# SemLytics: Semantic Enrichment for FAIR Analytics Architectures

**Thiago Zero Araujo[1], Cecília Nunes Sedenho[1] João Pedro de Carvalho Castro[2], Bruno Jesus dos Santos[1], Cristina Dutra Aguiar[1]**

[1]Department of Computer Science, University of São Paulo, Brazil

[2]Information Technology Board, Federal University of Minas Gerais, Brazil

thzeroa@usp.br, ceciliasedenho@usp.br, jpcarvalhocastro@ufmg.br, brunojsantos@usp.br, cdac@icmc.usp.br

***Abstract.*** *A key component of FAIR Analytics Architectures is the Semantic Enrichment Layer, which provides advantages such as enhancing interpretability and exploring complex relationships. However, developing this layer while addressing architectural specification, FAIR compliance, and semantic representation quality remains an unexplored challenge. In this paper, we propose SemLytics, a Semantic Enrichment Layer designed to address these issues. We present its logical level and demonstrate its implementation in a state-of-the-art FAIR analytics architecture. The results show that SemLytics improves FAIR compliance and semantic representation quality.*

## 1. Introduction

Analytics architectures support the ingestion, integration, storage, processing, and analysis of large volumes of heterogeneous data. They enable scalable and efficient data-driven decision-making and advanced analytical capabilities, such as business intelligence, machine learning, and statistical modeling. Data warehouses (DWs) played a central role in early data analytics architectures by providing support for batch-oriented analytical processing and query optimization [Kimball and Ross 2013]. Subsequently, driven by big data advances, data lakes (DLs) emerged to support heterogeneous data types (structured, semi-structured, and unstructured), real-time analysis, and flexible data science applications [Mathis 2017]. More recently, lakehouses have been proposed as unified architectures combining the advantages of DWs and DLs [Janssen et al. 2024].

The evolution of analytics architectures has also been driven by open science [Medeiros et al. 2020] and the FAIR principles [Wilkinson et al. 2016]. The FAIR Principles provide guidelines for making data Findable, Accessible, Interoperable, and Reusable, enabling digital assets to be shared and reused across organizational boundaries. As a result, developing FAIR Analytics Architectures requires not only scalable data infrastructures and advanced data management capabilities, but also semantic mechanisms capable of improving interoperability, contextualization, and data discoverability.

The Semantic Enrichment Layer is a key component of FAIR Analytics Architectures, representing semantic relationships through ontologies or knowledge graphs (KGs). It supports standards-based data retrieval, exploration of complex relationships, advanced analytics, semantic interpretability for machine learning, richer contextualization for large language models, and automated inference for knowledge discovery. However, designing a Semantic Enrichment Layer introduces several challenges. The layer should be generic enough to support integration with diverse FAIR analytics architectures while improving FAIR compliance and maintaining semantic representation quality.

In this paper, we propose SemLytics, a ***Sem****antic Enrichment* ***L****ayer for Ana****lytics*** designed to address these challenges. We present its logical design and demonstrate its implementation in a state-of-the-art FAIR analytics architecture, focusing on FAIR compliance and semantic representation quality.

The paper is organized as follows: Section 2 describes related work, Sections 3 and 4 presents SemLytics and its implementation, and Section 5 concludes the paper.

## 2. Related Work

Because a Semantic Enrichment Layer captures semantic associations through ontologies or knowledge graphs, this work is closely related to tabular-to-graph data mapping. This topic has received significant attention in the literature, particularly in the transformation of relational data into RDF representations (e.g., [Stvilia et al. 2024, Chen et al. 2025]).

In data analytics, [Bimonte et al. 2023] provide guidelines for modeling DW schemas (e.g., facts, dimensions, and hierarchies) using graph-based models. [Etcheverry and Vaisman 2012] propose QB4OLAP, an ontology-based approach aligned with the World Wide Web Consortium (W3C) Data Cube vocabulary. In our work, we do not introduce new mapping guidelines. Instead, motivated by QB4OLAP's alignment with W3C standards, we adopt it as the basis for SemLytics.

