A Symbolic/Sub-Symbolic-Based Framework for Indicators Management*

Cristina Rossetti[1,2,*,†]

[1]*Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, ITALIA*
[2]*Università Politecnica delle Marche, Piazza Roma, 22, 60121 Ancona*

Abstract

Indicators are a key tool for measuring and monitoring performances in several dimensions, such as economic, social and environmental areas. The integration and analysis of indicators data, collected from heterogeneous and distributed sources, pose considerable challenges. In this paper, a unified indicators management framework is proposed, which extends and enhances an existing Semantic Data Lake architecture. An automatic integration mechanism resolves heterogeneities by mapping data sources to a global schema, while a query-driven discovery system facilitates the exploration of the repository to locate information of interest. The data analysis process is further enriched with the explanation of query results and the inclusion of new functionalities for indicator analysis. In this regard, Symbolic Regression is employed to discover relationships between indicators and to support robustness analysis in the construction of composite indicators.

Keywords

Indicators management, Symbolic Regression, Semantic Data Lake, Robustness Analysis

1. Introduction

Indicators are essential tools for monitoring performances and supporting decision-making processes across different areas, such as economic, social and environmental. The use of indicators by organisations, institutions or countries allows them to measure the achievement of strategic goals in order to make informed decisions [1]. For instance, the United Nations (UNs) developed a framework based on 231 unique indicators aimed to monitor global progress towards the Sustainable Development Goals (SDGs), gathering data from diverse institutions and organizations worldwide [1]. Effectively managing indicator data presents several challenges, starting with data collection. Given that indicators originate from heterogeneous and scattered sources and are often produced using different methodologies, data integration becomes a complex task. A related issue concerns the standardization of indicators. In fact, stakeholders involved in data collection and decision-making processes could interpret the meaning of certain indicators differently, leading to ambiguity. For this purpose, semantic technologies, such as ontologies, may serve as models to formally represent and define indicators and their components [1]. Another major challenge lies in the analysis of indicators, which is essential to monitor their dynamics, the identification of hidden patterns, the rank of entities, the discovery of analytical relationships between different indicators, and so on. In this context, the use of advanced data analytics tools may help making the most of indicators data. As such, data management systems as Data Warehouses (DWs) or Data Lakes (DLs) provide ad hoc solutions to overcome data integration, exploration, and analysis challenges. These systems can be further enhanced to support the entire data analytics pipeline—from data ingestion to sophisticated analysis techniques, also with the integration of semantic approaches. A particularly important aspect of indicator-based analysis is the construction of composite indicators (CIs). When dealing with multidimensional phenomena, relying on different multiple indicators can be impractical. Instead, these indicators can be aggregated into a single synthetic index, simplifying comparisons and enhancing decision-making. The building of CIs has gained widespread adoption,

SEBD 2025: 33rd Symposium On Advanced Database Systems, Ischia (Italy), 16 Jun - 19 Jun 2025

*Corresponding author.

cristina.rossetti@polito.it (C. Rossetti)

[1] https://unric.org/it/agenda-2030/

with the OECD establishing a ten-step framework [2].

The main objective of my research is to build a unified and global framework for the management and analysis of (composite) indicators. In this regard, the idea is to rely on a Semantic Data Lake (SDL) [3], a data management platform endowed with semantic technologies. The architecture is provided with i) an automatic mapping mechanism for integrating heterogenous data sources; ii) a query-driven discovery system for the retrieval of relevant information; iii) indicator management functionalities. My research has contributed to improve the mapping process with neural symbolic architectures and to extend the query answering system providing the user a better understanding of the results. Despite existing modules in the SDL for indicator management, there is still limited support for advanced indicator analysis and the construction of CIs. My current focus is on addressing these gaps by employing Symbolic Regression (SR), which offers two key benefits in this context. First, SR algorithms enable the discovering of complex (non-linear) relationships between individual indicators. Moreover, SR can enhance the CIs building, in particular the robustness analysis phase, in order to assess variables importance.

This paper is structured as follows: in Section 2 related works are discussed; the Section 3 presents the current state of the indicators management framework, briefly describing the SDL (3.1) and the most recent developments (3.2); my main contributions regarding indicators' analysis and robustness analysis of CIs are addressed in Section 4; in Section 5 are the conclusions and future directions.

