005508700690080
In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (DATA-2015), pages 69-80
ISBN: 978-989-758-103-8

Nevertheless, in parallel with the development of advanced analytics platforms, inhibitors towards the integration of analytics tools and their daily usage with smaller companies, such as the current complexity in the usage of such tools and the necessity for existence of specialized personnel, have to be taken into account. Only if future research succeeds in extracting information from heterogeneous and distributed data in a user-friendly and effective way, it can constitute the basis for good operative and strategic business decisions and projections.

Thus, it could be argued that the following two challenges have to be faced towards the realization of advanced analysis and evidence-based decision making that optimize results for small scale businesses and organizations: (i) the design of advanced but user-friendly analytics tools that can be easily integrated within the daily business processes of organizations and enterprises and (ii) the adoption of techniques that permit the production and consumption of combined datasets that were previously closed in disparate sources and can now be appropriately interlinked.

Taking into account the aforementioned challenges, in this manuscript we present an approach for the exploitation of linked data principles towards the production of added-value business analytics, as it is developed within the framework of the FP7 project LinDA (http://linda-project.eu/). Linked data are about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. They regard actually a set of best practices for representing and connecting structured information on the web, which can be proven enablers towards the path to advanced analytics.

The production of linked data analytics has two notions in the proposed approach: (i) the interlinking of datasets prior to the realization of analysis, targeting at the preparation of datasets that can lead to unexpected and unexplored insights that were not possible previously and (ii) the interlinking of the analytic results output and the analysed input datasets for enriching the information at the input datasets in a clean and straightforward way. Thus, we refer to exploitation of linked data through the production of added-value business analytics and the exploitation of the produced business analytics for the production of high quality five star linked data (Zaveri, 2015).

The designed and developed business analytics and data mining tool –including its functional architecture and the developed ontology for the support of linked data analytics- as well as a pilot application scenario are presented. In more detail, the structure of the paper is as follows: in section two, the trends and challenges faced for the production of advanced business analytics are detailed along with the added value expected via the production and consumption of linked data and the positioning of the proposed approach with regards to related work in the era; in section three the proposed architectural and technological approach is described; in section four a pilot application scenario is presented while section five concludes the paper by referring to the path towards the adoption of such a tool by businesses as well as plans for future work.

2 BUSINESS ANALYTICS: CURRENT TRENDS AND NEEDS

2.1 Challenges and Trends

Advanced analytics are considered as a top business priority, since it enables business organizations to differentiate, significantly impact business performance and enhance their position compared to their competitors (Gartner report, 2014) (IBM report, 2014). Towards this direction, a new class of enterprises is emerging -called Generation D (Figure 1)- that routinely leverage varied data sources, including structured and unstructured data, and apply prescriptive analytics to optimise most processes and undertake strategic business decisions (IBM report, 2014). Thus, the current trend regards the development of advanced and self-service analytics tools that are going to facilitate enterprises to harvest complex and heterogeneous data, turn them into insights and make decisions.

As also stated in the introductory part of this manuscript, a set of challenges exist and have to be tackled towards the transition to a next generation of analytics tools. Given that these tools have to be made accessible by multiple end users –including data scientists and decision makers- the overall complexity in their usage, the overhead for the consumption of complex data, as well as the requirements for existence of personnel with specialized skills have to be minimized. Combination of data with sophisticated analytics have to be supported, providing the capacity for producing results leading to evidence-based answers and high business value.

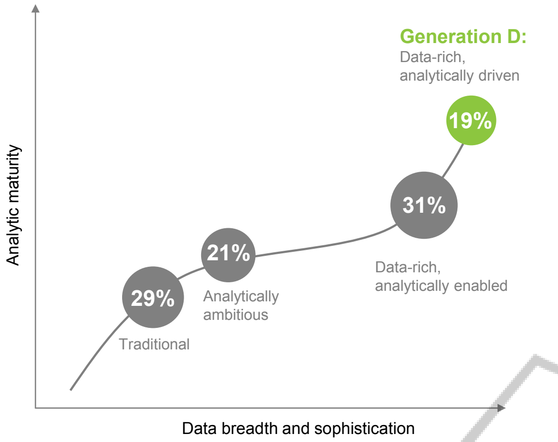

Figure 1: Evolution of enterprises with regards to analytics (IBM report, 2014).

Furthermore, in order to support the production of advanced analytics within the daily business processes of enterprises, analytics tools have to be extensible and support interfaces for interaction with existing or newly developed business applications. Business analytics have to lead to automation of processes and decisions, provide the capacity for identification of market trends, risks and new market opportunities. Real time analytics have to be supported in many cases, requiring access to the available data and resources when they are needed. Finally, scalability and performance issues have also to be tackled, since in many cases there is a need for processing of big data volumes or the production and storage of big data analytics (Hu, 2014).

2.2 Why Linked Data Analytics?

2.2.1 Consumption of Linked Data for Production of Business Analytics

Current analytics tools and data mining techniques are mostly restricted to isolated or “silo” datasets and therefore remain primarily stand alone and specialized in nature (Lausch, 2015). In many cases, such datasets impose limitations towards answering complex and interdisciplinary questions that require the establishment of association links among similar entities and concepts that are present in different datasets. In order to establish such associations, the corresponding information has to be represented in a format that is accessible and without any ambiguities regarding its structure and semantics.