A related work to ours, also based on QB4OLAP, is SETL [Nath et al. 2017], a programmable semantic ETL framework that integrates relational and RDF data sources to construct semantic DWs. SETL provides powerful constructs for transforming data into multidimensional semantic representations. In contrast to our work, SETL is not designed as a generic layer, does not address FAIR compliance, and does not provide a formal evaluation of mapping quality.

Current FAIR-compliant architectures include GADDS [Vazquez et al. 2022], BigFAIR [Castro et al. 2025a], CloudFAIR [Castro et al. 2022], and FLAMI [Castro et al. 2025b]. GADDS focuses on metadata quality control through blockchain-based cryptographic storage and decentralized validation and, consequently, does not include a Semantic Enrichment Layer. BigFAIR, CloudFAIR, and FLAMI define a Knowledge Mapping Layer (KML) with features related to SemLytics. However, KML is specified at a high level, lacking detailed specification and implementation. Aspects such as FAIR compliance and mapping quality assessment also remain unexplored.

## 3. Proposed Semantic Enrichment Layer

Figure 1 depicts SemLytics, which comprises six components: TBox Ontology, R2RML Mapping Files, ABox Graph Storage, FAIR Compliance Evaluation, Graph Mapping Quality Evaluation, and Score Storage. TBox (Terminological Box), ABox (Assertional Box), and R2RML[1] (Relational Database to RDF Mapping Language) are established Semantic Web standards [Baader and Nutt 2003]. The TBox Ontology defines the ontology

[1] https://www.w3.org/TR/r2rml/

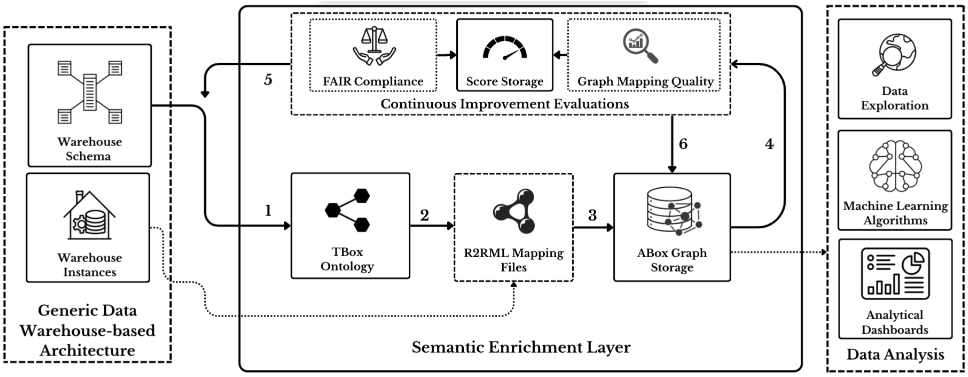


**Figure 1.** The logical design of the proposed Semantic Enrichment Layer, defining its components and data flows. The numbers represent the following stages: Stage 1: Ontology Construction; Stage 2: R2RML Mapping Construction; Stage 3: ABox and Knowledge Graph Generation; Stage 4: Quality Evaluation; Stage 5: Continuous Improvement; and Stage 6: Validated Knowledge Graph Return.

schema, including classes, properties, and relationships. The ABox Graph Storage stores instance-level data according to these definitions, such as individuals, their attributes, and their relationships. The R2RML Mapping Files component specifies mappings from relational databases to RDF using the W3C-standard language R2RML, ensuring a portable, auditable, and reproducible transformation process.

SemLytics is designed to be integrated with analytics architectures based on DWs or lakehouses, represented in Figure 1 as the Generic Data Warehouse-based Architecture. It ingests data from these architectures and transforms them into semantically enriched, FAIR-compliant data. The resulting data can then be consumed by later-stage analytical components, such as a Data Analysis layer built on top of a KG. By using the outputs generated by SemLytics, this layer exploits semantic enrichment to enable advanced analyses, including data exploration, machine learning techniques, and analytical dashboards.

