



Personalised Exploration Graphs on Semantic Data Lakes

Ada Bagozi, Devis Bianchini^(✉), Valeria De Antonellis, Massimiliano Garda,
and Michele Melchiori

Department of Information Engineering, University of Brescia, Via Branze 38,
25123 Brescia, Italy

adabagozi@gmail.com, {devis.bianchini, valeria.deantonellis,
m.garda001, michele.melchiori}@unibs.it

Abstract. Recently, organisations operating in the context of Smart Cities are spending time and resources in turning large amounts of data, collected within heterogeneous sources, into actionable insights, using indicators as powerful tools for meaningful data aggregation and exploration. Data lakes, which follow a schema-on-read approach, allow for storing both structured and unstructured data and have been proposed as flexible repositories for enabling data exploration and analysis over heterogeneous data sources, regardless their structure. However, indicators are usually computed based on the centralisation of the data storage, according to a less flexible schema on write approach. Furthermore, domain experts, who know data stored within the data lake, are usually distinct from data analysts, who define indicators, and users, who exploit indicators to explore data in a personalised way. In this paper, we propose a semantics-based approach for enabling personalised data lake exploration through the conceptualisation of proper indicators. In particular, the approach is structured as follows: (i) at the bottom, heterogeneous data sources within a data lake are enriched with Semantic Models, defined by domain experts using domain ontologies, to provide a semantic data lake representation; (ii) in the middle, a Multi-Dimensional Ontology is used by analysts to define indicators and analysis dimensions, in terms of concepts within Semantic Models and formulas to aggregate them; (iii) at the top, Personalised Exploration Graphs are generated for different categories of users, whose profiles are defined in terms of a set of constraints that limit the indicators instances on which the users may rely to explore data. Benefits and limitations of the approach are discussed through an application in the Smart City domain.

Keywords: Semantic data lake · Data exploration · Smart City

1 Introduction

Modern organisations are spending time and resources in understanding their Big Data, collected within heterogeneous sources, attracted by the opportunity

of turning it into actionable insights. A representative example is given by Smart Cities, whose infrastructures comprise both private and open data, legacy systems as well as new generation devices to collect city and citizens' consumption data, using different data formats, models and storage systems [5]. Data lakes, which follow a schema-on-read approach, have been proposed for storing both structured and unstructured data, regardless its structure. For end-users, data aggregation according to different analysis dimensions may help to take decisions in a more efficient manner. To this aim, indicators are powerful tools for meaningful data aggregation and exploration. However, they are usually computed based on the centralisation of data storage, according to a less flexible schema-on-write approach, and their definition in the context of data lakes deserves more investigation. Moreover, domain experts, who know data stored within the data lake, are usually distinct from data analysts, who define indicators, and users, who exploit indicators to explore data in a personalised way [15]. To this aim, Semantic Web technologies may come to the rescue, given their capability to enable interoperability over heterogeneous sources and improve data access; specifically, recent research efforts strove to combine the OBDA paradigm with data lakes, leading to the so-called *semantic data lakes*, with the development of semantic layers apt to uniformly access data lakes content [14]. In this context, the interest towards exploration techniques for multi-dimensional data (such as the ones collected within a Smart City), leveraging domain knowledge, is growing faster and faster [1]. Indeed, the objective is to capture users' analytical requirements, subject to a continuous and endless evolution over time.

Focusing on the exploratory aspect, the goal of this paper is to describe a semantics-based approach to enable personalised data exploration through the conceptualisation of proper indicators, built on top of a semantic data lake representation. In particular, the approach is structured over layers as follows: (i) at the bottom, heterogeneous data sources within a data lake are enriched with Semantic Models, defined by domain experts using domain ontologies; (ii) in the middle, a Multi-Dimensional Ontology is used by analysts to define indicators and analysis dimensions, in terms of concepts within Semantic Models and formulas to aggregate them, in order to extract what we call Smart City Exploration Graph; (iii) at the top, Personalised Exploration Graphs are generated for different categories of users, whose profiles are defined in terms of a set of constraints that limit the indicators instances on which the users may rely to explore data. The novel contribution of this approach mainly concerns the two upmost layers of the information model introduced above. These layers enable a progressive (and evolving) extraction of Smart City knowledge from the semantic data lake through the semantic modelling of indicators and the exploitation of users' profiles. Furthermore, the three layers enable the clear separation of roles, namely domain experts, data analysts and end-users.

The paper is organised as follows: in Sect. 2 a motivating example in the Smart City domain is provided; the approach overview is explained in Sect. 3, while Sects. 4, 5 and 6 describe in details the three layers of the information model; Sect. 7 reports personalised exploration scenarios over the semantic data

lake; Sect. 8 discusses benefits and limitations of our approach; cutting-edge features of the approach with respect to the literature are discussed in Sect. 9; finally, Sect. 10 closes the paper.

2 Motivating Example

Different categories of users may be interested in exploring urban data: mobility managers, energy managers, utility and energy providers, building managers and citizens. As a motivating example, let's consider John, the manager of three buildings located in different districts of a Smart City. John is interested in monitoring the electrical energy consumption of the buildings, in order to plan saving policies and effective renovation interventions for the buildings with highest consumption. Among buildings administered by John, `Building1` and `Building2` are located in `District1`, whereas `Building3` is located in a different district (`District2`). In the example, energy consumption of a building is computed by defining an indicator ($EC_{building}$) obtained as the sum of the electrical consumption of the stairs (EC_{stairs}) and other common spaces of the building (such as gardens, $EC_{gardens}$) and the electrical consumption of building elevators ($EC_{elevators}$), that is $EC_{building} = EC_{stairs} + EC_{gardens} + EC_{elevators}$. Precisely, the factors in the aforementioned sum are in turn indicators that aggregate data stored within distinct data sources. For example, $EC_{elevators}$ is extracted from a relational database containing the electricity bills for all the three buildings. The other two indicators are measured in a different way, using new generation smart meters to extract a new value and store it within a JSON file. Given the quantity and heterogeneity of data in this motivating example, indicators are effective tools to provide John with meaningful aggregation of data of interest. Moreover, for privacy preservation issues, John should not be allowed to inspect data at the highest granularity level. He should explore data at building level (and only for administered buildings).