This is where linked data principles come. The term linked data refers to a set of best practices for publishing and interlinking structured data on the Web. By following these practices, data from diverse sources can use the same standard format that allows them to be combined and integrated. Linked data specify that all data will be represented based on the Resource Description Framework (RDF) specifications. Conceptual description of data is realized based on specific vocabularies (and thus semantics) accessible over HTTP, allowing the user to interpret data from multiple vocabularies and query them in a uniform manner.

By adopting linked data principles, a set of advantages are provided towards the production of advanced business analytics. Combination of data from multiple and in many cases distributed sources, as well as from publicly available data (e.g. open governmental data) or privately owned data maintained by enterprises, can help businesses enhancing their experience of managing and processing of data, in ways not available before. Actually, linked data provide the capacity for establishing association links among concepts in different datasets, producing high-quality interlinked versions of semantically linked web datasets and promoting their use in new cross-domain applications by developers across the globe. Such interlinked datasets constitute valuable input for the initiation of an analytics extraction process and can lead to the realization of analysis that was not envisaged in the past.

Moreover, interlinked versions of datasets may be maintained or regularly updated facilitating the provision of access to latest versions of the available datasets, characteristic that is very helpful in cases of data with high volatility. Interoperability and re-use of linked data reduces significantly the data management overhead imposed to data scientists for maintaining up-to-date databases.

Linked data also provide mechanisms for assessing the quality of the available data, since there is a need to handle data with different quality characteristics regarding their accuracy, consistency, timeliness, completeness, relevance, interpretability and trustworthiness. Data quality assessment is very useful prior to an analysis, since it provides information that can be helpful during the interpretation of the analysis results (Zaveri, 2015).

2.2.2 Production of Linked Data through Business Analysis

In addition to the business value produced through the exploitation of linked data for the production of business analytics, the interlinking of the analysis results with the related input used for the analysis has also a set of advantages.

The value of the produced data is increased since, by appropriately interconnecting them with the input data, detailed information is provided to analysts regarding the exact datasets used in the analysis as well as the details of the process followed during the analysis. In cases of analyses that have to be updated in a periodical way and the results of the analysis highly depend on the instance of datasets used or in cases where comparison among the analysis in diverse time periods is required, such interlinking can provide further business value to the available data.

Moreover, with linked data analytics, the produced data are made discoverable for further private or public use in the future by the enterprise employees, clients or other data consumers. In case of public data, other data publishers may also link their data or analytics with them, adding further business value. Actually, data are accessible and usable at the point of creation and remain so indefinitely. Quality of the produced data may be also evaluated according to linked data evaluation criteria while the evolution of the datasets may be monitored.

2.3 State of the Art Analysis and Progress beyond

Several open source data mining tools are available today. The most commonly used include R (R project, 2015), Weka (Weka tool, 2015), Knime (Knime tool, 2015) and RapidMiner (RapidMiner tool, 2015). A comparative analysis of such tools is provided in (Hirudkar, 2013) and (Lausch, 2015) in terms of supported functions, efficiency, hosting platform needs, supported input formats, user friendliness etc. Furthermore, in (Lausch, 2015), a set of requirements are presented that have to be fulfilled by data mining and analytics tools in order to support future interdisciplinary research. These requirements include the need for linking various data mining types, the need for exploitation of available open data stemming from various areas, the need for use of open source software and the need for use of data mining tools that do not require a big time investment and in-depth programming knowledge (Hirudkar, 2013).

The distinguishing characteristic of the proposed approach in this paper –as compared to the state of the art tools- is the introduction of the notion of linked data analytics. It should be noted that – to the best of the authors' knowledge- no analytics framework exists that supports consumption and production of linked data (e.g. there is no support for RDF as an input/output format). Focus is given on the integration of existing open-source tools – namely R and Weka- into a holistic platform that facilitates the consumption and production of linked data, taking into account and fulfilling the afore-mentioned requirements for supporting future interdisciplinary research (Lausch, 2015).

3 LINKED DATA ANALYTICS ECOSYSTEM

In this section, we provide detailed description of the proposed approach for the production and consumption of linked data business analytics. As the designed tool consists part of an ecosystem (the LinDA workbench, available at http://linda.epu.ntua.gr/) aiming to assist enterprises in efficiently developing novel data analytical services that are linked to the available public data, in the first subsection we provide a short description of the overall ecosystem. Following, in the second subsection, detailed description of the business analytics and data mining component is provided, including its internal architecture and the supported functionalities, the designed Linked Data Analytics Ontology (LDAO) for the appropriate representation of an analytics process combined with the considered interlinking policy, the analytics production workflow based on a categorization of the supported algorithms, as well as the adopted performance evaluation framework.

3.1 The LinDA Workbench and Workflow

From a user perspective, the main LinDA workflow can be summarized in three simple steps, as illustrated in Figure 2. More specifically the three workflow steps are:

Step 1 – Explore Datasets/Turn Data into RDF: Using the LinDA toolset, users can publish their data as linked data in a few, simple steps. In cases where the data are not available in RDF format, the users can simply connect to their database(s), select the data table they want and make their mappings to popular and standardized vocabularies. LinDA assists even more by providing automatic suggestions to the mapping process. Based on the defined mappings, transformation from various formats (e.g. csv, relational database) to RDF is realised.

