Undefined 1 (2015) 1–11
IOS Press

LinDA – Linked Data for SMEs [1]

Editor(s): NN.
Solicited review(s): NN.
Open review(s): NN.

Spiros Mouzakitis [a], Judie Attard [b], Robert Danitz [c], Lena Farid [c,*], Eleni Fotopoulou [d], Michael Galkin [b,g], Barbara Kapourani [e], Fabrizio Orlandi [b], Dimitris Papaspyros [a], Michael Petychakis [a], Andreas Schramm [c], Anastasios Zafeiropoulos [d], and Norma Zanetti [f]

[a] *School of Electrical and Computer Engineering, National Technical University of Athens, Heroon Polytechneiou 9, 15773 Zografou, Athens, Greece*
E-mail: {smouzakitis,dpap,mpetyx}@epu.ntua.gr
[b] *Institut für Informatik III, Rheinische Friedrich-Wilhelms-Universität Bonn, Römerstraße 164, 53117 Bonn, Germany*
E-mail: {attard, orlandi}@iai.uni-bonn.de, mikhgalkin@gmail.com
[c] *Fraunhofer FOKUS, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany*
E-mail: {robert.danitz,andreas.schramm,lena.farid}@fokus.fraunhofer.de
[d] *Ubitech Ltd., Thessalias 8 & Etolias 10, 15231 Chalandri, Athens, Greece*
E-mail: {efotopoulou,azafeiropoulos}@ubitech.eu
[e] *Critical Publics, Garyttou St. 150, 15343 Chalandri, Athens, Greece*
E-mail: barbara@criticalpublics.com
[f] *Hyperborea Srl, Via Giuntini 25/6, 56023 Navacchio di Cascina, Italy*
E-mail: n.zanetti@hyperborea.com
[g] *Faculty of Computer Technologies and Control, ITMO University, 49 Kronverkskiy ave, 197101 Saint Petersburg, Russia*

Abstract. *Linked Data* is an active research field currently; new ideas, concepts, and tools keep emerging in quick succession. In contrast, relatively little activity is being seen towards aspects like ease of use and accessibility of tools for non-experts to promote Linked Data at SME level. This is a sign for the still developing maturity of said research field, and it is the motivation and starting point of the LinDA ("*Linked DAta*") project. Concepts and components of an integrated framework of tools will be presented, where both Linked Data provision and consumption are covered. It will be shown how these tools can be combined and employed to carry out various activities, whose scopes range from single actions to holistic workflows. The usability of these workflows will be demonstrated along concrete pilot application scenarios.

Keywords: Linked Data, workflows, small and medium enterprises (SMEs), semantic enrichment for SMEs.

1. Introduction

Linked Data is a term coined and animated by Tim Berners-Lee[1], referring to a data representation methodology that is machine readable and semanti-

[1]The LinDA project has received funding from the European Union's 7th FP for research, technological development and demonstration under grant agreement no. 610565.
[*]Corresponding author. E-mail: lena.farid@fokus.fraunhofer.de.

[1]http://www.w3.org/DesignIssues/LinkedData.html

cally enriched, and hence apt for semantic interlinking and automated processing of semantic queries. Linked Data draws some of its advantages from its building on RDF – Resource Description Framework [6] –, a standard data representation methodology based on labeled directed graphs, whose labels are literals or dereferenceable URIs that lay the semantic foundation in that they refer to ontology elements (RDF has come into existence long before the appearance of Linked Data).

The big potential of Linked Data lies in its unifying powers. They manifest both at the representational level, via the RDF format, as well as at the semantic level, by the establishment of relations to standard ontologies. Sophisticated application scenarios will, for instance, extract knowledge from combined data obtained from various places in the world, previously unforeseen by their providers. These applications will act as a kind of globally distributed inference engines. Indispensable prerequisites for this are machine readability and the ability of semantic interlinking, both of which are catered for by the Linked Data paradigm and appropriate ontologies.

When mapping the latter to enterprises, especially to SMEs, it unreservedly gives them the potential to efficiently develop novel data anlytical services that contribute to improving their competitiveness and stimulating the emergence of innovative business models.

So far for the prospects and expectations; the reality down here on earth is not quite as rosy. Linked Data is an emerging field, which has consequences for its practical use:

- There is a lack of established workflows and convincing use cases.
- The number and diversity of tools, and the pace of their appearance and replacement, makes them unmanageable for use within SMEs.
- Public data sources and useful vocabularies are still scarce and/or difficult to find.
- High initial costs and the lack of user friendliness.

