Personalised Exploration Graphs on top of Data Lakes

(Discussion Paper)

Devis Bianchini[1], Valeria De Antonellis[1] and Massimiliano Garda[1,*]

[1]*University of Brescia, Dept. of Information Engineering*
Via Branze 38, 25123 - Brescia (Italy)

Abstract

The volume, velocity and uncontrolled variety of Big Data are changing the way data exploration for data-driven decision making is performed on top of Data Lakes. As data grows, novel methods are needed for data aggregation by means of indicators and multi-dimensional analysis of Data Lakes content, enabling exploration of data according to various dimensions, thus empowering users with diverse roles and competencies to capitalise on the available information. In this paper, we present a computer-aided approach (named PERSEUS, PERSonalised Exploration by User Support) for data exploration on top of a Data Lake. The approach is structured over three phases: (i) the construction of a semantic metadata catalog on top of the Data Lake; (ii) the creation of an Exploration Graph, based on metadata contained in the catalog, containing the semantic representation of indicators and analysis dimensions; (iii) the enrichment of the definition of indicators with personalisation aspects (based on users' profiles and preferences) to identify Exploration Contexts, in turn delimiting portions of the Exploration Graph for a personalised and interactive exploration of indicators. Results of an experimental evaluation in the Smart City domain are presented with the aim of demonstrating the feasibility of the approach.

Keywords

semantic data lake, personalised data exploration, OLAP, Big Data

1. Introduction

In dynamic and rapidly evolving environments permeated by the volume, velocity and uncontrolled variety of Big Data, Data Lakes have been proposed as ground-breaking solutions to develop applications for data-driven decision making. Data Lakes ensure a suitable degree of flexibility for managing different types and formats of data sources, since data is loaded "as is" and transformed only when it becomes necessary [1]. However, as data grows, novel methods are needed to extract value from Data Lakes content, aggregating data into indicators according to various dimensions, thus empowering users with diverse roles and competencies to explore available information. In this paper, we present a computer-aided approach (named PERSEUS, PERSonalised Exploration by User Support) for data exploration on top of a Semantic Data Lake. The approach is structured over three phases: (i) the construction of a semantic metadata catalog on top of the Data Lake; (ii) the creation of an Exploration Graph, based on metadata catalog, containing the semantic representation of indicators and analysis dimensions; (iii) the enrichment of the definition of indicators with personalisation aspects (based on users' profiles

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
devis.bianchini@unibs.it (D. Bianchini); valeria.deantonellis@unibs.it (V. De Antonellis); massimiliano.garda@unibs.it (M. Garda)

and preferences) to identify Exploration Contexts, in turn delimiting portions of the Exploration Graph for a personalised and interactive exploration of indicators. An extended version of this work has been presented in [2], where we validated the approach in the scope of the Brescia Smart Living project [3]. The aim of the project was to enable citizens, energy providers and Public Administration to explore heterogeneous information available in the context of a Smart City, at different levels of aggregation, for making decisions and promoting virtuous behaviour in using private and public resources. The paper is organised as follows. Sections 2–4 describe the phases of the PERSEUS approach. An excerpt of the implementation details and of the experimental evaluation is reported in Section 5. Section 6 reviews the state of the art. Finally, Section 7 closes the paper, sketching future research directions.

2. Semantic Data Lake construction

We model a Data Lake as a set of N data sources $\mathcal{S}_i$, each one modelled as $\langle \mathcal{A}_i, \mathcal{DS}_i, \mathcal{M}_i \rangle$, where: (i) $\mathcal{A}_i$ is a set of *attributes*; (ii) $\mathcal{DS}_i$ is a collection of *data sets*, representing the content of the data source regardless its nature (i.e., structured, semi-structured, unstructured); (iii) $\mathcal{M}_i$ is a set of attribute-value pairs containing metadata apt to access the source (e.g., username, password) and other source-specific metadata. Each data set $ds_i^j \in \mathcal{DS}_i$ is defined over a set of attributes $\mathcal{A}_i^j \subseteq \mathcal{A}_i$. An attribute can be either: (i) a simple attribute or (ii) an attribute referencing another data set in the same data source (nesting).