Step 2 - Query/Link Your Data: With the LinDA toolset, users can perform simple or complex queries through an intuitive graphical environment that eliminates the need for SPARQL Protocol and RDF Query Language (SPARQL) syntax. In addition to the submission of queries, interlinking of instances is supported, where the designer lets the end user ignore its instance's data source and handle instances as if they were defined in the same data source. The possible types of interlinking vary according to the interlinking element that is used. More specifically, classes and object/datatype properties can be combined in a versatile way, during the interlinking procedure. Hereinafter, for the sake of homogeneous representation, all interlinking endpoints will be referred as interlinking types. The interlinking of instances of the types [A] and [B] can occur in several ways: (i) instances can be interlinked directly to each other, in which case an entity (URI) is fetched in the query results if belongs to both types [A] and [B] at the same time; (ii) an instance of type [A] can be interlinked to an instance of type [B] via a property, where [A].p is bound to be an instance of type [B] ("owl:same-as" interlinking) and (iii) instances can be interlinked by their properties, where [A].p = [B].q, given that [A].p and [B].q refer to the same URI or that [A].p and [B].q are literals (strings, numbers, dates etc.) with the same value.

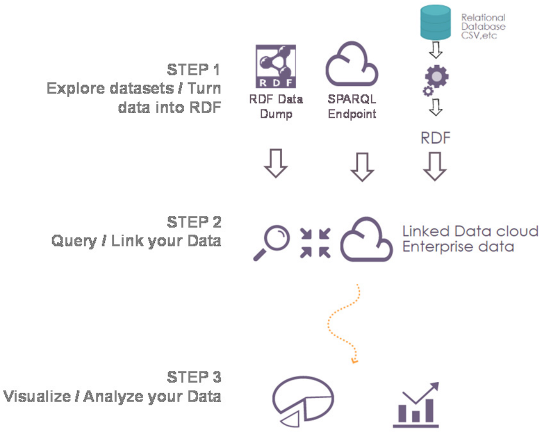

Figure 2: The LinDA Workflow.

Step 3 - Visualize/Analyse Your Data: the LinDA toolset can help enterprise users gain insights from the data that the company generates or consumes through the support of a set of visualization and analytics services. LinDA supports visualisations over different categories of data, e.g. statistical, geographical, temporal, arbitrary data, as well as a largely automatic visualization workflow for matching and binding data to visualizations. As far as the analytics services are concerned, they are presented in detail in the following subsection.

According to this workflow, the user can utilize either external public data or internal, private sources. If the initial data source is in RDF format, the user can directly insert the data source to the available data sources of the LinDA Workbench. If the initial data source is in another format (relational database, csv, etc.), the LinDA Workbench guides the user to the toolset responsible for transformations in order to transform the data into the RDF format, with the utilization of popular linked data vocabularies. Once in RDF, the user can then visit the list of data sources and activate one of the available LinDA services. More specifically, the user has the option to a) visualize the selected RDF data source, b) analyse it, c) query it and d) edit/update/delete it.

3.2 Business Analytics and Data Mining Component

The business analytics and data mining component supports the realization of analysis based on the consumption and production of linked data. A library of basic and robust data analytic functionality is provided through the support of a set of algorithms, enabling enterprises to utilize and share analytic methods on linked data for the discovery and communication of meaningful new patterns that were unattainable or hidden in the previous isolated data structures.

3.2.1 Architecture and Analytics Workflow

The business analytics and data mining component is based on an extensible and modular architecture that facilitates the integration of algorithms on a per request basis. The development of the component is based on open-source software while integration of algorithms is based on open-source analytics projects.

The business analytics and data mining component consists of the following sub-components: the Query selection component, the Algorithm selection and configuration component, the Algorithm execution component and the Linked data analytics management component, as they are illustrated in Figure 3.

The Query selection component is responsible for processing the output upon the execution of simple or complex queries and loading it as input for the initiation of an analysis process. For this purpose, appropriate interconnection interfaces with

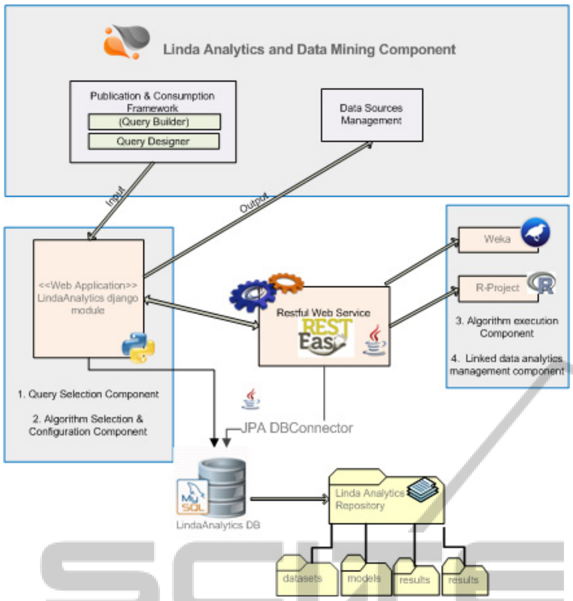
Linda Analytics and Data Mining Component
Publication & Consumption Framework
(Query Builder)
Query Designer
Data Sources Management
Input
Output
<<Web Application>> LindaAnalytics django module
1. Query Selection Component
2. Algorithm Selection & Configuration Component
Restful Web Service
REST Easy
Weka
R-Project
3. Algorithm execution Component
4. Linked data analytics management component
JPA DBConnector
LindaAnalytics DB
Linda Analytics Repository
datasets
models
results
results

Figure 3: Business analytics and data mining component's architecture.

the tools that serve the design and execution of SPARQL queries are designed and implemented. In order to facilitate the user to select a meaningful and coherent set of data, this component provides some hints based on the existing usage of the dataset within the LinDA workbench. The provision of such hints may be helpful for the end users in order to select the appropriate dataset and parameters for the realisation of the envisaged analysis.