3 Approach Overview

The three-layered information model proposed in this paper organises knowledge about Smart City data sources through the following phases: (i) semantic modelling of data lake sources, which defines a Smart City Semantic Data Lake, (ii) semantic representation of indicators, which define a Smart City Exploration Graph, and (iii) selection of Personalised Exploration Graphs. Figure 1 shows the components of the information model, the involved roles, and architectural modules that implement the three phases.

Starting from the bottom, Smart City heterogeneous data sources (e.g., IoT devices, sensors, energy providers databases) are conveyed, through proper ELT (Extract, Load and Transform) processes, into the Data Lake, where raw data is stored as values of a set of attributes defined in the source. Semantic annotations of data in a source, based on the concepts of one or more domain ontologies, define a corresponding *Data Source Semantic Model*, which is stored in the *Smart*

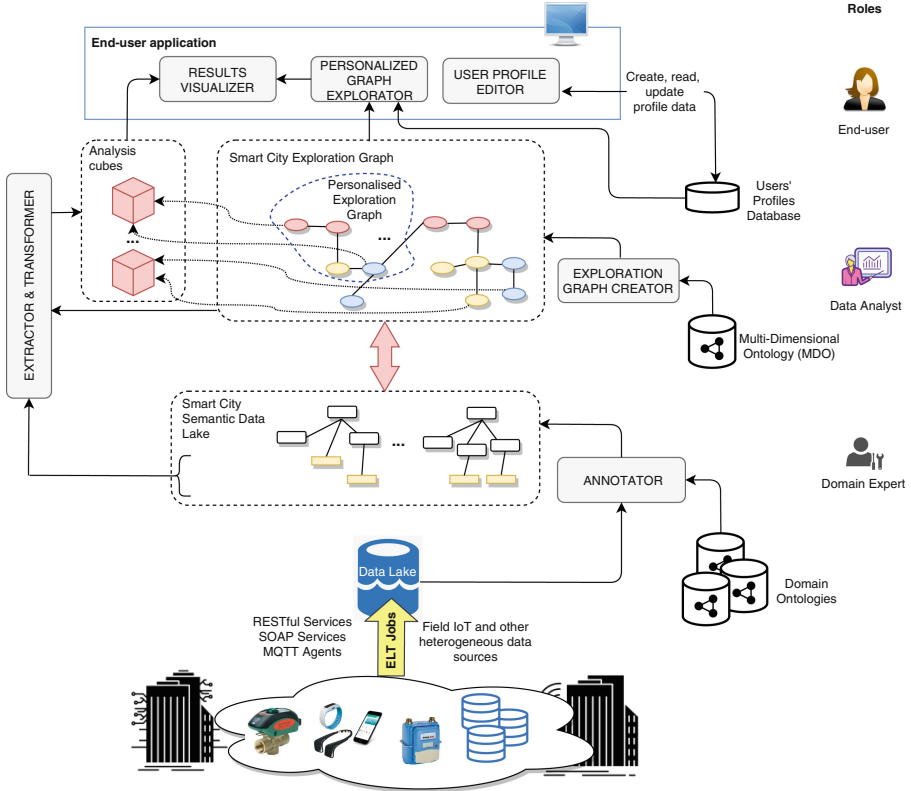


Fig. 1. Approach overview.

City Semantic Data Lake. Here, available data sources are described according to their meanings, rather than on their raw data attributes. Domain experts are in charge both to identify suitable domain ontologies and to use them for defining the Data Source Semantic Models, by adopting methodologies defined and validated in the literature [15]. A semi-automatic tool, ANNOTATOR, supports the domain expert in this task.

Semantic Models are the starting point for defining the *Smart City Exploration Graph*; in fact, by leveraging them, a data analyst can formally define Smart City indicators and detail their features (e.g., dimensions, computing formula, unit of measure). Base concepts and relationships required for modelling indicators are defined in a *Multi-Dimensional Ontology (MDO)*. Hence, the Smart City Exploration Graph contains knowledge about indicators, their dimensional characterisation and formulas. The semantic description of an indicator is obtained as an aggregation of other indicators (according to the formula) in a recursive way, until the most fine grained concepts are reached, corresponding to concepts in the Data Source Semantic Models. In this way, a set of mappings is established between concepts in the Smart City Exploration

Graph and concepts in the Semantic Models. An **EXPLORATION GRAPH CREATOR** tool supports the data analyst in semantic definition of indicators and creation of mappings. Analysis Cubes, shown on the left part of Fig. 1, provide actual values of indicators and related dimensions as defined in the Smart City Exploration Graph. Analysis Cubes are created and maintained with the support of an **EXTRACTOR & TRANSFORMER** tool. For each indicator, the tool uses both semantic definitions of indicators, mappings towards Semantic Models, and additional knowledge provided by the data analyst about transformation processes in order to materialise the cubes. Moreover, indicators (e.g., $EC_{building}$) and dimension levels (e.g., **District**) are mapped to corresponding attributes of the Analysis Cube schema (i.e., its star schema).

Portions of the Smart City Exploration Graph, the so-called *Personalised Exploration Graphs*, are dynamically selected, for different categories of users. Dynamic selection is based on both the current Smart City Exploration Graph content and on the users' profiles, and supports personalised and interactive exploration of Smart City indicators. The Personalised Exploration Graphs can be explored by end-users in various ways (e.g., as a interactive graph structure or by inspecting graph nodes, a node at a time, visualised as a structured record¹) through a GUI, which also enables users to register themselves and create/update their profile, stored in the users' profiles database. In particular, end-users may use proper tools to: (i) learn about indicators description in the Personalised Exploration Graph (using the **PERSONALIZED GRAPH EXPLORATOR**); (ii) select indicators and dimension levels, and see their values in a tabular format (using the **RESULTS VISUALIZER**); (iii) register to the platform by creating a profile (using the **USER PROFILE EDITOR**).

In the following sections, the key elements of the approach will be detailed also with reference to the motivating example.