In this paper, the LinDA[2] project is presented, which addresses these difficulties and strives for implementing the use of Linked Data within non-expert users such as SMEs. — Overview of the rest of the paper: Section 2 presents a summary of some of the related work and refers to existing surveys. In Section 3, the approach, its rationale and philosophy are introduced. Section 4 presents the individual components of integrated workbench (the LinDA Workbench). Based on that, Section 5, gives the unifying picture of the project and the LinDA workflows. In Section 6, the application scenarios that are applied to LinDA are presented. Finally, the outlook and conclusions are drawn in Section 7.

2. 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 [11,12]. The challenges can be split into three distinct categories: reviewing the datasets and preparing them in a proper format, interpreting or extracting knowledge from the data through interlinking, inferences as well as analytics extraction, and maintaining and updating the data regularly.

Several projects are actively handling parts of these challenges, many of these projects deliver standalone frameworks and do not provide a holistic workflow. For instance, Sparqlify[3] and SparqlMap [13] are two powerful libraries for transforming structured and semi-structured data into RDF, using standardized formats and protocols, such as the *Sparqlification Mapping Language* (SML)[4] and the W3C R2RML[5] respectively. Both libraries are unfortunately not explicitly designed for non-experts and assume technical adeptness as well as extensive knowledge in the semantic web field. Furthermore, they don't come with a use interface.

Openrefine[6] is another powerful tool that touches upon Linked Data, although initially designed for data wrangling, data augmentation and as its name suggests data refinement. Openrefine allows the docking of plugins to extend its workflow but this entails concrete knowledge of its code and would possibly cause SMEs high initial costs. This in the opinion of the authors does not solve the SME problem.

Furthermore, the LOD2[7] project that aims to contribute high-quality interlinked versions of public Se-

[2] http://www.linda-project.eu

[3] http://www.sparqlify.org
[4] http://sparqlify.org/wiki/SML
[5] http://www.w3.org/TR/r2rml/
[6] http://openrefine.org/
[7] http://lod2.eu/

mantic Web data sets, promoting their use in new cross-domain applications by developers across the globe. The DIACHRON[8] project takes on the challenges of evolution, archiving, provenance, annotation, citation, and data quality in the context of Linked Open Data 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[9] handles the uptake of open geographic information through innovative services based on Linked Open Data and LATC[10] and EUCLID[11] projects that aim to support people and organisations to better publish and consume Linked Open Data. With regards to enabling Linked Data analytics, up to knowledge of the authors, no holistic framework exists that is able to consume Linked Data towards the production of analytics and produce output data interlinked with the input data. Several open source frameworks exist for the production of data analytics (e.g. Weka[12], Knime[13], R[14], Pentaho[15] (community edition)), however without the support of RDF as an input/output format. Furthermore, large effort is 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 [11,10].

3. Overview of LinDA's approach

The aims of the LinDA project are not to add further to the tools of the kind presented in the previous section, but at selecting and improving upon existing ones and constructing a useful integrated tool chain while focusing on the non-expert end users, and thus enabling them to make their first steps in this field. Non-expert end-users are regard as those that are new to the realm of Linked Data but do posses sufficient knowledge of data science in their respective domain.

[8]http://www.diachron-fp7.eu/
[9]http://sdi4apps.eu/
[10]http://latc-project.eu/
[11]http://euclid-project.eu/
[12]http://www.cs.waikato.ac.nz/ml/weka/
[13]http://www.knime.org/knime
[14]http://www.r-project.org
[15]http://sourceforge.net/projects/pentaho/

Essentially, the LinDA approach consists in a number of tools, integrated in a workbench, that are accessed from a set of prefabricated minimalistic but holistic workflows.

Thus, the tool chain provided by LinDA covers both *data provisioning* and *data consumption,* including publishing and importing from external data sources.

The pivotal point is the *Linked Data* format.

For the data provisioning, there are tools for data renovation, i.e., conversion from various formats towards Linked Data in this case, specifically, the conversion to the RDF format); for the data consumption side, there are tools for querying, analytics, and visualizations.

Conceptual and technical simplicity are ranked higher than feature completeness. For instance, the data formats handled are confined to "semi-structured data", i.e., data from relational databases, CSV files and similar, as opposed to "unstructured data", which is the term commonly used to refer to natural-language data. The reason is that the respective complexities differ vastly, and Natural Language Processing (NLP) is an active research field in its own right, making it particularly unsuitable for toolchains that are supposed to subsume results rather than to support research.

4. The integrated workbench

This sections presents the components of the integrated workbench individually, each with its major features.