The Algorithm selection and configuration component coordinates the process of suggesting, selecting and configuring an analytics extraction process in a user friendly and quite explanatory way. Selection is realized based on a categorization of the supported algorithms in classification, association, regression/forecasting, clustering and geospatial analysis algorithms. For each category, a set of algorithms is supported along with their default configuration and guidelines regarding their proper usage. Based on the category of the selected algorithm, a specific workflow is being followed by the end user with guidance relevant to the appropriate configuration of the algorithm parameters' setup. For instance, part of the algorithms are based on one input dataset used for the analysis, while another part necessitates the existence of a training and an evaluation dataset (e.g. in cases of classification analysis). In the latter case, a further query is selected to be used for fetching the evaluation data. Finally, the output format is selected based on a list of supported formats per algorithm.

Upon finalizing the configuration phase of the algorithm, the next step regards the usage of the Algorithm execution component. This component realizes the execution of the analytics process. It should be noted that, at the current phase, integration of algorithms from the Weka open-source tool (Weka tool, 2015) and the R open-source project for statistical computing (R project, 2015) is realized. Thus, execution is based on the associated tools per algorithm, while the overall process is transparent to the end user.

The produced business analytics are then handled by the linked data analytics management component. This component is responsible for providing the output of the analysis in the appropriate format (e.g. RDF, text, graphs) in order to be meaningful for the end user, as well as easily exploitable for further usage (e.g. for visualization purposes). This component realizes also the interlinking of the input and output datasets, based on the defined interlinking policy, as it is described in the following subsection. Detailed information regarding the analytic process realized, the queries/datasets used for the analysis, a set of performance evaluation metrics as well as access to the script used in the associated open-source tool for the execution of the algorithm is provided. The performance evaluation metrics –as detailed in section 3.2.4- consist of a set of quantitative and qualitative metrics. The first ones are automatically computed by the data analytics management component, while for the latter ones, upon the execution of the analytics process, the user is kindly suggested through a star-based classification to evaluate the analytics result in terms of meaningfulness, produced added value, quality of experience etc. The computation of these metrics is used towards the provision of a set of recommendations to end users at the initiating phase of future analyses.

Furthermore, within the business analytics and data mining tool, customized information per user of the tool is provided regarding the realised analysis in the past. Such information may be proven helpful in the interpretation of changes in the results (meta-analysis of the analytics results), identification of trends (e.g. which are the most popular algorithms utilised by end users per business sector) as well as identification of significant changes at the input data (e.g. upon an update with newly collected data).

With regards to the technologies used for the development of the business analytics and data mining component, the front-end is based on the Django Python Web Framework, while the Query selection and the Algorithm selection and configuration components are developed as a

Django module, pluggable in every Django installation. The Algorithm execution and the Linked data analytics management components are implemented as a RESTful web service that is consumed by the Django analytics module. The aforementioned RESTful web service is responsible for realising the analytics and data mining part and offers a loosely coupled connection with the overall Linda workbench, permitting its easy integration to third party software. The Java-based RESTful web service is developed with RESTEasy, an implementation of the JAX-RS specification and deployed by using a JBoss server.

3.2.2 Algorithms Categorization and Integration

The supported algorithms in the business analytics and data mining component aim to support a variety of cross-sectorial studies and business needs. In order to achieve it, a categorization of the supported algorithms is provided along with the integration of a limited set of algorithms per category, as shown in Table 1. The supported set of algorithms per category is based on the identified needs of the LinDA pilot cases. However, as already mentioned, the integration of further algorithms is supported, especially from the Weka and R tools.

Table 1: Supported algorithms per category.

Category	Supported Algorithms	Tool
Classification	J48 (decision making) M5P (piecewise linear fit to the dependent variable)	Weka
Association	Apriori (decision making)	Weka
Clustering	KMeans - Partitioning (unsupervised learning) Ward Hierarchical Agglomerative (unsupervised learning) Model Based Clustering (unsupervised learning)	R
Regression	Linear/Multiple linear regression (identify relationships and trends)	Weka
Forecasting	Arima (market analysis trends and seasonality patterns)	R
Geospatial	Morans I (detect spatial autocorrelation) Kriging (describe spatial autocorrelation) NCF correlogram (describe spatial autocorrelation)	R

The goal of classification analysis algorithms is to find functions and models in order to identify to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The goal of association analysis algorithms is to discover interesting relations between variables in large datasets. It is intended to identify strong rules discovered in datasets using different measures of interestingness (Piatetsky-Shapiro, 1991). The identified relations can be quantified, while, by using specific indicators (e.g. support, confidence and lift values), the strength of the identified associations can be evaluated. The goal of clustering analysis algorithms is to group a set of objects in such a way that objects in the same group (called a cluster) are more similar (according to a set of metrics) to each other than to those in other groups (clusters). As far as the regression analysis algorithms are concerned, the goal is to model the relationship between a scalar dependent variable and one or more explanatory independent variables. Forecasting algorithms are also supported that facilitate the process of making statements about events whose actual outcomes (typically) have not yet been observed. Finally, the goal of geospatial analysis algorithms is to apply statistical analysis and other informational techniques to data which has a geographical or geospatial aspect.

As already stated, high priority is given to the user friendliness of the provided interfaces based on the design of specialized workflows per algorithm category.