Algorithm 1 formalizes SemLytics. Its inputs are: (i) a warehouse schema $S_{DW}$; (ii) its corresponding instances $I_{DW}$; (iii) a set of external vocabularies $V$; (iv) an optional set of dereferenceable Uniform Resource Identifiers (URIs) $URI$; (v) a framework for assessing graph mapping quality; (vi) a threshold $\tau_{map}$ representing the minimum acceptable graph mapping quality; (vii) a framework for assessing FAIR compliance; and (viii) a threshold $\tau_{FAIR}$ representing the minimum acceptable level of FAIR compliance. The algorithm outputs a knowledge graph $KG = (T, A)$, where $T$ denotes the TBox Ontology and $A$ represents the ABox Graph Storage, together with the graph mapping quality score $q_{map}$ and the FAIR compliance score $q_{FAIR}$ of $KG$.

Regarding the inputs, the set $V$ corresponds to terms defined in ontologies and employed to improve data interoperability. The set $URI$ contains dereferenceable URIs that enable resources to be uniquely identified and accessed on the Web, thereby supporting interoperability and Linked Data principles. When no URIs are provided, $URI = \emptyset$. Different frameworks may be employed to assess graph mapping quality and FAIR compliance, depending on the application requirements. The values of $\tau_{map}$ and $\tau_{FAIR}$ range from 0% (no compliance) to 100% (full compliance).

The algorithm starts in Stage 1 (line 1) by obtaining $S_{DW}$ from the target archi-

**Algorithm 1:** SemLytics Algorithm

**Input:** warehouse schema $S_{DW}$, instances $I_{DW}$, external vocabularies $V$, dereferenceable URIs $URI$, mapping quality framework $F_{map}$, mapping quality threshold $\tau_{map}$, FAIR framework $F_{FAIR}$, FAIR threshold $\tau_{FAIR}$

**Output:** knowledge graph $KG = (T, A)$, mapping quality $q_{map}$, FAIR compliance $q_{FAIR}$

`// Stage 1: Ontology Construction`
1 $T \leftarrow$ construct_TBox($S_{DW}$, QB4OLAP, $V$, $URI$)
`// Stage 2: R2RML Mapping Construction`
2 $M \leftarrow$ construct_R2RML_Mapping($S_{DW}$, $T$)
`// Stage 3: ABox and KG Generation`
3 $A \leftarrow$ execute_R2RML_processor($M$, $I_{DW}$)
4 $KG \leftarrow (T, A)$
`// Stage 4: Quality Evaluation`
5 $q_{map} \leftarrow$ evaluate_mapping_quality($KG$, $M$, $F_{map}$)
6 $q_{FAIR} \leftarrow$ evaluate_FAIR_compliance($KG$, $F_{FAIR}$)
`// Stage 5: Continuous Improvement`
7 **while** $q_{map} < \tau_{map}$ ***or*** $q_{FAIR} < \tau_{FAIR}$ **do**
8 $\quad G \leftarrow$ identify_gaps($T$, $M$, $F_{map}$, $F_{FAIR}$)
9 $\quad T \leftarrow$ refine_TBox($T$, $G$, $S_{DW}$, QB4OLAP, $V$, $URI$)
10 $\quad M \leftarrow$ update_mappings($M$, $G$, $S_{DW}$, $T$)
11 $\quad A \leftarrow$ execute_R2RML_processor($M$, $I_{DW}$)
12 $\quad KG \leftarrow (T, A)$
13 $\quad q_{map} \leftarrow$ evaluate_mapping_quality($KG$, $M$, $F_{map}$)
14 $\quad q_{FAIR} \leftarrow$ evaluate_FAIR_compliance($KG$, $F_{FAIR}$)
`// Stage 6: Validated Knowledge Graph Return`
15 **return** $KG$, $q_{map}$, $q_{FAIR}$