2. Related Works

Indicator management requires the development of ad hoc frameworks and architectures to meet the challenges of data integration, discovery and analysis. A few works proposed implementations of indicators frameworks with data management architectures (e.g., [4], [5]). An important challenge related to the management of indicators is to provide a common representation of them, thus facilitating their standardization and usability. For this purpose, semantic technologies, such as ontologies and Knowledge Graph (KG) have demonstrated their usefulness. For example, [6] enhanced a DL architecture with the use of ontologies for policy support systems towards SDGs. The UNs [7] proposed an ontology-based organization system for the representation and analysis of SDGs indicators. [1] presented a comprehensive survey on the use of semantic approaches in indicators management frameworks. The authors claim that more efforts are needed to develop indicators management tools. Also, the results of the survey show a lack of data management systems-based frameworks. For example, [8] proposed an ontology-based DW system for business intelligence applications; [9] developed a smart city ontology to support indicators storing and exploration in a DL. From this perspective, the SDL proposed by [3] is a semantic-based DL with tools to support indicators uniform representation, their management and related data sources discovery and exploration. As mentioned before, the SDL has still room for improvement concerning indicators' analysis support. The idea is to enhance the data analytics pipeline with functionalities to discover new relationships among indicators. In this regard, SR algorithms can be employed to find mathematical expressions representing explicit relations among a set of indicators. SR differs from traditional regression techniques for its ability to simultaneously find the functional structure and its parameters' values, thus returning the analytic formula which best fits a given dataset. SR has been widely used in many contexts of application, from material sciences and physics to environmental and healthcare dimensions [10]. Some works focused on applications of SR in the analysis of performance indicators. [11] used SR to generate country-specific confidence indicators estimating the economic growth. [12] employed SR to model the relationships between large-scale temperature and several climate factors, such as greenhouse gases emissions. [13] aimed to identify correlations between outdoor and indoor particulate emissions by means of SR in order to discover behavioural trends in air pollutants. Hence, SR methods can support the retrieval of interesting formulas within indicators data and can be especially useful to identify complex non-linear relationships [14].

Concerning the construction of composite indices, to the best of our knowledge, no work has employed SR algorithms to support robustness analysis.

3. Current state of the framework

3.1. The Semantic Data Lake

The SDL [3] manages heterogeneous and distributed data sources by ingesting them without any pre-processing step, as typically in the case of DLs. Its main components are the Knowledge Layer and the Metadata Layer. The first aims to describe the semantics of information stored in the data sources. A set of ontologies provides the vocabularies to represent these information, and comprehends a KPIOnto ontology[2] specifically tailored to define performance indicators and their formulas. A KG built on the ontologies represents in a graph-based structure the concepts stored in data sources. Additionally, the Knowledge Layer includes a set of *logic programming rules*, which allows to reason over indicators' formulas. The main functionalities are: formula manipulation, including formula rewriting and equation solving; consistency check, allowing to evaluate whether a new formula is coherent, i.e. does not contradict any already stored formulas, or is equivalent to the stored ones [15]. The Metadata Layer manages metadata of data sources by storing them in a Metadata Graph (MG). A mapping mechanism involving the KG and the MG links data sources metadata to KG concepts, thus enabling data integration. Mappings are built automatically with the LSH Ensemble [16] algorithm and only works with categorical attributes. At present, the data sources we are working with are multidimensional cubes, which include indicators measured on a set of dimensions, i.e. perspectives of analysis, such as temporal or geographical. Dimensions are structured on hierarchies of levels, allowing data to be aggregated at different granularities (e.g., year, month or day for the temporal dimension). On the top of the SDL architecture, a query-driven discovery module allows the end user to query the SDL according to KG concepts. Thus, one can search for data sources containing indicators data measured on certain dimensional levels, e.g. aggregate GDP at country and year level. As already mentioned, the semantics of the indicators are stored in the KG, and their formulas are structured as *formula graphs*. This, together with the reasoning engine, supports a *query rewriting* mechanism which aims to generate additional queries for a given query based on the presence of formulas for specified indicators. In fact, when a formula has other indicators as arguments, it is necessary to search for other data sources containing these indicators. When query rewriting is activated, or there is simply more than one indicator in the query, it may happen that the information on the requested indicators is contained in different sources, thus requiring joins. Performing joins between different data sources can be time-consuming and computationally expensive, especially when multiple joins are required and the resulting data sources are many and huge. For this reason, the query answering system of the SDL is provided with a joinability index (JI) which estimates join cardinality, also when the join is multi-attribute, by means of the LSH Ensemble algorithm.