3.2.3 Linked Data Analytics Ontology and Interlinking Policy

In order to support linked data analytics, an ontology is being designed aiming at the representation of the interlinking among input and output datasets, as well as the conceptual representation of the overall analytic process. Main benefits from the usage of such an ontology include the consolidation of a unified schema of terms and content relationships that describe the business analytics realisation process, the mapping of the acquired information in this schema and the usage of queries to flexibly investigate the available content and the capacity to maintain quality and trace changes made over time.

The designed Linked Data Analytics Ontology (LDAO) is based on a set of existing ontologies along with specific extensions (LDAO, 2015). Specifically, LDAO inherits concepts from the Friend Of A Friend (FOAF) ontology (Brickley, 2014) and the PROV ontology (Lebo, 2013). Concepts of *Agent, Person, OnlineAccount* and *Homepage* from the FOAF ontology and concepts of

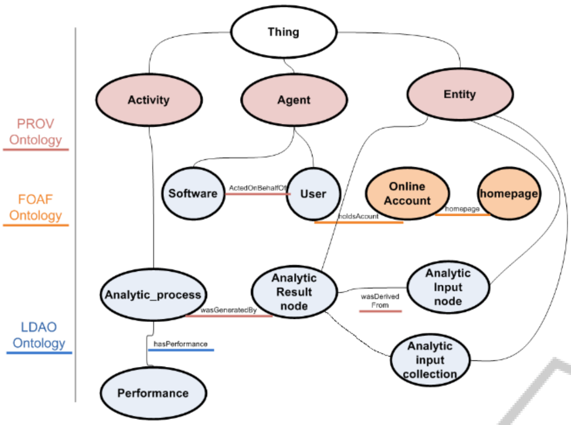

Figure 4: Linked Data Analytics Ontology.

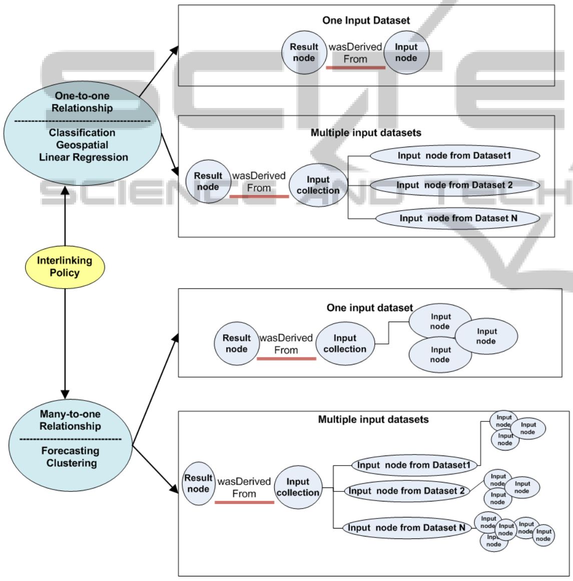

Figure 5: Interlinking policy.

Agent, Activity and *Entity* from the PROV ontology are used. Specific extensions are provided for the representation of the analytics input and result node, the analytic process along with performance characteristics. The concepts and relationships included in the LDAO ontology are shown in Figure 4.

In addition to the design and specification of the LDAO ontology, an interlinking policy is implemented for the creation of linked data business analytics. Different types of interlinking are supported taking into account the peculiarities of each type of algorithm, as illustrated in Figure 5. These types can be classified in "one-to-one relationship" and "many-to-one relationship" interlinking. In the first case, there is "one-to-one relationship" among each analytics result node with an analytics input node, while in the latter case there is "many-to-one relationship" among each analytics result node with a set of analytics input nodes. In the "one-to-one relationship", we refer to classification, geospatial and linear regression algorithms, while in the "many-to-one relationship", we refer to forecasting and clustering algorithms. In both cases, the input dataset can be derived from one or multiple input data sources, depending on the executed SPARQL query and the associated sources. In case of multiple data sources, the direct relationship(s) is realised on the interlinked data set, however references to the input data sources are provided.

3.2.4 Performance Evaluation Framework

High importance is given on the evaluation of the performance of the developed component with a twofold objective: the evaluation of the efficiency with regards to data management and algorithms execution as well as the evaluation of the overall usability of the component. It should be noted that several studies exist for evaluating the performance of data mining techniques (Hirudkar, 2013) (Jia, 2014). We are focusing on the differentiation in the performance that is related with the consumption and production of linked data within the analytics process. The performance evaluation is based on an assessment model that includes a set of quantitative and qualitative metrics extracted mainly from the ISO/IEC 25010:2011 standard (ISO/IEC 25010, 2011).

The quantitative metrics refer to a set of indicators for evaluating the efficiency in handling different linked data workloads per algorithm category and the time required for the realisation of the various steps within an analytic process. A characterization of the data mining and analytics algorithms based on their performance on handling linked data is envisaged to be produced. The following indicators are considered: time to fetch the required linked data, time for execution of an analytic process (per algorithm and per volume of processed data) and time for producing the output linked data (per algorithm and per volume of produced data). Such indicators are accompanied by system performance metrics such as processor utilization, cache behaviour, memory access statistics, and bus usage, considering that the performance evaluation is realised in the same or similar systems.

The qualitative metrics refer to a set of usage metrics, such as effectiveness (accuracy and completeness with which users achieve specified goals), satisfaction (the degree to which users are satisfied with the experience of using a product in a specified context of use) and usability (the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use).