The domain expert is in charge of creating the semantic metadata catalog by means of a web-based tool supporting basic annotation tasks. The annotation procedure regards only attributes names and not their values, thus reducing the annotation burden. The steps for the creation of the catalog are performed incrementally, as soon as new data sources are added.

Lexical enrichment of data source attributes. Each attribute a_k of a data source is associated with a label referred to as *Entity Property*, to reduce the gap between the attribute name and names of concepts used for semantic annotation. To this aim, domain experts are supported by two external linguistic APIs, conceived to complement each other: (i) an Abbreviations API [4], providing a dictionary of acronyms and their expansion, and (ii) WordNet [5], the widely adopted lexical database.

Semantic annotation of data source attributes. Starting from the Entity Property, the web-based tool retrieves a suitable *concept* describing the meaning of the attribute a_k. To this aim, a set of domain ontologies stored within an open access repository (LOV - Linked Open Vocabularies [6]) is accessed through a proper API to search for semantic concepts whose names match the Entity Property label. The top-ranked concept is automatically proposed for the annotation to the domain expert, who may revise the annotation.

Semantic metadata catalog population. The semantic metadata catalog constructed over the Data Lake contains: (i) the set of concepts annotating attributes of data sources; (ii) equivalence relationships between pairs of concepts, either associated with the same data source or different data sources, which are suggested relying on the metadata set $\mathcal{M}_i$ (e.g., when the concepts annotate attributes belonging to two tables of a relational database) or manually defined by the domain expert (e.g., when involved attributes belong to different data sources).

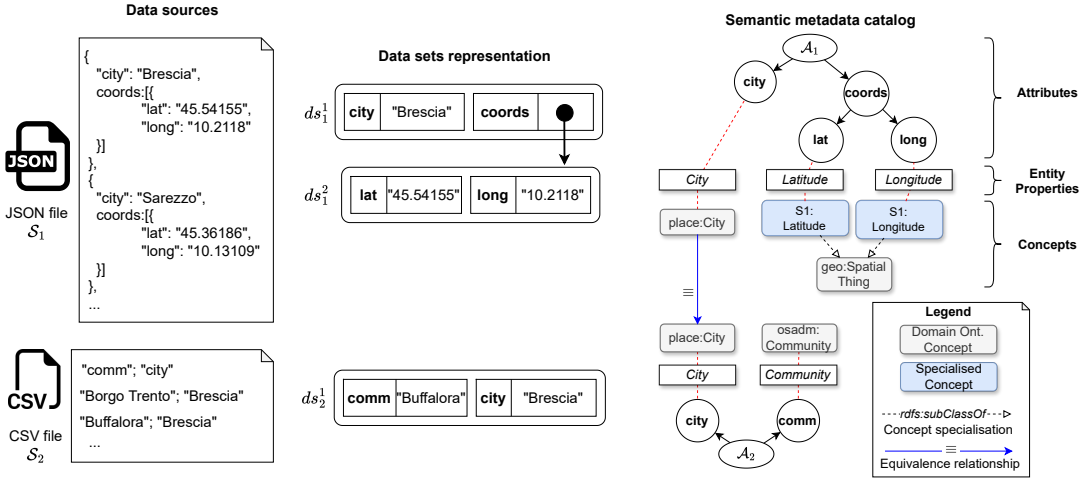

Figure 1: Examples of data sources, data sets and semantic metadata catalog.

Example. The left side of Figure 1 illustrates examples of two Smart City data sources and their representation as attributes and data sets. The two sources contain geospatial information of cities and related administrative areas. The right side of Figure 1 shows the semantic representation of the sources in the semantic metadata catalog. The Entity Properties are retrieved from WordNet (e.g., for `country` attribute) and the Abbreviations API (e.g., `lat` and `long` attributes). To find suitable concepts for semantic annotation, the LOV Search Term API is invoked using the Entity Properties as query parameters. Two concepts (`Latitude` and `Longitude`) have been obtained as a specialisation of the ones extracted from LOV ontologies (through the `rdfs:subClassOf` semantic relationship). In the figure, blue arrows denote equivalence relationships between concepts.