These separate tools are integrated into a common web application for easy installation in various deployment scenarios. Following this the overall high-level architecture for the LinDA Workbench is presented.

4.1. Transformation Engine

The Transformation Engine (TE) is the pivotal part of the provisioning framework. The purpose of the LinDA TE is renovating (semi-)structured private and Open Data by transforming them into semantically described and enriched data. Two strains of transformation are currently supported: the transformation of relational data from SQL databases, and the transformation of tabular data from CSV files. In the user interface, the user is guided through the process of the transformation by a wizard-style "Mapping Designer" that includes explanatory text to aid with the renovation process. The Mapping Designer generates a map-

ping file as a result. This mapping file captures the necessary relations from input data elements (i.e. table, column, row, and literal) to RDF elements (i.e. <rdfs:Class>, <rdfs:Property>). The mapping file is expressed in the SML, which provides a SPARQL-like syntax to address elements that are foreign to the Resource Description Framework, i.e. tables and columns.

1. RDB-to-RDF transformation:
   When connected to a SQL database as input, the TE presents the schema of the database and lets the user specify the mapping according to the SML. The TE leverages the functionality of the Sparqlify project to execute the transformation and providing the RDF output data. The resulting RDF data can therefore either be *dumped* to a triple store, or be served by the SPARQL-SQL-rewriter engine of Sparqlify as a SPARQL endpoint. The rewriter expects a SPARQL query on the RDF data basis, and rewrites it to a SQL query which can be optimized and executed efficiently within the RDBMS.
2. CSV-to-RDF transformation of tabular data from CSV files:
   When providing a CSV or TSV file as data input, the TE presents the data accordingly, treating it as relational data with just one table. In this case, Sparqlify can be leveraged as well. However, an ad-hoc query compilation is not reasonable, as the resulting extraction operation would not perform well on plain text files. The resulting data can be dumped to disk/to a triple store instead

When dealing with relational or tabular data, a straight-forward way to interpret the structure of the data is to regard tables or tabular files as data about a specific *subject;* database or CSV columns describing a certain *property* about the subject; and where each row in the table constitutes a statement about one *individual.* Therefore, tables should be annotated with one or more RDF classes, whereas columns should be annotated with at least one RDF property.

The end user is assisted in the production of Linked Data by using the Mapping Designer in a guided and user-friendly manner. As the quality of the produced Linked Data is paramount, a user-supervised, semi-automated process is adopted as opposed to a fully automated process, where the domain expert has several means to adjust the transformation and give preferences for choosing RDF elements and vocabularies. The user is encouraged to select classes and properties from already existing "higher" vocabularies. The mapping designer provides an interface to search for relevant RDF elements among a selection of frequently used and well-described vocabularies using the external Linked Open Vocabularies web service[16]. On the other hand, the Mapping Designer is based on simple assumptions concerning the structure of the output data (as outlined above). This limits the expressiveness of the RDF graph yet positively simplifies the complexity of the user interface, thus lowering the complexity of usage for the domain expert.

4.2. Vocabulary and Transformation Metadata

LinDA Vocabulary and Metadata repository fulfils the need for a repository of Linked Data vocabularies, accessible by users and other tools in the LinDA ecosystem. Linked Data vocabularies are collections of relationships between real world data, known as properties, and object categories, known as classes, typically expressed in RDF [3]. Based on Linked Data vocabulary repositories that already exist and are open to the public [5,14], the LinDA vocabulary repository presents many advantages for enterprises and non-technical users. The presented architecture alongside with its implementation facilitate both enterprise users as well as other LinDA applications and Linked Data applications in general in the processes of creating and publishing semantically rich linked datasets in a more robust and expandable way. The repository can be installed within the LinDA workbench. Each vocabulary repository installation retains an independent metadata database that is synchronized to a central, master vocabulary repository. Every vocabulary repository instance periodically communicates to the master LinDA vocabulary repository to search for new vocabularies and metadata or updates to existing vocabularies. Knowledge inside the master repository is gathered from different catalogues around the web, with local repositories also having the ability to add repositories locally or remove unwanted ones. Thus the vocabulary repository of an enterprise can be enriched with private vocabularies that are stored in a private database, from where they are exposed to all other LinDA tools and modules inside the enterprise. In order to facilitate the exposure of vocabulary metadata to Linked Data and Semantic Web applications in general, and especially for the Transformation Engine,