4 BUSINESS ANALYSIS OVER LINKED GOVERNMENTAL, SOCIETAL AND HEALTH DATA

In this section, we provide description of a business analysis realized through the implemented business analytics and data mining component. The provided analysis is indicative, related with governmental, societal and health parameters and aims to prove the validity of the designed solution. It should be noted that a series of cross-sectorial analysis is envisaged to be realized within the framework of the LinDA project, based on the implementation of the LinDA pilots (business intelligence, environmental and media analytics sectors) as well as the usage of the LinDA ecosystem from interested SMEs and organizations. An overview of the analysis is provided in the first subsection focusing on the identified business needs, followed by the description of the consumption and production of linked data along with the primary analysis results in the second and third subsections respectively.

4.1 Business Analysis Overview

The realized analysis regards the examination of the relationship among governmental, societal and health parameters across several areas in Italy. The primary subject of the analysis is related with the health impact of the intensification of the industrial activity, taking into account the respect of the existing laws and guidelines from regulatory authorities during a set of years. The influence imposed through the consumption of mass media towards the public opinion’s formation and control is also taken into account in the analysis.

Based on existing datasets –as they are detailed in the following subsection- the business analysts examine the evolution of industrial activity in specific areas in relation with the rates of carcinogens exposure violations monitored in the various industries. The objective is to identify any correlation among the industrial activity and its impact on the health of workers. Since the industrial activity should be related with development indicators in a country and the realized violations could be also related to corruption levels (e.g. through greater tolerance by public officers to violation instances), examination of the variation in the collected data based on changes on the Gross Domestic Product (GDP) growth rate and corruption indexes are examined.

Locality characteristics are also taken into account in the study, from various perspectives. On one hand, the incident of violation of the exposure limits in carcinogens in a region is examined in relationship with the particular political party in power in this area. On the other hand, given the power that the mass media have in modern societies, the analysis includes examination of the relationship that may exist among the high consumption of mass media and the political parties that are dominant per geographical area in Italy. Last but not least, the number and type of the industries that are operating close to the areas with high number of carcinogens exposure violations are examined. The identification of such relationships may help the business analysts or the public officers to propose holistic solutions for reducing the negative health impact that the industrial activity might have.

4.2 Linked Data Consumption

The input datasets for the realization of the analysis are provided from a diverse set of sources, as shown in Table 2. Interlinking of datasets is realized based on the location and time parameters in order to prepare unified datasets, ready to be fed as input to the analysis process. An indicative example of interlinking is shown in Figure 6.Further interlinking may be realized by the business analysts for the collection of information from available web resources, such as information related to a specific Italian province (e.g. information from DBpedia), the positioning of the members of the national or the European parliament of a specific political party (e.g. information from the datasets of the Italian parliament - http://data.camera.it/data/en/datasets/- or the European parliament - http://www.talkofeurope.eu/data/) or the number and type of the industries that are operating close to specific areas (e.g. information via http://www.openstreetmap.org).

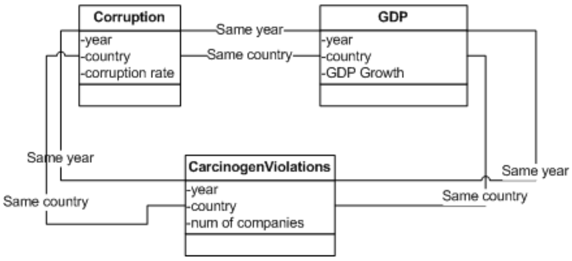

Figure 6: Preparation of interlinked datasets for analysis.

Table 2: Data sources and parameters description.

Data Source	Parameter Description	Data access
Istat (http://linkedstat.spaziodati.eu/)	Political parties leadership & mass media consumption per Italian region (2005-2012)	SPARQL endpoint
Inail (http://dati.inail.it/opendata/elements/RegistriEsposizione)	Number of companies that surpass the established exposure levels to carcinogens subjects (1994-2012)	RDF dump
Worldbank (http://worldbank.270a.info/sparql)	GDP growth rate and corruption levels (1996-2008)	SPARQL endpoint
Linked Geodata (http://linkedgeodata.org/sparql)	Industrial sites per region (current status). The query is realised based on the http://www.openstreetmap.org project.	SPARQL endpoint

4.3 Primary Analysis Results and Linked Data Production

The first part of the analysis regards the examination of the impact that the GDP growth rate and corruption indexes variation have to the total carcinogens exposure violations in country level. A multiple linear regression was realized for this purpose -having as dependent variable the number of violations and as independent variables the GDP growth rate and corruption indexes- resulting though to a high p-value and not permitting the extraction of conclusions regarding potential associations among the examined variables. Then, the M5P algorithm was executed for producing a piecewise linear fit to the dependent variable. Based on the produced results, it seems that high GDP growth rate values (GDP growth rate > 1.708) are associated with smaller number of violations (55 instances), while low GDP growth rate values (GDP growth rate < -0,095) are associated with high number of violations (from 1239 to 3305 instances). When the GDP growth rate fluctuates between these values (-0.095 ≤ GDP growth rate ≤ 1.708), the trend followed depends on the corruption levels. An increase in the corruption levels leads to an increase in the number of the violations, as it could be envisioned. The lowest number of violations is given for the following combination: GDP growth rate ≤ 1.708 and 4.25 < corruption rate <= 4.75.

The second part of the analysis regards the relationship among the consumption of mass media and the political preferences of people per Italian area. A clustering analysis is realised leading to the following highlights. Mass media consumption and political party preferences are highly correlated, as they are grouped in less and well defined clusters in comparison with the political party preferences and the Italian province variables or the mass media consumption and Italian province variables, as illustrated in Figure 7. Furthermore, it is noted that the provinces associated with lower consumption of mass media vote in a more differentiated way. As shown in Figure 8, in regions with high consumption of mass media there is a dominant cluster (the red one) where its instances are mostly composed by the dominant political party, while in the case of regions with low mass media consumption, the red cluster is smaller and the rest of them are more intensely represented, as illustrated in Figure 9.