3. Creation of the Exploration Graph

Indicators are modelled by data analysts starting from the knowledge retained in the semantic metadata catalog and through the *specialisation* of concepts and relationships of a *Multi-Dimensional Ontology* (MDO), containing the conceptual elements that must be taken into account to model indicators. In the design of the MDO, pivotal concepts from available foundation ontologies have been exploited to: (i) represent users' activities (Schema.org ontology), (ii) characterise indicators and dimensions as analytical data entities (Data Cube ontology) and (iii) model units of measure for indicators (OM ontology). Further details regarding the conceptual elements of the MDO can be found in [2]. The result of this phase is an *Exploration Graph* $\mathcal{G}$ (an example is given in Figure 2, whose construction follows the steps reported below and it is accomplished with the support of the Protégé tool [7].

Creation of indicator concept. In this step the `Indicator` concept of the MDO is specialised to extend the indicators hierarchy. Using the `takesDataFrom` semantic relationship, composite indicators can be defined starting from other fine-grained indicators. For a newly created

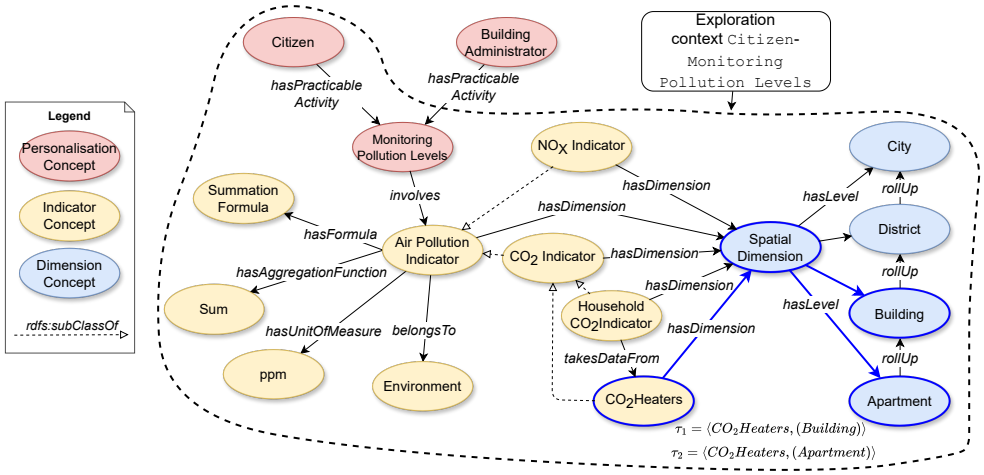

Figure 2: Portion of the Exploration Graph with an example of Exploration Context.

indicator, a `Formula` (that, for a composite indicator, reports how to calculate it in terms of its component indicators), the `UnitOfMeasure` and the `AggregationFunction` are specified.
Link to dimensional hierarchies. Once an indicator has been modelled, it must be bound to one or more dimensional hierarchies. The data analyst may reuse previously created hierarchies or define new ones, relying on the pivotal concepts `Dimension` and `Level` from the MDO.
Definition of personalisation concepts. The semantic representation of indicators is further enriched by associating them with their target domains (e.g., environment, health) through the `belongsTo` relationship. Personalisation concepts derived from the MDO are employed to affirm that the awareness of certain indicators impacts particular tasks, requiring end-users to base their decisions on these indicators (e.g., building monitoring, check air pollution). This is achieved by binding the indicator to a `UserCategory` and an `Activity` (or one of their sub-concepts) from the MDO. In particular, the `hasPracticableActivity` relationship binds a `UserCategory` to an `Activity`. Finally, the `involves` semantic relationship links an `Activity` to one or more `Indicators`.
Validation of the created indicator. To assist the data analyst in the modelling task, several constraints are checked through validation rules defined in the MDO: (i) a valid activity involves at least one indicator; (ii) a valid dimension hierarchy, being associated with an indicator, must gather at least one dimension level; (iii) a valid indicator belongs to at least one domain, is explorable according to at least one dimension hierarchy, possibly has a unit of measure and is involved in at least one activity. The interested reader can find the formulation of the validation rules in [8].