3.2. Architecture improvements

Data integration is performed via automatic mappings computed with LSH Ensemble applied over Minhash signatures of data sources attributes and KG concepts. Despite the efficiency of the algorithm, the process can be improved both in terms of effectiveness and the possibility of mapping non-categorical attributes. One of my on-going work focuses on employing a neuro-symbolic architecture, namely Logic Tensor Network (LTN) [17], to perform data integration. The training process is enhanced by the ingestion of a prior knowledge, which in our case is represented by the KG, and the prediction output is the degree of satisfaction of a certain attribute mapped to a KG concept.

The SDL has also been recently extended with the enrichment of the query-driven discovery module. Since the JI allows to rank the returned combinations based on their estimated cardinalities, the user has the possibility of choosing the best solutions in terms of the amount of information contained.

[2] https://kdmg.dii.univpm.it/kpionto/specification

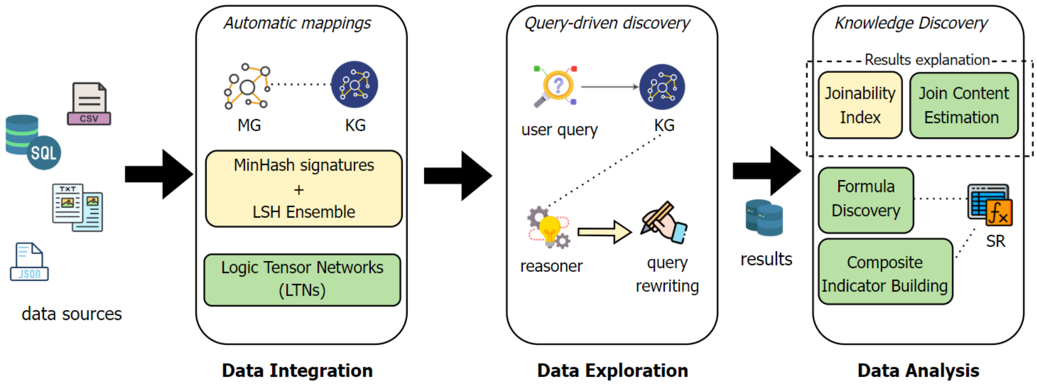

Figure 1: Data Analytics pipeline for indicators management and analysis within the SDL

In order to enhance the understanding of results, we proposed to employ data sources profiles (i.e., metadata on data distribution) in order to estimate the informative content of the joins. This, together with the JI-based classification, provides the user with a comprehensive overview of the discovered data sources, thus enhancing the exploratory power of the query answering system.

Figure 1 shows the data analytics pipeline supported by the SDL. Yellow boxes represent previous research, while green boxes are on-going works. Hence data sources of any type can be ingested and integrated with automatic mappings. Data exploration is performed via queries based on KG, which represents the global access schema. A reasoner engine allows to eventually find formulas for indicators in the query, thus rewriting multiple queries. Then the user can have a better understanding of results with the conjunction of JI and profile-based content estimation of joins. Additionally, data analysis can be enriched with the discovery of new analytical expressions for indicators (*formula discovery*) and the support of the building of CIs (with focus on robustness analysis), by means of SR, as discussed in Section 4.

4. Indicator analysis and building

4.1. Formula discovery

The query-driven discovery system of SDL allows to find the best data sources containing the indicators of interest. The reasoning engine enables the retrieval of formulas of the indicators in the query, thus discovering additional data sources. This allows to further support the discovery of interesting relations in data, enriching the analysis of indicators. Anyway, the availability of formulas for an indicator is not always straightforward. In this regard, we propose the use of SR in the context of formulas discovery, as it allows to retrieve explicit and interpretable analytical relationships between multiple indicators without forcing a pre-defined functional structure. Furthermore, unlike traditional regression techniques, SR allows the discovery of hidden (non-linear) relationships within the data without the need for prior knowledge. For example, suppose that there is a user who is interested in analyzing air pollution in Italian cities and wants to look for possible relationships with the traffic of vehicles. The SDL returns data sources containing values of the air quality index (AQI) averaged by city and day. The KG does not contain formulas related to AQI, but the user suspects some relations with a stored indicator, namely the average number of vehicles (ANV). With another query, the user retrieves a data source containing ANV data aggregated by Italian cities and day. The user presumes that as average traffic increases, so does the level of pollution in the air, but does not know the mathematical model relating the two indicators. Using ANV as the independent variable and AQI as the dependent variable (target), the SR algorithm returns the best expression representing the relations between ANV and AQI,