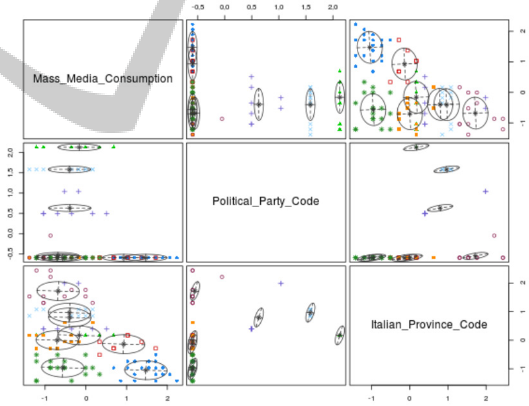

Figure 7: Clustering analysis results.

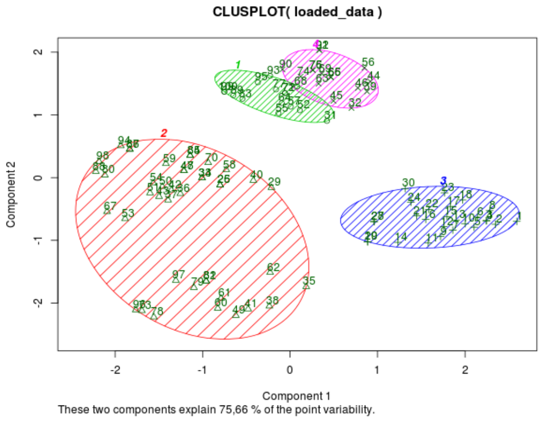

Figure 8: Clustering of political parties' preferences to regions with high mass media consumption.

Following, an instance of the RDF files of the produced linked data analytics through the execution of the aforementioned clustering analysis is shown in Table 3. The interconnection of the analytics input and output files based on the concepts and relationships defined in the LDAO ontology is provided.

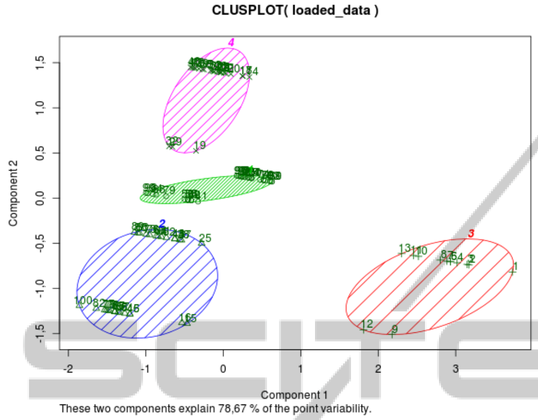

Figure 9: Clustering of political parties' preferences to regions with low mass media consumption.

Table 3: Instance of linked data analytics.

```
Indicative Input Dataset:
<rdf:Description
rdf:about="http://linkedstat.spaziodati.eu/dataset/61_132/
ITE4/TOTAL/20/9/9/1/A/2000">
        <rdf:type
rdf:resource="http://purl.org/twc/vocab/conversion/RawConv
ersionProcess"/>
        <dbpedia-owl:media>90.7</dbpedia-owl:media>
        <dbpedia-owl:place>Lazio</dbpedia-owl:place>
        ………..
</rdf:Description>
Analytics Result Node:
<rdf:Description
rdf:about="http://linda.epu.ntua.gr:8000/analytics115V2Dat
e15022015#/81">
        <rdf:type
rdf:resource="http://linda.epu.ntua.gr:8000/analyticsontol
ogy#analytics_result_node"/>
        <analytic_process:predictedValue>3</analytic_proce
ss:predictedValue>
        <prov:wasGeneratedBy
rdf:resource="http://linda.epu.ntua.gr:8000/analyticsontol
ogy#analytic_process/115/2"/>
        <prov:wasDerivedFrom>http://linda.epu.ntua.gr:8000
/analyticsontology#analytic_input_collection/115/2/3</prov
:wasDerivedFrom>
        <rdfs:subClassOf
rdf:resource="http://www.w3.org/ns/prov#Entity"/>
</rdf:Description>
Analytics Input Collection
<rdf:Description
rdf:about="http://linda.epu.ntua.gr:8000/analyticsontology
#analytic_input_collection/115/2/3">
        <rdf:type
rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#Bag"/>
        <rdf:_1
rdf:resource="http://linkedstat.spaziodati.eu/dataset/61_1
32/ITF4/TOTAL/20/9/9/1/A/1998"/>
        ………..
        <rdf:_3
rdf:resource="http://linkedstat.spaziodati.eu/dataset/61_1
32/ITE4/TOTAL/20/9/9/1/A/2000"/>
</rdf:Description>
<rdf:Description
rdf:about="http://linkedstat.spaziodati.eu/dataset/61_132/
ITE4/TOTAL/20/9/9/1/A/2000">
        <rdf:type
rdf:resource="http://linda.epu.ntua.gr:8000/analyticsontol
ogy#analytic_input_node"/> </rdf:Description>
```

5 CONCLUSIONS AND FUTURE PLANS

By taking into account current business trends and challenges towards the production of advanced business analytics that can lead into insights and facilitate businesses to make analytical-driven decisions, an approach for exploiting linked data towards the production of added-value business analytics has been provided.