[16] http://lov.okfn.org/dataset/lov/

calls to an exposed RESTful Web based API allow applications to specify search terms and categories. The API shall then operate as an oracle, suggesting a list of entities that may be suitable in the specified context, ranked according to factors like community rating and other vocabulary statistics provided by the master vocabulary repositories, like statistics from the LODStats framework [8]. The repository also offers a web interface where users can browse the various vocabularies, along with all the entities defined by them, including classes and properties. Interconnections between different entities are described, and new information like vocabulary usage examples has been added, in order to allow users better understand the intended usage of each vocabulary. It should be noted that all information is automatically extracted by the vocabulary definitions, using SPARQL queries. Features like auto-detection of search term language and immediate translation also facilitate non-English speakers' needs, and can also improve the suggestions produced for databases and CSV files with terms in other languages. Vocabulary definition source documents are also accessible in all major RDF formats, and a visualization of the RDF graph allows users to better perceive each vocabulary's structure. The repository uses various libraries in its web application's implementation. Django is a lightweight framework written in python that emphasizes the DRY principle [9]. The jQuery library [2] is used to facilitate AJAX requests and to make the application more dynamic all around. Finally, with ElasticSearch [1] it becomes possible to support multiple search requests to the web application and the Oracle API at the same time.

4.3. Publication and Consumption Tools

The Linked Data Publication and Consumption Framework aims to assist SMEs and data publishers and consumers in analysing and interlinking public sector information with enterprise data. In Figure 1 the basic workflow of the processes within the framework is shown. Our process starts with dataset exploration. Here a user can explore existing open datasets, both with regards to their content and also with regards to the underlying schema. The user then has two options, either directly executing a SPARQL query in the *SPARQL Query Tool* or use the *Query Builder Tool* to generate the query. Finally, the user can proceed to convert the result set in a number of different formats and export results. This conversion allows users to import data from linked open datasets into their native system. The results can also be exploited further through the LinDA Visualisation and Exploration Ecosystem, and the LinDA Analytics and Data Mining Services.

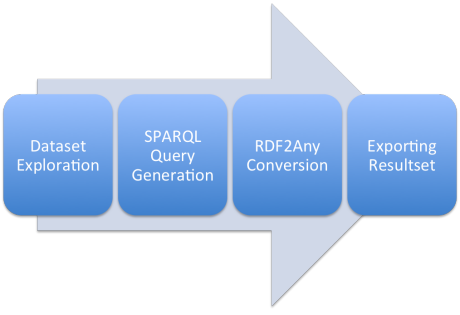

Fig. 1. Workflow within the Linked Data Publication and Consumption Framework.

The main tools provided within this framework, which provide the functionality described above, are the following:

RDF2Any API: This API provides the functionality for both the dataset exploration and converting of RDF data into a number of target formats, including JSON, RDB, CSV, and a generic format. The latter conversion allows a user to upload a template and execute a conversion specific to his/her needs. The RDF2Any API can also be easily extended with other converters.

SPARQL Query Tool: Intended for experienced users, this tool allows users to directly input SPARQL queries which are executed on defined external endpoints. The users can then preview, convert the results in a number of formats, and export the resultset. The format conversion functionality is obtained through the consumption of the RDF2Any API.

Query Builder Tool: As opposed to the SPARQL Query Tool, this tool is aimed for users who are not familiar with the SPARQL query language. It provides drag and drop and auto-complete features that allow a user to easily build the desired query and access the required data. This tool also caters for users who either don't know the specific dataset(s) that contains the data that suits their needs, or otherwise would like to explore existing open datasets. Through this tool, the user can navigate through classes, subclasses, instances, and properties. Similarly to the SPARQL Query Tool, the user can then preview, convert, and export the resultset. The Query Builder tool also operates through the consumption of the RDF2Any API which provides the required functionality.

4.4. Visualization Tool

The role of LinDA Visualization is to provide a largely automatic visualization workflow that enables SMEs to visualize data in different formats and modalities. In order to achieve this, a generic web application is being developed based on state-of-the-art Linked Data approaches [4,7,15] to allow for visualizing different categories of data, e.g. statistical, geographical, temporal, arbitrary data, and a largely automatic visualization workflow for matching and binding data to visualizations. The Visualization tool consists of two main components:

- The Explore and Select Data component;
- The Visualization component.

The Explore and Select Data component allows users to pre-select the data to be processed and provides a concise preview of the data through a tabular representation. The input data formats supported are RDF (any serialization and vocabulary) and CSV. The selection performed with this component is used as an input of the Visualization component.