tecture to design a TBox ontology $T$, based on QB4OLAP to ensure independence from proprietary formats, and on external vocabularies $V$ to enrich data. When $URI \neq \emptyset$, dereferenceable URIs are incorporated to enhance interoperability and Web accessibility. This stage generates the TBox component. In Stage 2 (line 2), the algorithm employs R2RML to define the set of mappings $M$ representing the transformations between the features of the warehouse schema $S_{DW}$ and $T$. Stage 2 generates the R2RML Mapping Files component. In Stage 3 (line 3), the ABox $A$ is generated from $M$ and the warehouse instances $I_{DW}$. Furthermore, in line 4, the knowledge graph $KG$ is generated from the ontology $T$ and the ABox $A$, and stored in the ABox Graph Storage component.

In Stage 4, the algorithm performs a dual evaluation of $KG$, assessing mapping quality ($q_{map}$) using the mapping quality framework $F_{map}$ (line 5) and FAIR compliance ($q_{FAIR}$) using the FAIR compliance framework $F_{FAIR}$ (line 6). The evaluations are conducted by the Graph Mapping Quality and FAIR Compliance components, respectively, and the resulting scores $q_{map}$ and $q_{FAIR}$ are stored in the Score Storage component.

A continuous improvement process is initiated in Stage 5 by comparing $q_{map}$ and $q_{FAIR}$ with their respective thresholds $\tau_{map}$ and $\tau_{FAIR}$ (line 7). If any metric falls below its threshold, an iterative refinement begins by identifying the gaps $G$ affecting performance through the analysis of $T$ and $M$ with respect to $F_{map}$ and $F_{FAIR}$ (line 8). The TBox $T$ is refined using inputs from Stage 1, the previous $T$, and $G$ (line 9), while the mapping files $M$ are updated using inputs from Stage 2, the previous $M$, and $G$ (line 10). Stages 3 and 4 are then repeated: $A$ and $KG$ are regenerated (lines 11–12), and $q_{map}$ and $q_{FAIR}$ are re-evaluated (lines 13–14), with updated values stored in the Score Storage component, looping back to Stage 5. When both $q_{map}$ and $q_{FAIR}$ exceed their respective thresholds, the algorithm proceeds to Stage 6, returning the resulting $KG$, $q_{map}$, and $q_{FAIR}$.

| Vocabulary | Generated Instances |
|---|---|
| qb | 27,901 |
| dct | 27,842 |
| qb4o | 24,745 |
| time | 19,763 |
| skos | 10,525 |
| geo | 78 |
| sosa | 6 |
| prov | 1 |

**Table 1.** External vocabularies.

| Component | | Count |
|---|---|---|
| DW | Fact Records | 13,900 |
| | Dimension Records | 23,156 |
| TBox | Classes | 43 |
| | Object Properties | 16 |
| | Datatype Properties | 42 |
| | Reused Vocabularies | 8 |
| ABox | Observations | 13,900 |
| | Level Members | 23,156 |
| | External Instances | 110,861 |
| | Total RDF Triples | 306,460 |

**Table 2.** Characterization of $T$ and $A$.

## 4. Implementation and Demonstration

In this section, we present an implementation of SemLytics[2] based on a seismology instance of FLAMI, a state-of-the-art FAIR analytics architecture. We instantiate Algorithm 1 using the schema $S_{DW}$ and instances $I_{DW}$ derived from FLAMI CSV files. The schema, omitted here due to space limitations, contains one fact table with five numeric measures and nine dimension tables with 60 attributes in total [Castro et al. 2025b]. The external vocabularies $V$ are listed in Table 1, and $URI$ is initialized as $\emptyset$. For evaluation, we adopt [Tarasowa et al. 2015] as $F_{map}$ with $\tau_{map} = 50\%$, and FAIREST [D'Aquin et al. 2022] as $F_{FAIR}$ with $\tau_{FAIR} = 84.5\%$.