4 Smart City Semantic Data Lake

We represent the Data Lake DL as a set of *data sources*, namely $\langle \mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_n \rangle$. Each data source \mathcal{S}_i is in turn composed of a collection of *data items*, whose definition is given as follows.

Definition 1 (Data item). *A data item $\mathcal{D}_j^i \in \mathcal{S}_i$ can be formalised as $\mathcal{D}_j^i = \langle \mathcal{A}_j^i, \mathcal{O}_j^i \rangle$ where: (i) \mathcal{A}_j^i is a set of attributes and (ii) \mathcal{O}_j^i is a set holding attribute-value pairs, with attributes belonging to \mathcal{A}_j^i , representing the content of the data item.*

Example 1. In the considered motivating example, a data item \mathcal{D}_j^i may contain the electrical energy consumption sensed by a smart meter installed at a specific building. Coherently, for the data item, candidate attributes for the set \mathcal{A}_j^i would be timestamp, smart meter configuration parameters (e.g., the physical

¹ <http://vowl.visualdataweb.org/protegevowl.html>.

position of the meter) and so forth, that is, $\mathcal{A}_j^i = \{\text{Electr_Rdng}, \text{eMeter_ID}, \text{building_ID}, \dots\}$, where `eMeter_ID` denotes the unique ID associated with a particular electricity meter, `building_ID` denotes the identifier of the building where the meter is located and `Electr_Rdng` denotes the electrical energy reading at a specific time on the meter. Given the attributes set \mathcal{A}_j^i , a sample of the content of set \mathcal{O}_j^i is: $\{\langle \text{Electr_Rdng}, 10023.5 \rangle, \langle \text{eMeter_ID}, \text{RGO16} \rangle, \langle \text{building_ID}, \text{Building001} \rangle, \dots\}$.

4.1 Data Sources Semantic Models

Once data has been ingested in the Data Lake, domain experts are enabled to build proper *Semantic Models*, created during the addition of the data source into the Data Lake. These models serve as a semantic description of data attributes and their relationships, allowing data analysts, who might not be familiar with the data, to understand data meaning. In fact, concepts and relationships used within each Semantic Model represent the domain expert’s view of the data source, based upon one or more (external) domain ontologies. In our approach, data lake sources are semantically annotated in order to define the *Smart City Semantic Data Lake* in terms of Semantic Models and their instances. For this purpose, we rely on techniques and models developed and tested in [15]. In particular, we design a Semantic Model for each data source $\mathcal{S}_i \in DL$ according to the following steps: (1) at the beginning, the domain expert considers all the attributes of the data source \mathcal{S}_i and then seeks suitable semantic concepts and relationships amongst the ones available in the domain ontologies; (2) after that, the domain expert may specialise, if needed, the existing concepts and relationships in the domain ontologies by adding specific ones; (3) mappings are defined between concepts and attributes to automatically extract instances (i.e., individuals) of the aforementioned concepts from the underlying data sources.

A fragment of Semantic Model for the motivating example is depicted in Fig. 2. The domain ontologies chosen by the domain expert for this Semantic Model are: GeoNames², for representing spatial entities, SOSA³ (acronym for Sensor, Observation, Sample and Actuator) containing concepts apt to describe sensors and measures produced by them, and Time⁴, devoted to model temporal entities.

5 Smart City Exploration Graph

Semantic Models provided by domain experts are leveraged by data analysts to assure a formal definition of Smart City indicators. Specifically, data analysts are in charge of designing the aforementioned indicators (jointly with additional knowledge) by specialising concepts and relationships of a *Multi-Dimensional*

² <http://www.geonames.org/> (prefix: `GEO`).

³ <http://www.w3.org/ns/sosa> (prefix: `sosa`).

⁴ <http://www.w3.org/2006/time> (prefix: `time`).

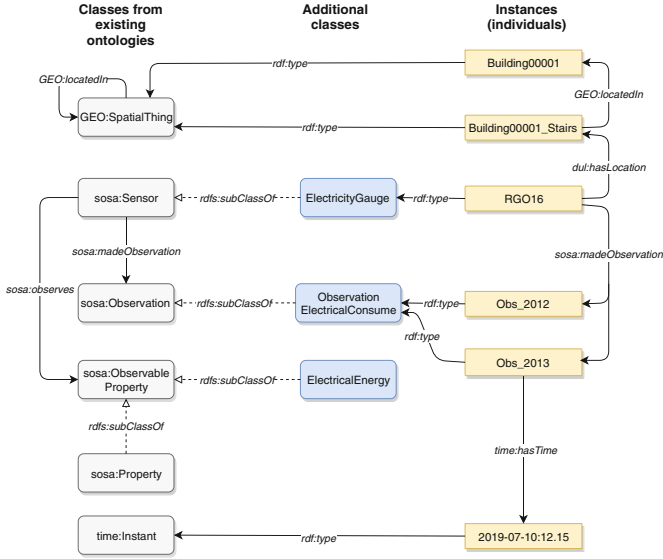


Fig. 2. Semantic Model designed for a data source in the motivating example.

Ontology (MDO, in brief) conceived as a modelling framework to provide a semantic representation of the indicators which have to be defined.

The semantic definition of indicators, obtained by specialising MDO concepts and relationships, weaves the *Smart City Exploration Graph*. In the following, the description of the MDO, emphasising its semantic areas, is provided, along with an example of indicator definition, related to the motivating example in Sect. 2.

5.1 The Multi-Dimensional Ontology

As shown in Fig. 3, the MDO contains baseline concepts and relationships regarding indicators and exploration personalisation aspects. Hereafter, the three conceptual areas of the MDO are described, emphasising the semantic role they fulfill.

Personalisation Concepts. Owing to the wide variety of data that can be explored, data exploration can be personalised taking into account user’s category (e.g., building manager, citizen) and activities performed by users (e.g., building monitoring, check air pollution), the former abstracted through the `UserCategory` concept, whereas the latter through the `Activity` concept. An activity may involve one or more indicators to be observed, compatible with its purpose (`involves` relationship). Activities can be performed by users, according to their role inside the Smart City context (`hasPracticableActivity` relationship).