Example. In Figure 2, the `AirPollutionIndicator` is described as a sum of other indicators, has ppm as unit of measure and is linked with the `Environment` domain. `HouseholdCO2` is an example of composite indicator, specialised from `CO2Indicator` and computed starting from `CO2Heaters` indicator. All the indicators are associated with `SpatialDimension`, articulated over the `Apartment`, `Building`, `District` and `City` levels and connected each other by

rollUp relationship. Similarly, indicators are associated with the TimeDimension (not shown here). Lastly, indicators can be explored by both citizens and building administrators (modelled through corresponding concepts) while performing MonitoringPollutionLevels activity.

4. Identification of Personalised Exploration Contexts

Once the Exploration Graph $\mathcal{G}$ has been created, the Data Lake can be explored by relying on: (i) *Multi-Dimensional Descriptors*, apt to model multi-dimensional basic elements on which exploration is performed; (ii) *Exploration Contexts*, that identify portions of the Exploration Graph containing indicators compliant with users' activities; (iii) *contextual preferences*, to suggest to the user the most promising indicators to start the exploration from within an Exploration Context.

Multi-Dimensional Descriptors. Navigating across the Exploration Graph may be unpractical with a growing number of nodes and edges. Hence, to explore the indicators in $\mathcal{G}$, we foster a strategy grounded on the assumption that users inherently explore data according to a multi-dimensional organisation. In this respect, we defined proper *Multi-Dimensional Descriptors* over $\mathcal{G}$ (MDDs), to provide a compact representation of indicators and their dimensional levels. Figure 2 highlights two examples of MDDs for the CO2Heaters indicator.

Exploration contexts. Personalised exploration of MDDs is modelled through a set of *soft constraints* contained in users' profiles $p(u)$ for each u in the set of users $\mathcal{U}$. Soft constraints are modelled as *preferences*, organised according to *Exploration Contexts*, that represent the situations in which the user explores the MDDs, influenced by both his/her roles and goals. An Exploration Context ctx_u^i is used to delimit a portion $\mathcal{G}^i$ of the Exploration Graph $\mathcal{G}$, explorable by the user u. Available contexts are derived from $\mathcal{G}$ considering all the distinct pairs of UserCategory and Activity (sub-)concepts. At exploration time, a context ctx_u^i can be bound to one or more users' profiles. Users can manage their profile by selecting/changing the context of interest, choosing it from the ones compliant with their role(s).

Contextual preferences. The portion $\mathcal{G}^i$ delimited by an Exploration Context ctx_u^i may contain a high number of indicators, especially when considering a generic activity (such as "pollution monitoring"). To cope with this issue, *contextual preferences* help suggesting the user the indicators which best fit his/her demands. Contextual preferences can be either: (a) *short-term preferences*, expressed by the user at exploration time, representing imminent exploration needs; (b) *long-term preferences*, stored in user's profile, which are assumed to be static or change slowly over time. Contextual preferences are expressed on the set of MDDs derived from $\mathcal{G}$ through different *constructors* that rank indicators based on: (i) the distance that indicators have in the hierarchy induced by rdfs:subClassOf relationships (IND constructor); (ii) the distance that dimensional levels have in the hierarchy induced by rollUp relationships, focusing on a specific dimension (LEV constructor); (iii) the fact that an indicator belongs to a given domain (DOM constructor). Formalisation details of the constructors are available in [2]. The rationale behind these constructors is that the user can express his/her preferences on MDDs by relying on the relationships between MDDs and other concepts within his/her Exploration Contexts. Base constructors can be in turn combined using the *Pareto* composition ($\otimes$), composing two preferences with equal priority, and the *prioritization* ($\triangleright$) operator [9].