The Visualization component is responsible for the creation of various plots, charts and maps and offers opportunities to customize visualization options or save and share the graphical results. The visualization is powered by a recommendation algorithm which aims at suggesting the most suitable visualization type according to different features of the input data. The algorithm consists of the following fundamental steps:

1. Defining the dimensions and scales of measurement of the selected data;
2. Building possible allocations, or combinations, of the dimensions;
3. Comparing the constructed allocations with premade patterns, which describe the required parameters of a visualization component (e.g. the formats for the X and Y axes of a line chart);
4. Ranking of the most relevant allocations and building the list of preferred visualizations.

An essential part of the recommendation engine is the visualization ontology: a high-level model of the visualization workflow, its components and its particular instances (e.g. line chart, map, etc.).

The implemented tool adopts *Ember.JS* as a main operational platform combining a comprehensive user interface with a relatively lightweight backend. By adopting existing open source libraries the following visualizations to an extensible modular infrastructure have been integrated: Line Chart, Bar Chart, Column Chart, Pie Chart, Scatter Chart, Area Chart, Bubble Chart, and Map.

4.5. Analytics and Data Mining

The LinDA Analytics and Data Mining component supports the realisation 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 SMEs to utilise 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. High priority is given to the user friendliness of the provided interfaces based on the design of specialized workflows per algorithm category (e.g. workflows for supervised learning techniques such as classification and regression/forecasting algorithms and unsupervised learning techniques such as clustering and pattern discovery (association) algorithms). The LinDA Analytics and Data Mining component supports RDF as an input and output format, while the provided output from the analysis is interlinked with the input RDF source or sources. In addition to the interlinking, information regarding the type of the analytic process executed – including configuration and description issues – is saved and made available to the end users for further use. RDF input is supported through the selection of existing queries that have been prepared through the LinDA Query Builder or the LinDA Query Designer, their execution and the production of the input RDF data source for the analytic process. The next step regards the selection of the appropriate algorithm to be executed and the configuration of the analytic process to be followed. As already mentioned, a set of algorithms are integrated and custom workflows are prepared based on default parameters. Specifically, integration of the Weka open-source tool and the R open-source project for statistical computing is realised, while workflows are developed for algorithms that are being used in the analysis at the LinDA pilots (business intelligence analytics sector, environmental analytics sector, media analytics sector). A default configuration is proposed at each case, while there is flexibility for executing the algorithm with customized configuration. Upon the finalisation of the configuration phase, the algorithm is executed in a transparent way for the end user and the output results are produced. Different types of output formats are supported based

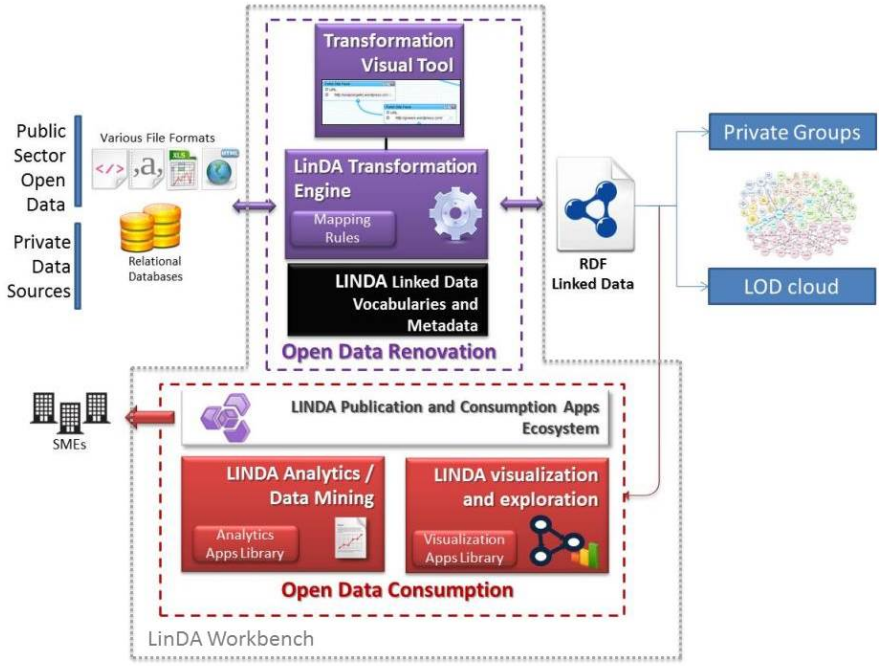

Fig. 2. The LinDA Workbench high-level architecture.