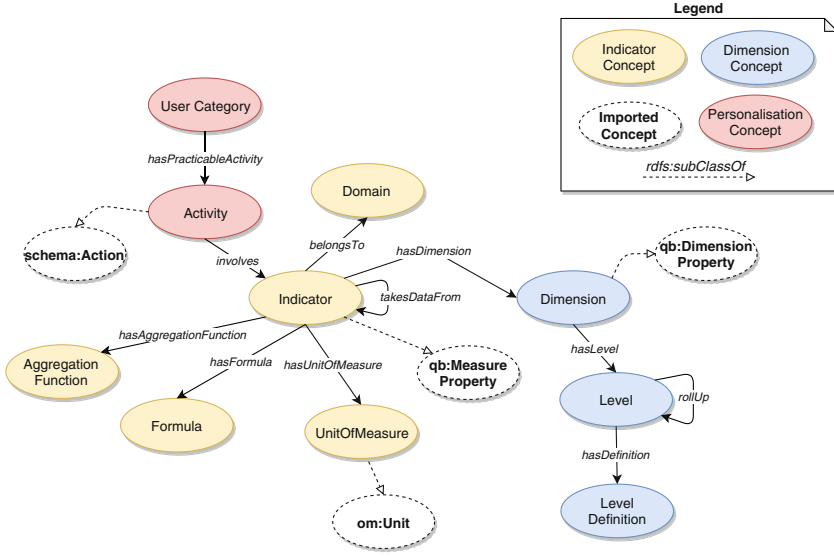


Fig. 3. TBox of the Multi-Dimensional Ontology (in the figure, the three semantic areas are highlighted with different colours).

Dimensions Concepts. Indicators can be segmented and calculated according to several dimensions (`hasDimension` relationship), which are in turn organised into levels, modelled through the `Level` concept. The `rollUp` semantic relationship enables a hierarchical organisation of levels, thus ensuring the navigability from the lowest to the highest aggregation level, resembling the renowned OLAP *roll-up* operator. Moreover, `Level` concept is linked in the MDO to `LevelDefinition` concept through the `hasDefinition` relationship, in order to represent an abstraction for the procedure devoted to calculate level instances starting from their counterparts in the semantic models.

Indicators Concepts. Indicators are designed to aggregate urban data according to several perspectives and represent the ultimate target for data exploration, as their value may give valuable insights to users. In the ontology, indicators can be modelled as simple or compound, i.e. calculated exploiting other indicators (through `takesDataFrom` semantic relationship). Being measurable quantities, they are endowed with a unit of measure (abstracted through the `UnitOfMeasure` concept) and have a calculation formula (`Formula` concept). For an indicator, it is also possible to define an aggregation function (`AggregationFunction` semantic concept), adopted as standard operation to apply whenever switching from a dimensional level to another (through the `rollUp` semantic relationship). Finally, indicators may be associated with domains, such as environment, health, energy and so forth; this is represented inside the ontology through the `belongsTo` semantic relationship, connecting `Indicator` concept with `Domain` concept.

The design of the MDO (especially for **Dimension** concepts) has been inspired by the existing literature concerning OLAP paradigm constructs [6]. Moreover, for defining the ontology, pivotal concepts from available foundation ontologies have been exploited to: (i) represent users' activities ([Schema.org](https://schema.org/) ontology⁵), (ii) characterise indicators and dimensions as analytical data entities (Data Cube ontology⁶) and (iii) model units of measure for indicators (ontology of units of measures⁷).

5.2 Semantic definition of indicators through the Multi-Dimensional Ontology

Given the knowledge structure in the MDO, data analysts are enabled to semantically describe indicators through the specialisation of MDO concepts and relationships. In the following, an example of semantic definition is provided (Fig. 4), referring to the indicator **EC_Building** in the motivating example. For clarity purposes, in the figure, specialisation relationships are omitted and concepts are represented using colours that match the colours of the semantic areas in Fig. 3.

Example 2. According to the motivating example, the **EC_Building** indicator is compound, as it depends on (i.e., **takesValueFrom**) three indicators, namely **EC_Stairs**, **EC_Gardens** and **EC_Elevators**, whose sum contributes to produce the value for **EC_Building**. Indicators concepts are linked to a **SpatialDimension**, suitable for defining their spatial coverage, through the specialisation of **Level** concept. In the motivating example, the following levels of spatial dimensions are considered: **City**, **District**, **Building** and **ElectricalMeter**. Analogously to what happens for OLAP analysis, **rollUp** relationships ensure to establish a hierarchical organisation of levels concepts, assuring the navigability from the lowest (i.e., **ElectricalMeter**) to the highest (i.e., **City**) extent of aggregation. Focusing on **Building** concept, it is linked to **BuildingLevelDefinition** concept; this is exploited to formalise the calculation of **Building** instances starting from their conceptual counterpart in semantic models (that is, referring to Fig. 2, instances of **SpatialThing** concept). Noteworthy, different indicators concepts may share the same dimensional organisation (that is, **EC_Building**, **EC_Stairs**, **EC_Gardens** and **EC_Elevator** are all connected to **SpatialDimension** concept).

Mappings Towards Semantic Models and Analysis Cubes. We distinguish among two types of mappings, manually defined by data analysts, binding: (i) a concept in the Smart City Exploration Graph with a concept in a Semantic Model and (ii) a concept in the Smart City Exploration Graph with an attribute in a specific Analysis Cube.

⁵ <https://schema.org/> (prefix: **schema**).

⁶ <https://www.w3.org/TR/vocab-data-cube/> (prefix: **qb**).

⁷ <http://www.ontology-of-units-of-measure.org/resource/om-2/> (prefix: **om**).