without the need of assuming any functional structure.
When the SR algorithm finds a new expression relating a given indicator to other indicators, this formula can be stored in the KG. The reasoning engine can support this process by evaluating the goodness of the expression in terms of coherence and consistency with existing formulas.

4.2. Robustness Analysis

CIs are synthetic indices obtained through the aggregation of multiple single indicators, with the aim of measuring some complex multidimensional phenomenon. Among the ten main steps of CIs building, we are currently focusing on robustness analysis. When a developer constructs a composite index, the choices made (e.g., weighting method) can heavily influence the soundness of the result. Unsuitable choices can lead to meaningless CIs values and thus to misleading conclusions by decision- and policy-makers [18]. Robustness analysis evaluates the quality of the resulting index by assessing how the inputs of the process (e.g., methodological choices) impact on the output. The main idea consists of employing SR to perform robustness analysis by computing the CI formula and identifying the contributions of the input variables, i.e. the single indicators. Target values are obtained through well-established formulas for CIs, such as the Adjusted Mazziotta Pareto Index (AMPI) [19], and the SR tries to retrieve the corresponding expressions. Here, robustness analysis can be performed in two main steps. The first concerns the analysis of formulas retrieved by the SR algorithm, in order to identify the selected variables and their functional form in the formula, as well as their polarity (whether they contribute positively or negatively to the output). The idea is that variables with simpler functional forms (e.g., linear contributions) and returned with the right polarity might be the most influential. The second step consists of using a machine learning model, more likely a neural network (NN), to evaluate whether the selected subset of indicators represents the most important variables for the CI. If the prediction error of the NN is smaller when the predictors are the selected indicators compared to the error obtained with the remaining ones, then it means that the chosen subset of indicators has a high probability of being the ones with the highest predictive power, thus the most influential.

5. Conclusions and future directions

The main goal of my work is to build a (composite) indicator framework with symbolic and sub-symbolic approaches, starting from an existing SDL. A KG built on ontologies makes it possible to formally define and represent indicators and their formulas and to facilitate the process of data integration and exploration. The automatic mappings between the KG and MG, designed for data integration and originally performed with the LSH Ensemble algorithm, can be improved with the use of neuro-symbolic architectures. The query-driven discovery system enables the discovery of data sources containing indicators of interest, also communicating with the reasoner engine to retrieve additional data if formulas are available. The results can be interpreted in terms of both cardinality and information content estimation, thus improving user understanding. In order to support the analysis of indicators, the SR can be employed for a twofold purpose. First, it enables formula discovery, which consists in retrieving new relationships among indicators stored in the KG. Then, it is used in robustness analysis to construct CIs, finding out which variables are most important. Future directions of my research relates different parts of the indicators management framework. It is intended to further test, and if possible improve, the application of LTNs for data integration. The other objective is to better integrate the reasoning engine with the application of SR, for example by using logic programming rules to validate, simplify and compare resulting expressions from SR algorithms. Concerning indicators' analysis, the proposal of SR-based formula discovery has to be implemented and tested, while the SR-based robustness analysis still has to be structured at the methodological and experimental level.

6. Acknowledgments

Cristina Rossetti has received funding from the MUR – DM 118/2023 as part of the project PNRR-NGEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. Diamantini, T. Khan, D. Potena, E. Storti, Semantic models of performance indicators: A systematic survey, ACM Computing Surveys (2025).

[2] J. R. Centre, Handbook on constructing composite indicators: Methodology and user guide, OECD publishing, 2008.

[3] C. Diamantini, D. Potena, E. Storti, Analytic processing in data lakes: A semantic query-driven discovery approach, Information Systems Frontiers (2024). doi:10.1007/s10796-024-10471-4.