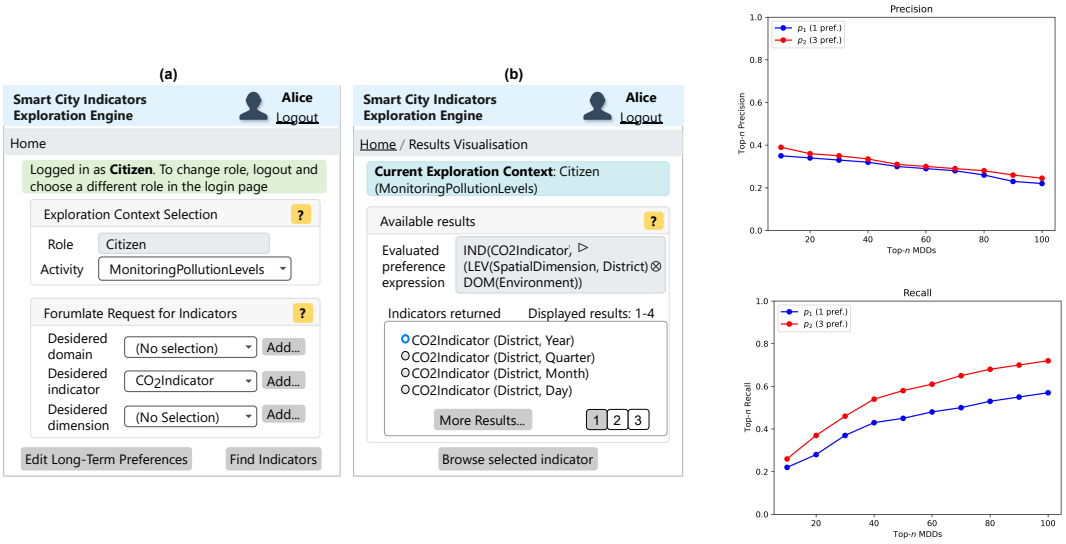

Figure 3: Representative web pages of the GUI for personalised exploration of indicators (left). Top-n precision and recall for assessing personalisation effectiveness (right).

5. Implementation and experimental evaluation

In this section, we present an excerpt of the implementation and the experimental evaluation conducted in the scope of the Brescia Smart Living project, wherein three different typologies of end-users have been identified as targets for the personalised exploration of indicators: (i) *citizens*, willing to explore aggregated data related to their neighbourhood (for example, average energy consumption, air quality, neighbourhood safety); (ii) *property managers*, administering one or more apartment buildings; (iii) *technical user plant managers*, responsible for heat distribution in buildings. In particular, we focus here on presenting the procedure for personalised exploration of indicators, achieved through a prototype web-based GUI (Figure 3). The usability of the GUI (in terms of facility in finding and exploring indicators) has been tested by a representative group of 10 users, belonging to the three aforementioned typologies, completing a standard System Usability Scale (SUS) questionnaire [10]. Averaged results from SUS questionnaires positioned the prototype in the 90-95 percentile range of the SUS score curve. Personalised indicators exploration has been performed on the top of a Data Lake infrastructure, relying on the Apache Hadoop File System (HDFS). The Data Lake internally adheres to a *zone-based* organisation [11], where each zone represents a stage of data processing. However, since the PERSEUS approach is meant to be employed on the top of the Data Lake infrastructure, it is agnostic with respect to the inner organisation strategy of the Data Lake, that depends on the non-functional requirements of application domain. For instance, when data quality issues have to be considered, a *Bronze-Silver-Gold* organisation may be more suitable with respect to a zone-based organisation.

GUI for personalised indicators exploration. Personalised indicators exploration is articulated over four main steps executed through a web-based GUI (Figure 3). To present the four

steps (S1-S4), we consider Alice, a citizen interested in monitoring air pollution levels to decide whether or not to practise outdoor activities, since pollution has effect on this kind of activities.
(S1) *Exploration context selection* – The exploration platform proposes Alice to select one of the Exploration Contexts available for her profile. Figure 3(a) shows the selection of the `AirPollutionMonitoring` activity.
(S2) *Short-term preferences formulation.* – In this step, Alice chooses the desired indicators, domains and dimensional levels, and the corresponding concepts are mapped by the platform to DOM, IND and LEV base preference constructors. For instance, in Figure 3(a), when Alice selects the `CO2Indicator`, the corresponding IND preference constructor will be automatically included in the request. The obtained constructors constitute the short-term preferences.
(S3) *Short-term and long-term preferences combination.* – Short-term preferences in the request are combined with long-term preferences in the profile $p(u)$ of the user, holding within the Exploration Context, thus leading to the compound preference $\overline{\mathbf{P}}$. Long-term preferences are automatically combined using the Pareto composition operator, since they all assume an equal importance for Alice. Short-term preferences are combined with long-term ones according to the prioritization operator ($\triangleright$), as they address an immediate need. After the request formulation has been finalised, Alice confirms her choices by clicking the "Find Indicators" button.
(S4) *Preference evaluation and indicators exploration.* – The compound preference $\overline{\mathbf{P}}$ from the previous step undergoes an evaluation process to identify the set of best (optimal) MDDs according to $\overline{\mathbf{P}}$. Such MDDs are proposed to the user, who can select any of them to explore indicator values. For example, Alice's preference evaluation result is displayed in the first page of the list in Figure 3(b). Finally, Alice selects one of the MDDs (by clicking on the "Browse selected indicator" button) and the multi-dimensional query apt to retrieve indicator values will be issued over the underlying Semantic Data Lake. As detailed in [2], the query is automatically generated from the selected MDD and contains: (i) a *projection* clause, with target indicator and analysis dimensions; (ii) the *aggregation function*; (iii) a *selection* clause (to restrict data access, according to the user's profile); (iv) the calculation *formula*. A set of *mappings* associated with the MDD (defined by the data analyst) allows to circumscribe a portion of the catalog over concepts that annotate the attributes involved in the query. The approach proposed in [12] is leveraged to create a query plan aimed at retrieving indicators values from Data Lake sources.