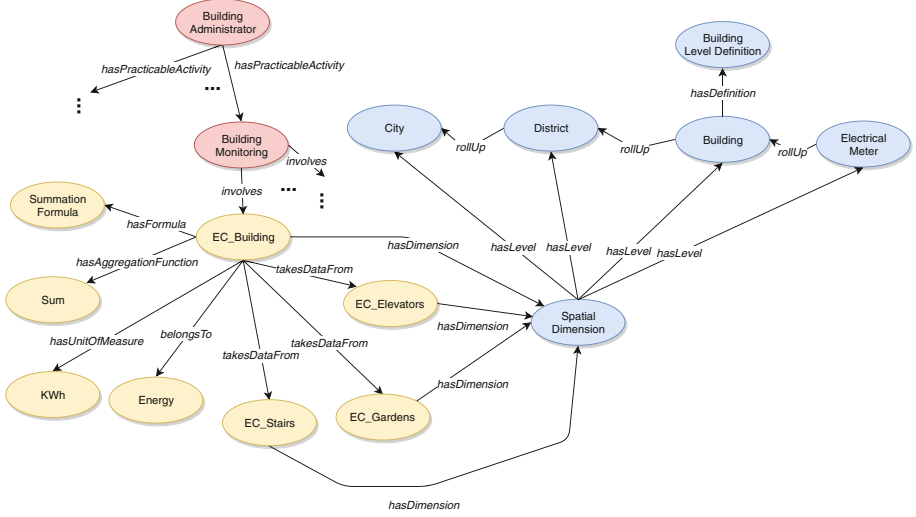


Fig. 4. Excerpt of the Smart City Exploration Graph for the description of $EC_{building}$ in the motivating example.

Specifically, the first kind of mapping is a tuple $\langle C_{EG}, C_{SM}, R \rangle$ binding a concept C_{EG} contained in the Smart City Exploration Graph to a concept C_{SM} belonging to a Semantic Model, through the relationship R (e.g., equivalence or subsumption relationship). In this case, data analysts may be supported in their definition by ad-hoc tools⁸, which may be equipped with similarity-based matching algorithms, and ensure the possibility of exporting such mappings in various formats.

Similarly, mappings between concepts in the Smart City Exploration Graph and Analysis Cubes are expressed through the tuple $\langle C_{EG}, c_{ID}, c_{ATTR} \rangle$, where C_{EG} represents a reference to a concept in the Smart City Exploration Graph, c_{ID} is the Analysis Cube unique identifier and c_{ATTR} is an attribute contained in the Analysis Cube schema. These mappings are retained in a dedicated repository and processed by the EXTRACTOR & TRANSFORMER tool.

6 Personalised Graphs Extraction

In this section, we describe how to extract Personalised Exploration Graphs, containing a portion of the Smart City Exploration Graph targeted to a specific end-user. Specifically, the aim of this task is to assure a personalised exploration of Smart City indicators, achieved by considering users' profiles. In the following, a formalisation of users' profiles is given, together with the description of the stepwise procedure leading to the extraction of Personalised Exploration

⁸ For instance, Protégé (<https://protege.stanford.edu/>) and COMA tool (<https://dbs.uni-leipzig.de/Research/coma.html>).

Graphs containing the set of indicators to be explored, along with their semantic characterisation.

6.1 Users' Profiles

The profile of a user $p(u)$ can be abstracted through the tuple $\langle ID_u, cat_u, C_l, I_l \rangle$, composed of: the unique identifier associated with the user (ID_u), a concept representing the category of the user, i.e. his/her role inside the Smart City (cat_u), a set of dimensional level concepts constraints (C_l) and a set of dimensional levels instances (I_l). Users' profile data is stored in a database; all the elements in $p(u)$ are used during the Personalised Graph extraction process, detailed in the following section, except from the set I_l , exploited at exploration time.

Precisely, the category concept cat_u is used to retrieve, from the Smart City Exploration Graph, activities concepts compliant with user's role (from the motivating example, **BuildingMonitoring** is an activity concept available for the role of John, as he belongs to **BuildingAdministrator** category), which will in turn determine the indicators concepts involved in such activity. The set C_l represent concepts deemed as constraints, applicable to delimit the dimensional levels to which the user is allowed to access. In this way, data privacy preservation is enabled, for instance granting building managers to visualise data up to building level (therefore not at lower levels, such as single electrical meters).

When users register themselves through the GUI, the registration wizard, starting from the user's category, prompts to the user proper masks to insert in C_l concepts references and in I_l instances references, which will be in turn validated by a third-party authority, in order to assure the exploration of indicators only for the dimensional levels the user has an explicit authorisation and for the buildings under his competency (e.g., buildings in the motivating example).

6.2 Personalised Graphs Derivation Procedure

As previously remarked, from the Smart City Exploration Graph, different personalised graphs can be derived, by considering users' profiles. The extraction process can be summarised by these two steps:

- (1) *Activity-based indicators concepts selection*, in which, besides the user selects the activity he/she wants to perform, relying on his/her role, candidate indicators concepts (and related semantic knowledge) involved in the activity are identified and retrieved from the Smart City Exploration Graph;
- (2) *Level-based dimension pruning*, in which concepts specialising the MDO **Level** concept, retained in the set C_l , are used to constrain the visibility of the dimensional organisation levels concepts only to those levels a user is allowed to access.

In the following, each step will be further described, considering the motivating example and with the support of Fig. 5.

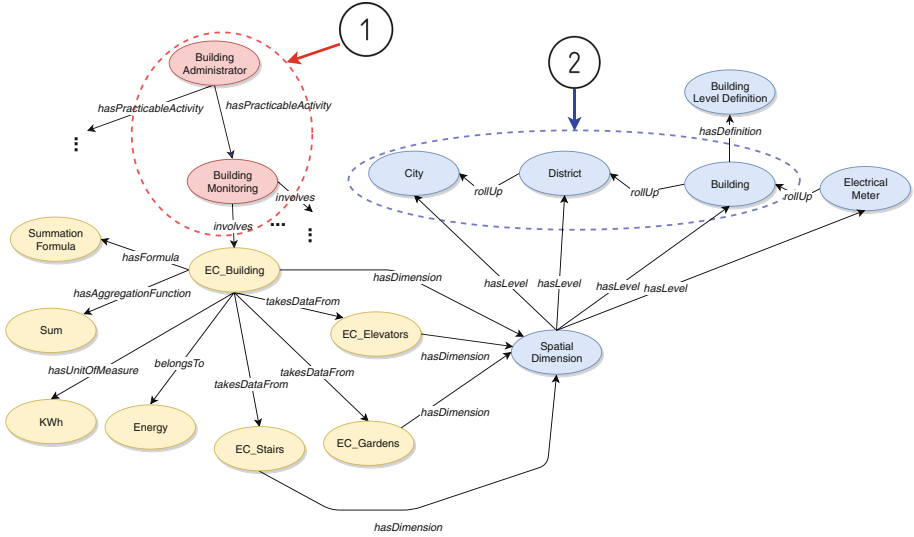


Fig. 5. Stepwise derivation of a Personalised Graph. In the figure, the focus is on a specific portion of the Smart City Exploration Graph (numbers in the circles represent the order of extraction steps).