on the peculiarities of each algorithm. Thus, the analytics output can be an RDF file, plain text or even specialized graphs. As already mentioned, in case of RDF output, interconnection of the output dataset to the input dataset and the executed analytics process is provided. This interconnection is based on the design of a specific ontology that describes the overall analytic process followed and the relationship among the input and output data sources. Thus, the end users are able to have access to the analytics processes executed in the past along with the input and output files used or produced. This functionality is considered critical in cases that re-evaluation of algorithms/processes have to be realised in the examined business scenarios where input data are updated. As stated earlier, the design and deployment of the analytics and data mining component is realised with main objective its user-friendliness towards the SMEs employees. However, even if specific workflows are designed per algorithm category that hide part of the complexity, detailed knowledge of the algorithm's functionality and configuration parameters is required. Since, in most cases, a set of analysis has to be realised by interplaying also with the parameters, the end users are responsible for properly configuring the considered parameters per algorithm according to their needs.

4.6. Architectural properties of the LinDA Workbench

From a technical perspective, the LinDA tools presented earlier in this chapter are modular and can be deployed and used separately in other projects, as independent tools. With a view to achieving a seamless workflow between the tools, however, the LinDA Workbench has been developed. The LinDA Workbench orchestrates the tools and handles the main communication with the selected triple store. Moreover, it acts as an integrated environment with the ability to host new tools through an extensible python-based system. The LinDA Workbench provides an authentication and authorization framework above all integrated tools, handling user accounts, groups, permissions and cookie-based user sessions. Data Source Management is the core submodule of the LinDA Workbench, where users can create, edit or delete graph-based RDF models or provide a link to a public SPARQL endpoint. Moreover LinDA workbench provides a context-menu with all available services (explore, visualize, etc.) for each data source and global configuration page for all hosted tools. Lastly LinDA Workbench provides an administration interface for the management of all entities of the system (Data Sources, Queries, Analytics, Users, Groups, etc.) As such the main workflow of the tools is illustrated in Figure 2.

5. Workflow

From a user perspective the main LinDA workflow can be summarized in three simple steps as illustrated in Figure 3.

More specifically the three workflow steps are:

Step 1: Turn data into RDF: Using LinDA Transformation Engine, users can publish their data as linked data in a few, simple steps. They 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 through its Oracle API.

Step 2: Query/link your data: With the LinDA Query Builder, users can perform simple or complex queries through an intuitive graphical environment that eliminates the need for SPARQL syntax.

Step 3: 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.

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 (N3, RDF/XML, Turtle, Trix and TriG) 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, tsv, etc.) the LinDA Workbench guides the user to the Transformation Engine 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 through the Query Builder,
d) edit/update/delete it.

By querying internal or external RDF data sources, the user can extract specific information and save the query for future use. Similar to the list of data sources, the user is able to perform specific operations to the list of the saved queries including visualizing, analysing and editing a query item. By clicking the "Visualize" or "Analyse" button the user is transferred to the LinDA visualization or Analytics engine respective with the required data source input pre-selected, according to the chosen query output. The user can also further edit the query through the Query builder tool. Visualizations configurations and Analytic processes are also saved in a dedicated list of respective items. Lastly, the LinDA Workbench provides a ubiquitous full-index search for all entities of the system (Vocabularies, Data Sources, Visualization configurations, saved Analytic processes and Queries). Overall, the workflow made possible via the LinDA Workbench allows users and enterprises to seamlessly transition from traditional data formats into the world of Linked Open Data.

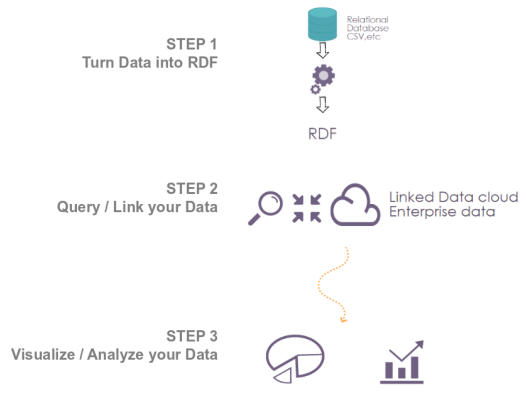

Fig. 3. Simplified Workflow from a user perspective.

6. Scenarios with the LinDA approach

In this section two application scenarios that have been designed along the LinDA project are presented. These test the feasibility as well as the aptness of the tools and ultimately aid in showing the value of Linked Data for SMEs.