Experiments on personalisation effectiveness. To demonstrate the benefits (effectiveness) of personalisation in suggesting *relevant* indicators (MDDs) to the user, we used the two renowned metrics of Top-n precision and Top-n recall, as they are the most widely used metrics for the evaluation of retrieval systems. The effectiveness of a personalised search for indicators depends on users' profiles and, more specifically, on the preferences contained within. In this respect, two types of profiles, differing in the number of preferences, have been considered for ranking ≈ 3000 MDDs generated from 223 indicators: (i) p_1, containing only a single preference and (ii) p_2, a richer profile containing three preferences. Results for different values of Top-n MDDs are reported in the right side of Figure 3. In particular, the Top-n recall increases as long as the value of n increases and, for the same value of n, the profile with more preferences achieves a higher recall. Thus, a richer profile (i.e., with more personalisation elements) enables a more effective retrieval of relevant indicators (higher Top-n recall) and, as witnessed by the experiments conducted in [2], delivers a higher selectivity of MDDs.

6. Related Work

In this section, we will analyse an excerpt of the literature based on the requirements demanded by each phase of the PERSEUS approach. Regarding Semantic Data Lake modelling research, the focus of the latest years has been on the formalisation of models for supporting knowledge extraction from Data Lakes, building a semantic overlay with different techniques (e.g., by grouping similar attributes for easing querying data sources [13] or by building thematic views on the data sources, annotating their attributes [14]). Concerning the design of indicators, ontologies have been widely used due to their shared and machine-understandable conceptualisation. Recent efforts propose ontology-driven approaches to model KPIs, emphasising the importance of correlation between indicators values [15], possibly including personalisation concepts to drive the exploration of indicators (e.g., in [16], to explore sensors network data). Shifting towards data exploration issues, the usage of qualitative preferences yields higher expressiveness with respect to quantitative ones in assuring a (strict) partial order of search results. In [17], SPARQL qualitative preference queries are translated into query over relational databases systems, whereas in [18] preferences are formulated over aggregation levels of facts in a Data Warehouse ecosystem.

Novel contributions. PERSEUS aims at proposing a combined engineering of different techniques for addressing Semantic Data Lake exploration. With respect to [13, 14], PERSEUS fosters a preliminary lexical enrichment of data sources using both a lexical database and an abbreviation dictionary for building the semantic metadata catalog. In the second phase, the approach supports the definition of indicators also considering the activities performed by users while exploring data. These personalisation aspects in indicators modelling are only partially treated in [15, 16]. In the third phase, with respect to [17, 18], PERSEUS exploits users' preferences to rank indicators relying on their semantic definition, instead of actual values, which only at a later time are retrieved, thus saving cost and resources to query data sources.