Activity-based indicators concepts selection. Due to the fact that John is a building administrator, he must select only activity concepts compliant with his role. Amongst the available ones, he selects from the proposed ones **Building-Monitoring**. Such activity involves several indicators and **EC_Building**, as previously described, is one of them. Being **EC_Building** a compound indicator, the **takesDataFrom** relationship connects it to the concepts **EC_Stairs**, **EC_Gardens** and **EC_Elevators**. Noteworthy, the following dimensional pruning step is iterated over the aforementioned indicators. Overall, for these composing indicators and for the ancestor one, the related semantic characterisation (dimensional organisation, formula, aggregation function, domain and so forth) is retrieved from the Smart City Exploration Graph.

Level-based dimension pruning. Retained in John’s profile data, the set C_l expresses the dimensional levels (that is, references to concepts specialising **MDO Level** concept) to which he may access. Referring to John, the set C_l in his profile holds the concepts $\{\text{City, District, Building}\}$. Furthermore, such elements are linked in a **rollUp** chain (starting from **Building** up to **City**), defining the dimension levels navigation path.

7 Personalised Graphs Exploration

Exploration within Personalised Graphs can be performed according to different perspectives (highlighted in Fig. 6), thus enabling the Explorative Search

paradigm [4]: (a) exploration over the indicators involved in the performed activity, (b) exploration over the indicators dependencies and (c) exploration over the indicators dimensional organisation. Regardless the type of exploration scenario, the user selects an indicator node and one or more dimensional levels nodes from his/her Personalised Exploration Graph. According to this selection, the content of the set $I_l \in p(u)$ is exploited to limit the set of visible instances related to the selected levels concepts. As a result, a query over the Analysis Cube corresponding to the selected indicator is issued, in order to show (e.g., through a tabular layout) the instances of the indicator and the selected dimensions.

Exploration Over the Indicators Involved in the Performed Activity. An activity performed by a user, such as **BuildingMonitoring**, involves one or more indicators to be inspected (related to energy and water consumption). Nevertheless, for clarity purposes, we will focus only on the **EC_Building** and its associated indicators, adhering to the motivating example, which have been considered for the next two exploration scenarios.

Exploration Over the Indicators Dependencies. In this exploration scenario, John selects the **EC_Building** indicator and, starting from that, he can delve into other indicators by traversing the **takesDataFrom** semantic relationship, connecting a compound indicator concept with its dependencies (which in turn may be compound or not). In this respect, John can inspect the more specific indicators **EC_Stairs**, **EC_Gardens** and **EC_Elevators**. For instance, if John’s focus is on evaluating the electrical consumption of the elevators of the building, he may select and inspect details about **EC_Elevators** indicator.

Exploration Over the Indicators Dimensional Organisation. This exploration scenario exploits the semantic relationships between an indicator and its dimensional organisation (through the **hasDimension** relationship). In particular, knowledge on dimensional organisation for indicators considers levels concepts (**hasLevel** relationship). Focusing on **EC_Building** indicator or one of its dependencies, John can position himself over one dimensional level (e.g., **Building**) and then he can change the granularity of the aggregation level for the chosen indicator, by exploiting the **rollUp** semantic relationship, thus switching to **District**. Therefore, John could choose to visualise at a higher aggregation viewpoint (in this case, district) the values of the **EC_Building** indicator; whenever such **rollUp** is performed, the aggregation function associated with the indicator, through the **hasAggregationFunction** relationship (e.g., “sum” in the case of **EC_Building**), is applied.

8 Preliminary Validation

Advantages on the use of a semantic layer on top of the data lake, composed of different Semantic Models managed by domain experts who know data stored within the data lake, have been already investigated in [15], both in terms of flexibility, expressiveness and expandability. Following the suggestions given

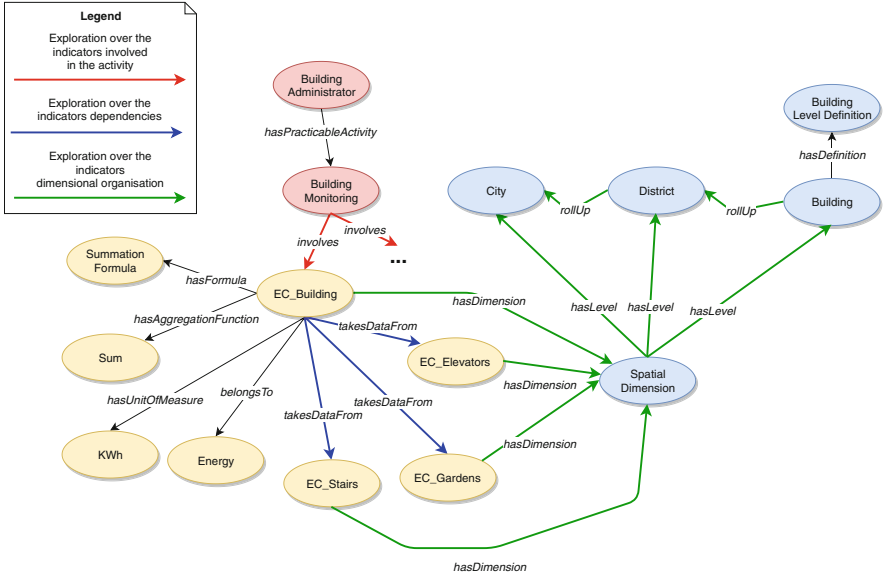


Fig. 6. Subset of a Personalised Graph on which exploration scenarios are considered.

in [15], possible inconsistencies and heterogeneity across different Semantic Models, already mitigated due to the adoption of reference domain ontologies, can be further reduced through the application of similarity-based alignment and reconciliation techniques. The goal of this section is to highlight pros and cons of the three-layered approach proposed in this paper. In particular, the focus will be on the preliminary proof that the Smart City Exploration Graph and Personalised Graphs are effective in supporting different categories of users for indicators exploration.