[4] A. Maté, J. Trujillo, J. Mylopoulos, Specification and derivation of key performance indicators for business analytics: A semantic approach, Data & Knowledge Engineering 108 (2017) 30–49.

[5] A. Nasiri, W. Ahmed, R. Wrembel, E. Zimányi, Requirements engineering for data warehouses (re4dw): From strategic goals to multidimensional model, in: Advances in Conceptual Modeling: ER 2017 Workshops AHA, MoBiD, MREBA, OntoCom, and QMMQ, Valencia, Spain, November 6–9, 2017, Proceedings 36, Springer, 2017, pp. 133–143.

[6] A. Kulkarni, P. Bassin, N. S. Parasa, V. E. Venugopal, S. Srinivasa, C. Ramanathan, Ontology augmented data lake system for policy support, in: International Conference on Big Data Analytics, Springer, 2022, pp. 3–16.

[7] A. Joshi, L. G. Morales, S. Klarman, A. Stellato, A. Helton, S. Lovell, A. Haczek, A knowledge organization system for the united nations sustainable development goals, in: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18, Springer, 2021, pp. 548–564.

[8] G. Xie, Y. Yang, S. Liu, Z. Qiu, Y. Pan, X. Zhou, Eiaw: towards a business-friendly data warehouse using semantic web technologies, in: International Semantic Web Conference, Springer, 2007, pp. 857–870.

[9] A. Bagozi, D. Bianchini, V. De Antonellis, M. Garda, M. Melchiori, Personalised exploration graphs on semantic data lakes, in: On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings, Springer, 2019, pp. 22–39.

[10] D. Angelis, F. Sofos, T. E. Karakasidis, Artificial intelligence in physical sciences: Symbolic regression trends and perspectives, Archives of Computational Methods in Engineering 30 (2023) 3845–3865.

[11] O. Claveria, E. Monte, S. Torra, Economic forecasting with evolved confidence indicators, Economic Modelling 93 (2020) 576–585.

[12] K. Stanislawska, K. Krawiec, Z. W. Kundzewicz, Modeling global temperature changes with genetic programming, Computers & Mathematics with Applications 64 (2012) 3717–3728.

[13] R. R. Karri, B. Heibati, Y. Yusup, M. Rafatullah, M. Mohammadyan, J. Sahu, Modeling airborne indoor and outdoor particulate matter using genetic programming, Sustainable Cities and Society 43 (2018) 395–405.

[14] L. Kammerer, G. Kronberger, B. Burlacu, S. M. Winkler, M. Kommenda, M. Affenzeller, Symbolic regression by exhaustive search: Reducing the search space using syntactical constraints and efficient semantic structure deduplication, Genetic programming theory and practice XVII (2020) 79–99.

[15] C. Diamantini, D. Potena, E. Storti, Sempi: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators, Future Generation Computer Systems 54 (2016) 352–365.

[16] E. Zhu, F. Nargesian, K. Q. Pu, R. J. Miller, Lsh ensemble: Internet-scale domain search, arXiv preprint arXiv:1603.07410 (2016).

[17] S. Badreddine, A. d. Garcez, L. Serafini, M. Spranger, Logic tensor networks, Artificial Intelligence 303 (2022) 103649.

[18] S. Greco, A. Ishizaka, M. Tasiou, G. Torrisi, On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness, Social indicators research 141 (2019) 61–94.

[19] M. Mazziotta, A. Pareto, On a generalized non-compensatory composite index for measuring socio-economic phenomena, Social indicators research 127 (2016) 983–1003.

A. Appendix

Participating in this Doctoral Consortium represents a unique opportunity for me to enrich my research career. First of all, it would allow me to introduce myself to a national community of experts and colleagues, who could give me valuable feedback and advice for my field of study, the work done so far and my possible future directions. I also believe that confronting an audience of experts can improve my communication and presentation skills. Participating in this event also gives me the opportunity to expand my network of contacts and is an ideal environment to establish collaborations.

As the advisor I, Claudia Diamantini confirm that participating to the Doctoral Consortium would be an invaluable opportunity for Cristina to gather useful feedback from senior researchers, meet and discuss with peers about their research and seek possible collaborations, and breathing in the lively and inspiring atmosphere of the conference. Cristina is at the second year of her PhD, and the defense is scheduled for spring 2027.