7. Concluding remarks

In this paper, we presented PERSEUS, a computer-aided approach for data exploration on top of a Semantic Data Lake. The approach is structured over three phases: (i) the construction of a semantic metadata catalog on top of the Data Lake; (ii) the creation of an Exploration Graph, based on metadata catalog, containing the semantic representation of indicators and analysis dimensions; (iii) the enrichment of the definition of indicators with personalisation aspects (based on users' profiles and preferences) to identify Exploration Contexts, in turn delimiting portions of the Exploration Graph for a personalised and interactive exploration of indicators. Results of an experimental evaluation in the scope of the Brescia Smart Living project are presented with the aim of demonstrating the feasibility of the approach. Each phase of the PERSEUS approach paves the way to further investigation. For instance, regarding preference-based indicators exploration, we will enhance the preference model by considering the propagation of preferences across Exploration Contexts, as proposed by [19], thus establishing how preferences holding in a more generic Exploration Context are propagated to a more specific context.

References

[1] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, P. C. Arocena, Data Lake Management: Challenges and Opportunities, Proceedings of the VLDB Endowment 12 (2019) 1986–1989.

[2] D. Bianchini, V. De Antonellis, M. Garda, A semantics-enabled approach for personalised data lake exploration, Knowledge and Information Systems 66 (2024) 1469–1502.

[3] D. Bianchini et al., Data Management Challenges for Smart Living, in: Proc. of Cloud Infrastructures, Services, and IoT Systems for Smart Cities (IISSC 2017), 2017, pp. 131–137.

[4] STANDS4 Web Services: Abbreviations API, 2024. URL: https://www.abbreviations.com/abbr_api.php, Accessed on March 2024.

[5] G. A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (1995) 39–41.

[6] P.Y. Vandenbussche et al., Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web, Semantic Web 8 (2017) 437–452.

[7] Protégé: a free, open-source ontology editor and framework for building intelligent systems, 2024. URL: https://protege.stanford.edu/, Accessed on March 2024.

[8] M. Garda, A Semantics-Enabled Approach for Personalised Data Lake Exploration, Ph.D. thesis, University of Brescia - Italy, 2021.

[9] W. Kießling, Foundations of Preferences in Database Systems, in: Proceedings of the 28th International Conference on Very Large Databases (VLDB 2002), 2002, pp. 311–322.

[10] A. Bangor, P. T. Kortum, J. T. Miller, An Empirical Evaluation of the System Usability Scale, Intl. Journal of Human–Computer Interaction 24 (2008) 574–594.

[11] C. Giebler, et al., A zone reference model for enterprise-grade data lake management, in: 2020 IEEE 24th Int. Enterprise Distributed Object Computing Conf., 2020, pp. 57–66.

[12] H. B. Hamadou, E. Gallinucci, M. Golfarelli, Answering GPSJ Queries in a Polystore: A Dataspace-Based Approach, in: Proceedings of the International Conference on Conceptual Modeling (ER 2019), 2019, pp. 189–203.

[13] M. N. Mami, D. Graux, S. Scerri, H. Jabeen, S. Auer, J. Lehmann, Squerall: Virtual Ontology-Based Access to Heterogeneous and Large Data Sources, in: Proceedings of 18th International Semantic Web Conference (ISWC 2019), 2019, pp. 229–245.

[14] C. Diamantini, P. Lo Giudice, D. Potena, E. Storti, D. Ursino, An Approach to Extracting Topic-guided Views from the Sources of a Data Lake, Information Systems Frontiers 23 (2021) 243—-262.

[15] M. del Mar Roldán-García, J. García-Nieto, A. Maté, J. Trujillo, J. F. Aldana-Montes, Ontology-driven approach for KPI meta-modelling, selection and reasoning, International Journal of Information Management 58 (2019) 102018.

[16] C. Kuster, J.-L. Hippolyte, Y. Rezgui, The UDSA ontology: An ontology to support real time urban sustainability assessment, Advances in Engineering Software 140 (2020) 102731.

[17] M. Goncalves, D. Chaves-Fraga, O. Corcho, Handling qualitative preferences in sparql over virtual ontology-based data access, Semantic Web 13 (2022) 659–682.

[18] M. Golfarelli, S. Rizzi, P. Biondi, myOLAP: An Approach to Express and Evaluate OLAP Preferences, IEEE Transactions on Knowledge and Data Engineering 23 (2010) 1050–1064.

[19] P. Ciaccia, D. Martinenghi, R. Torlone, Foundations of Context-aware Preference Propagation, Journal of the ACM (JACM) 67 (2020) 1–43.