To this aim, a prototype implementation of the approach has been created, focusing on the GUIs designed to accomplish the exploration tasks. Specifically, a web-based dashboard has been built in order to: (a) let users register themselves to the exploration platform, specifying elements in $p(u)$ and (b) explore their Personalised Graphs, generated according to their profile data. In the prototype implementation, data sources containing energy consumption in the context of the Brescia Smart Living project⁹, including sources considered in the motivating example, have been collected. A sample of users has been involved, including citizens without prior knowledge on Smart City indicators, representatives of Public Administration (in particular, the Municipality of Brescia, Italy), utility and energy providers. Citizens live in two districts of the city, a modern one (Sanpolino), where new generation smart meters have been installed, and a district in city downtown, more densely populated and presenting older buildings. Semantic Models, the Smart City Exploration Graph and Personalised Graphs

⁹ <https://www.bresciasmartliving.eu/index.php>.

extracted for users have been deployed in OWL using the Stardog¹⁰ Triplestore. The involvement of domain experts and data analysts in the definition of Semantic Models and semantic representation of indicators based on the MDO has been supported through the Protégé graphical editor. Moreover, visualisation of Personalised Graphs has been achieved by relying upon well-known libraries apt to provide a graphical representation of semantic graphs. Given the versatility of Stardog module, we used the *d3sparql*¹¹ library, suitable to represent dynamic and interactive visualisation starting from data (in this case, the Personalised Graphs). Once indicators and dimensions concepts are selected by the user from the Personalised Graph, the inspection of the associated analysis cube is achieved, in a tabular way, by means of the OpenCube Toolkit¹². Scalability issues concerning the querying of data lake based on the semantic layer have been already studied in several approaches according to the OBDA paradigm [14] and are out of the scope of this preliminary evaluation. They will be considered for future work.

Workload and Performance Analysis. To demonstrate the effectiveness of three-layered information model in supporting personalised indicators exploration, we simulated an exploratory search scenario, composed of three tasks of increasing difficulty, where users were allowed to explore their Personalised Graphs according to the exploration methods introduced in Sect. 7. Specifically, the three aforementioned tasks included the exploration of: (i) a simple indicator, (ii) a compound indicator with a single level of dependencies (like the EC_Building of the motivating example) and (iii) a compound indicator with more than two nested levels of dependencies. Users attended an initial training of 30 min, in order for them to get acquainted with the GUI and, at the same time, let them create their profile. For evaluating the performance of our prototype, participants filled a likert scale NASA TLX questionnaire, a multidimensional assessment tool that rates perceived workload in order to assess a task or system. Questionnaires analysis has revealed that, for our proposal, the “Mental Demand” sub-scale has been rated with the highest score whereas “Performance” with the lowest. We impute such outcome to the fact that our proposal increases the success of users while accomplishing data exploration, despite a slight growth of mental and perceptual activity required to understand how to effectively navigate across Personalised Graphs.

Usability Tests. Usability experiments are being carried on within the Brescia Smart Living project until the end of 2019. Participants were assigned a task to be accomplished using the exploration tool, without imposing time-up or any particular exploration constraint. Specifically, the task was a broad exploratory search scenario for inspecting indicators in the Personalised Graph concerning the energy and the environmental domain (compliant with user’s profile). Then, we asked participants to fill a standard System Usability Scale (SUS)

¹⁰ <https://www.stardog.com/>.

¹¹ <https://github.com/ktym/d3sparql>.

¹² <http://opencube-toolkit.eu/>.

questionnaire, as an instrument to assess software usability. In this kind of questionnaire, for each statement, respondents can indicate the degree of agreement and disagreement on a five point scale. SUS scores a software in a range between 0 to 100, where 0 indicates the least usability and 100 represents a high usability, respectively. We averaged over all participants questions and the resulting mean score amounted to 87.5, which situates our prototype in the 90–95 percentile range in the curve matching SUS score with the corresponding rank. We also calculated the time participants spent to carry out their task. Using the prototype, participants accomplished their task in a shorter timer compared to a plain keyword-based interface. This relieves participants of a detailed knowledge about indicators definition, being guided in the exploration of indicators by the semantic relationships in the Personalised Graph.

9 Related Work

In the last decades, Semantic Web technologies have been suggested to effectively combine data from multiple heterogeneous sources, with the aim of providing a unified view of data resulting from integration processes. In this respect, in-depth studies concerning the renowned OBDA (Ontology-Based Data Access) paradigm have been applied in the field of data lakes, paving the way to the so-called semantic data lakes [14], wherein heterogeneous data is seamlessly accessed and queried through a proper semantic layer.

Generally speaking, the role of semantics in data lake approaches is manifold and goes beyond integration issues. For instance, in [3, 11] semantic enrichment is exploited to link data with external knowledge bases, also envisaging probabilistic techniques [13]. Moreover, ontologies and knowledge graphs have been fostered in literature as a promising solution to offer a comprehensive view over the underlying data sources, modelling their relationships and dependencies [2, 11]. Likewise, in [17] knowledge graphs store data fragments, represented as different typologies of nodes inside the graph. The tool presented in [14] embraces Semantic Web facilities to answer on-demand queries over heterogeneous data, ensuring high levels of scalability when treating significant data volumes. Nonetheless, in the former approaches, there is no mention about fostering such knowledge to achieve a personalised Data Exploration experience, and the focus is more on assessing the performance of the systems. Furthermore, the separation of roles and competencies of the actors involved in the aforementioned frameworks is less evident, with respect to our multi-layer approach. Sharing some issues with the proposal of this paper, the approach presented in [7] harnesses thematic views upon the datasources of the data lake, but again personalisation aspects are treated only at a high level of abstraction, without a fully-fledged semantic support. Lastly, a semantic data platform has been presented in [15], adopting a flexible data ingestion pipeline, but exploration aspects have been slightly neglected.