The approach is based on a designed and developed business analytics and data mining tool that facilitates the consumption of linked data for the realisation of analysis, as well as the production of data analytics interlinked with the input data in the analytics process making them discoverable and further exploitable in the future. The functional architecture of the proposed approach, the designed ontology for the representation of the defined concepts and relationships as well as a pilot application scenario for validation purposes are presented in detail.

Based on the provided results, it could be argued that the proposed approach can help enterprises enhancing their experience of managing and processing of data, in ways not available before. It can provide them the potential to produce advanced knowledge, leveraging the power of linked data analytics, acquire a significant competitive advantage in the decision making process and increase their overall effectiveness. However, in order to be able to easily adopt and integrate the usage of such an ecosystem in their daily business processes, they have also to take into account the need for an initial learning curve as well as the involvement of data scientists in the specification of the analysis and the interpretation of the analysis results.

With regards to open issues for further investigation, a set of cases are identified. These include the need for realising a detailed evaluation of algorithms efficiency with regards to the usage of linked data taking into account a wide set of algorithms per category, the performance evaluation of the proposed business analytics and data mining tool based on the defined performance evaluation framework, the tackling of challenges related to the management of big data and the adoption of a distributed nature of the execution mode as well as the implementation of a set of cross-sectorial use cases aiming at the validation of the overall approach and the identification of the produced business value in each case.

ACKNOWLEDGEMENTS

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

REFERENCES

Brickley, D., Miller, L., 2014. FOAF Vocabulary Specification 0.99. *Available Online*: http://xmlns.com/foaf/spec/

eBay report, 2013. How to Build Trust and Improve the Shopping Experience. *Available Online*: http://knowwpcarey.com/article.cfm?aid=1171.

Gartner report, 2013. Gartner Says Business Intelligence and Analytics Need to Scale Up to Support Explosive Growth in Data Sources. *Gartner press release*, *Available Online*: http://www.gartner.com/newsroom/id/2313915.

Gartner report, 2014. Gartner Says Advanced Analytics Is a Top Business Priority, *Gartner press release*, *Available Online*: http://www.gartner.com/newsroom/id/2881218.

Hirudkar, A.M., Sherekar, S., 2013. Comparative Analysis of Data Mining Tools and Techniques for Evaluating Performance of Database System, *International Journal Of Computer Science And Applications*, Vol. 6, No.2, Apr 2013, ISSN: 0974-1011.

Hu, H., Wen, Y., Chua, T., Li, X., 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial, *IEEE Access*, vol.2, pp.652, 687, doi: 10.1109/ACCESS.2014.2332453.

IBM report, 2014. Inside the mind of Generation D, What it means to be data-rich and analytically driven, *Available Online*: http://www.ibm.com/smarterplanet/us/en/centerforappliedinsights/article/gen_d_insights.html.

IDC report, 2014. Worldwide Business Analytics Software 2014–2018 Forecast and 2013 Vendor Shares, *Available Online*: http://www.idc.com/getdoc.jsp?containerId=249926.

Intel report, 2015. Achieving Intel Transformation through IT Innovation, 2014–2015 Intel IT Business Review – Annual Edition, *Available Online*: http://www.intel.com/content/dam/www/public/us/en/documents/best-practices/intel-it-annual-performance-report-2014-15-paper.pdf.

ISO/IEC 25010, 2011. Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models, *Available Online*: http://www.iso.org/iso/catalogue_detail.htm?csnumber=35733.

Jia, Z., Zhan, J., Lei, W., Han, R., McKee, S.A., Yang, Q., Luo, C., Li, J., 2014. Characterizing and subsetting big data workloads, *In 2014 IEEE International Symposium on Workload Characterization (IISWC)*.

Knime tool, 2015. KNIME Analytics Platform, *Available Online*: https://www.knime.org/knime.

Kreissl, R. 2013. Datenspuren: Komplette Umkehr der Beweislast. New Scientist, *Available Online*: http://irissproject.eu/?p=325.

Lausch, A., Schmidt, A., Tischendorf, L., 2015. Data mining and linked open data – New perspectives for data analysis in environmental research, *Ecological Modelling*, Volume 295, 10 January 2015, Pages 5-17, ISSN 0304-3800, http://dx.doi.org/10.1016/j.ecolmodel.2014.09.018.

LDAO, 2015. Linked Data Analytics Ontology. *Available Online*: http://linda.epu.ntua.gr:8000/vocabulary/122/linked-data-analytics-ontology/

Leavitt, N., 2014. Bringing big analytics to the masses, *Computer*, vol.46, no.1, pp.20-23, Jan. 2013, doi: 10.1109/MC.2013.9.

Lebo, T., Sahoo, S., McGuinness, D., 2013. PROV-O: The PROV Ontology, *W3C Recommendation*. *Available Online*: http://www.w3.org/TR/2013/REC-prov-o-20130430/

Piatetsky-Shapiro, G., 1991. Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., *Knowledge Discovery in Databases*, AAAI/MIT Press, Cambridge, MA.

R Project, 2015. The R Project for Statistical Computing, *Available Online*: http://www.r-project.org/

RapidMiner tool, 2015. *Available Online*: https://rapidminer.com/

Weka tool, 2015. Weka 3: Data Mining Software in Java, *Available Online*: http://www.cs.waikato.ac.nz/ml/weka/

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S., 2015. Quality Assessment for Linked Data: A Survey, *Semantic Web journal*, IOS Press.