6.1. Business intelligence analysis scenario

In the business intelligence analysis scenario an SME is considered that operates at the business intelligence sector and is providing consultancy services at the pharmaceutical domain. Within a specific project,

the employees of the SME are examining whether a pharmaceutical company should invest in the Over-The-Counter (OTC) drugs market in a set of countries in Europe. The decision for proceeding to an investment is depending on a set of conditions related to the existence of a liberalized market for OTC (in terms of price, entry and/or distribution-retail) as well as socio-economic indicators. The realisation of analysis is required in order to gain insights regarding the opportunities for entering in such a market in a specific country. By conceptualising the domain under investigation and by using linked data analytics, the SME is able to build a cohesive knowledge totality based on intelligence, insight and purpose, understanding whether OTC liberalisation is beneficiary, for which countries, for which actors and stakeholders, in which point of time and under which circumstances. Through the usage of the LinDA Ecosystem, the SME envisages to facilitate the analysis phase, enhance the provided intelligence services to its clients, and increase the productivity, while reducing the effort and the associated costs. The first and more crucial step for initiating the analysis is the conceptualisation of the pharmaceutical domain under investigation, as it serves as a backbone for the identification and assessment of the operating environment, leading to the situation review and analysis, through LinDA tools. Throughout extensive research and communication with the healthcare scientists and advisors, the most important indicators and parameters that might affect the OTC liberalisation have been identified. These include business-oriented parameters (e.g. number of OTC medicines consumed per country, revenues from OTC medicines, healthcare expenditures), governmental oriented parameters (e.g. corruption, political stability, Purchasing-Power-Parity) and socio-economic oriented parameters (e.g. GDP per capita, level of education, unemployment). Going one step further, a set of public and private datasets are prepared along with their appropriate interlinking in order to be served to LinDA tools as input for the desired analysis. Interlinking is required for combining information (e.g. OTC indicators, demographic indicators, socio-economic indicators) per country and year. Given the heterogeneity in the considered datasets as well as the volatility of data in many cases, the creation and maintenance of interlinked datasets is beneficiary for the SME since it reduces significantly the complexity and the required effort for the realisation of the analysis. Upon the creation of the interlinked datasets, the analysis starts with the preparation phase were association algorithms are being used on the pool of the identified parameters, for an initial identification of parameters that are highly correlated as well as clustering algorithms for identifying groups of interest within the datasets. In the next phase, classification, regression analysis and forecasting algorithms are executed for extracting trends and estimating impacts of the OTC liberalisation. It should be mentioned that a set of iterations may be required for the first two steps in order to manage to acquire meaningful results. At the final step, based on the acquired results as well as visualised insights, a focused analysis takes place for examining the impact of specific parameters and proceeding to decision making.

6.2. Environmental analysis scenario

In this scenario consider an SME is considered that works on elaborating datasets that describe the ecological status of provinces in a specific area based on various parameters related to polluting emissions, waste production, water discharge, water sample etc. The main objective is to use these data in order to proceed to analysis and extract conclusions about the health impact of environmental pollution in this area taking into account the climate conditions. In more detail, environmental scientists within the SME are in contact with environmental institutes and Non-Governmental Organisations (NGOs) targeted to the protection of the environment, universities and research institutes in order to acquire access to specific datasets and extend their database. Prior to preparing the appropriate dataset for proceeding with analysis, consultation with the interested parties is realised for better understanding the peculiarities of the considered case as well as the challenges that are faced. Based on exchange of information with scientists working on the era, it could be argued that a scientist actually needs: access to series of data covering different decades of years; capacity to register cases of mortality and diseases as well as pollution relevant data (e.g. wastes, air pollution), capacity to geo-localise those cases and, finally, capacity to interlink the available data. Based on the acquired information and an overall view on the available datasets, specific theoretical reports are compiled per geographical area with indications regarding the climate conditions, the pollution level and possible consequences for the health of the citizens. A set of hypotheses regarding the correlation among the examined parameters and a set of causations that have the potential to explain the observed results are produced. These hypotheses serve as the starting point for the re-