For what concerns approaches focused on Data Exploration tasks in data lakes, the starting point for users is almost always a GUI, where they can

issue queries to perform a keyword-based search [3,8] or where a direct interaction through ad-hoc visualisation tools can be achieved [9]. Other approaches, instead, focus more on favouring the user in getting acquainted with the exploration interface, taking into account his interaction waiting tolerance [12] or by implementing proper caching strategies, to ensure high responsiveness levels [16]. The ultimate goal of Data Exploration is also to provide suitable instruments to obtain actionable insights related to the observed data; in [10], for instance, current sensor data is compared against simulated data, computed from previously stored data, in order to predict future behaviours and trends.

In the majority of these approaches, the focus is more on the visualisation of data, instead of proposing techniques to detect and attract user's attention on data suitable and compliant with user's exploration interests, also suggesting effective exploration directions based on user's profile.

10 Concluding Remarks

In this paper, we described a semantics-based approach to extract Exploration Graphs for enabling personalised data lake exploration. In particular, the approach is structured over three layers: (i) at the bottom, heterogeneous data sources within a data lake are enriched with Semantic Models, defined by domain experts using domain ontologies, to provide a semantic data lake representation; (ii) in the middle, a Multi-Dimensional Ontology is used to describe indicators and their analysis dimensions, in terms of concepts within semantic models and formulas to aggregate them (iii) at the top, Personalised Exploration Graphs are generated for different categories of users, whose profiles are defined in terms of a set of constraints that limit the indicators concepts on which the user may rely to explore data. We presented the approach, discussing its benefits and limitations through an application in the Smart City domain.

Future efforts will be devoted to provide a full-fledged implementation of the approach, detailing the experimentation steps and the enabling technologies. In this respect, scalability issues, deriving from the potential vast amount of annotated data coexisting in the Smart City Semantic Data Lake, deserve a thorough dissertation and a comparison with other similar approaches (for instance, [14]) that will be under the lens of forthcoming research studies. Likewise, the impact of Personalised Exploration Graphs evolution (due to modifications in users' profiles) will be further investigated, focusing on the influence it has on the exploration tasks.

References

1. Abelló, A., et al.: Using semantic web technologies for exploratory OLAP: a survey. *IEEE Trans. Knowl. Data Eng.* **27**(2), 571–588 (2014)
2. Alserafi, A., Abelló, A., Romero, O., Calders, T.: Towards information profiling: data lake content metadata management. In: *Proceedings of IEEE 16th International Conference on Data Mining Workshops (ICDMW 2016)*, Barcelona, Spain, pp. 178–185 (2016)

3. Beheshti, A., Benatallah, B., Nouri, R., Tabebordbar, A.: CoreKG: a knowledge lake service. *PVLDB* **11**(12), 1942–1945 (2018)
4. Buoncristiano, M., Mecca, G., Quintarelli, E., Roveri, M., Santoro, D., Tanca, L.: Database challenges for exploratory computing. *SIGMOD Rec.* **44**(2), 17–22 (2015)
5. Chauhan, S., Agarwal, N., Kar, A.: Addressing big data challenges in smart cities: a systematic literature review. *Info* **18**(4), 73–90 (2016)
6. Diamantini, C., Potena, D., Storti, E., Zhang, H.: An ontology-based data exploration tool for key performance indicators. In: *Proceedings of 22nd OTM Conference on Cooperative Information Systems (CoopIS 2014)*, Amantea, Italy, pp. 727–744 (2014)
7. Giudice, P.L., Musarella, L., Sofo, G., Ursino, D.: An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Inf. Sci.* **478**, 606–626 (2019)
8. Hai, R., Geisler, S., Quix, C.: Constance: an intelligent data lake system. In: *Proceedings of the 2016 International Conference on Management of Data (SIGMOD/PODS 2016)*, San Francisco, California, pp. 2097–2100 (2016)
9. Halevy, A.Y., et al.: Managing Google’s data lake: an overview of the GOODS system. *IEEE Data Eng. Bull.* **39**(3), 5–14 (2016)
10. Kasrin, N., Qureshi, M., Steuer, S., Nicklas, D.: Semantic data management for experimental manufacturing technologies. *Datenbank-Spektrum* **18**(1), 27–37 (2018)
11. Lytra, I., Vidal, M., Orlandi, F., Attard, J.: A big data architecture for managing oceans of data and maritime applications. In: *Proceedings of International Conference on Engineering, Technology and Innovation (ICE/ITMC 2017)*, Madeira, Portugal, pp. 1216–1226 (2017)
12. Maccioni, A., Torlone, R.: KAYAK: a framework for just-in-time data preparation in a data lake. In: Krogstie, J., Reijers, H.A. (eds.) *CAiSE 2018*. LNCS, vol. 10816, pp. 474–489. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_29
13. Malysiak-Mrozek, B., Stabla, M., Mrozek, D.: Soft and declarative fishing of information in big data lake. *IEEE Trans. Fuzzy Syst.* **26**(5), 2731–2747 (2018)
14. Mami, M.N., Graux, D., Scerri, S., Jabeen, H., Auer, S., Lehmann, J.: Squerall: virtual ontology-based access to heterogeneous and large data sources. In: *Proceedings of 18th International Semantic Web Conference (ISWC 2019)*, Auckland, New Zealand (2019, in press)
15. Pomp, A., Paulus, A., Kirmse, A., Kraus, V., Meisen, T.: Applying semantics to reduce the time to analytics within complex heterogeneous infrastructures. *Technologies* **6**(3), 86–114 (2018)
16. Skluzacek, T.J., Chard, K., Foster, I.: Klimatic: a virtual data lake for harvesting and distribution of geospatial data. In: *Proceedings of 1st Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS 2016)*, Salt Lake City, Utah, pp. 31–36 (2016)
17. Walker, C., Alrehamy, H.: Personal data lake with data gravity pull. In: *Proceedings of 2015 IEEE Fifth International Conference on Big Data and Cloud Computing (BDCLOUD 2015)*, Dalian, China, pp. 160–167 (2015)