alisation of analysis later on, while they also lead to the precise specification of the parameters that are going to be used in the analysis phase. Thus, the next step regards the preparation of the envisaged datasets. Interlinking of datasets is being required for the preparation of the input datasets for the analysis by combining information (e.g. health-oriented, pollution-oriented, and climate-oriented) for specific areas and time periods. The creation of Linked Data is considered as an enabler for the realisation of the study since it facilitates the interconnection of existing datasets (a process that was not straight forwards and had a lot of complexity and administration overhead), it permits the interconnection of parameters to other available international environmental or health-oriented datasets as well as the automatic update of datasets in case of addition of further data in the various data sources. It should be noted that in the existing situation, the capacity for realizing advanced analysis is very limited (or requires lot of effort on behalf of the scientists) due to the overhead imposed for processing, interconnecting and maintaining the existing datasets. By having declared the hypotheses to be tested and successfully prepared the appropriate datasets, the next step regards the selection and configuration of the appropriated algorithms to be applied for the production of Linked Data analytics. The objective is to manage to provide partial or full responses to the following questions: is there any correlation among the pollution levels in an area, the climate conditions, demographic characteristics and the incidence of diseases in the population? How may one find abnormalities in the instances of diseases during the years? When the change in the number of instances can be considered to have epidemiological characteristics? What estimates may one produce regarding the evolution of the instances of diseases the upcoming years? Can one assume the existence of clusters with similar characteristics or trends with regards to the expansion of diseases? What kind of preventive actions can be undertaken in cases with large number of incidents or pessimistic predictions in order to restrict the spreading of specific diseases? Given that the geographical location is considered to be highly relevant to the observed incidents, in addition to generic statistical algorithms, the extraction and interpretation of spatial statistics are considered as very helpful for such type of studies. Towards this direction, spatial data handling, display and statistical analysis algorithms are being used. These include, among others, facilities for point pattern analysis, geostatistics and geographic correlation studies, disease mapping and analysis and spatial regression analysis. In addition to statistics, clustering and association algorithms are applied for detecting areas that present similar trends and commonalities in the identified incidents. It should be noted that in cases that the amount of data at disposal is not enough for carrying out an accurate statistical analysis, bootstrapping methodologies are applied for creating appropriate samples.

7. Conclusion and outlook

The paper has introduced a number of tools that successfully implement a holistic yet simple workflow, motivated by the lack of viable alternatives that enable SMEs (whom are most often novices in this field) to benefit from the concepts and ongoing research in Linked Data.

The workflow that has been introduced is bundled in an open source cross-platform webbased platform, the LinDA Workbench. The workflow is made up of three distinct steps that allow for the renovation of datasets, construction of queries, visualisation and conducting analytics. The workflow gives the freedom of docking into a specific step based on the data and the formats at hand.

Further work and research to be conducted is i) the continuous improvement of the user experience in terms of intuitiveness, guidance and simplicity, ii) conducting extensive user studies and evaluations by the pilots, iii) supporting additional semi-structured data formats, such as XML, JSON, and NoSQL databases, iv) and investigating whether Linked Data, in its essence, requires specific data analytics techniques.

References

[1] Banon, S. (2011). Elasticsearch: An open source, distributed, RESTful search engine.

[2] Bibeault, B., & Kats, Y. (2008). jQuery in Action. Dreamtech Press.

[3] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2007). Linked Data on the Web. Workshop Summary. Beijing. Accessed on 30.01.2015 at http://linkeddata.org/docs/how-to-publish.

[4] Brunetti J. M., Auer, S., García, R., Klímek, J. & Nečaský, M. *Formal Linked Data Visualisation Model,* in Proceedings of International Conference on Information Integration and Web-based Applications & Services, Vienna, Austria, 2013.

[5] Cyganiak, R. (2010). About prefix.cc. Accessed on 1.12.2014 von prefix.cc: http://prefix.cc/about.

[6] Cyganiak, R., Wood, D., & Lanthaler, M. *RDF 1.1 Concepts and abstract syntax,* W3C Recommendation, 25 February 2014.

[7] Dadzie, A. & Rowe, M., *Approaches to Visualising Linked Data: A Survey,* Semantic Web, Nr. 2, p. 89–124, 2011.

[8] Demter, J., Auer, S., Martin, M., & Lehmann, J. (2012). LODStats - An Extensible Framework for High-performance Dataset Analytics. EKAW.

[9] Holovaty, A., & Kaplan-Moss, J. (2009). The Definitive Guide to Django.

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

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

[12] 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 .

[13] Unbehauen, J., Stadler, C. & Auer, S. (2012), Accessing Relational Data on the Web with SparqlMap, in 'JIST' .

[14] Vatant, B., & Vandenbussche, P.-Y. (27. 11 2014). Linked Open Vocabularies (LOV). Accessed on 1.12.2014 at Linked Open Vocabularies (LOV): http://lov.okfn.org/dataset/lov/.

[15] Voigt, M., Pietschmann, S., & Meißner, K., *A Semantics-based, End-user-centered Information Visualisation Process for Semantic Web Data,* in Semantic Models for Adaptive Interactive Systems, Springer, p. 83–107, 2013.