In Stage 1, the TBox $T$ is designed by combining FLAMI's schema $S_{DW}$ with the selected external vocabularies $V$, using QB4OLAP as the core vocabulary. Table 2 provides a quantitative summary of $T$, which includes 43 classes, 16 object properties, and 42 datatype properties, while reusing terms from 8 external vocabularies. All classes and properties are annotated with descriptive metadata, such as labels, domains, and ranges.

The mapping files $M$ are constructed in Stage 2 from $T$ and $S_{DW}$. The resulting mappings cover 53 out of the 60 attributes in $S_{DW}$, excluding 7 attributes from FLAMI's schema due to sensitive information. In addition, common analytical aggregation patterns typically computed at query time in DWs are explicitly represented in $M$. This enables the R2RML processor in the next stage to directly materialize pre-computed aggregated values as RDF triples.

In Stage 3, the ABox $A$ is generated by executing $M$ over the DW instances $I_{DW}$ using an existing R2RML processor[3], consolidating the knowledge graph $KG$. Table 2 quantitatively summarizes the generated triples. The process produced 13,900 observation instances and 23,156 dimension-level instances, matching the original DW records and indicating that no data loss occurred during the transformation. Additionally, 110,861 instances linked to external vocabularies were generated, detailed by vocabulary in Table 1. In total, 306,460 triples were stored in the Ontotext GraphDB[4] graph database.

With the resulting $KG$, the dual formal evaluation is performed. The Graph Mapping Quality evaluation in Stage 4 assesses $M$ and $KG$ using the framework $F_{map}$. Table 3 presents the results for the 13 metrics defined in [Tarasowa et al. 2015], yielding an average score of $q_{map} = 73.3\%$. For FAIR compliance ($q_{FAIR}$), we apply the frame-

[2] https://github.com/thzero0/SemLytics-Seismology-Implementation
[3] https://github.com/chrdebru/r2rml
[4] https://www.ontotext.com/products/graphdb/

| Metric | Score | Explanation |
|---|---|---|
| Data accessibility | 0.500 | System relies solely on ETL for the R2RDF transformations. |
| Standard compliance | 1.000 | Strictly adheres to R2RML standards rather than proprietary formats. |
| Coverage | 0.883 | 53 out of 60 relational columns were mapped, the remaining contain sensitive information. |
| Accuracy of data representation | 1.000 | Exact equivalent data, numeric averages and literal integrity preserved. |
| Incorporation of domain semantics | 0.000 | No values outside of the DW corresponding tables were added. |
| Simplicity | 1.000 | Graph abstractions were done by pre-aggregating data and including them in the KG, optimizing and reducing complexity in all proposed querys. |
| Data quality | 0.718 | Arithmetic mean of applicable Linked Data quality dimensions specified in [Tarasowa et al. 2015]. |
| Data integration | 0.000 | This metric measures the incorporation of external datasets into the mapping process. As no external data was incorporated in the mapping files, the resulting value is zero. |
| Reuse of existing ontologies | 0.626 | Average of 0.977 for class reuse (42 out of 43) and 0.276 for property reuse (16 out of 58). |
| Quality of reused vocabulary elements | 0.796 | Average of the documentation ($i_{doc}$) and Web popularity ($i_{pop}$) indices of the reused terms from the vocabularies listed in Table 1. |
| Accuracy of reused properties | 1.000 | Every reused property strictly obeys its original $rdfs:domain$ and $rdfs:range$ constraints and leaves no ambiguity. |
| Accuracy of reused classes | 1.000 | Reused classes used without ambiguity. |
| Quality of declared classes/properties | 1.000 | Comprehensive documentation (labels, domains, ranges) for all newly declared custom elements. |

**Table 3.** Evaluation of R2RDF mapping metrics.

work $F_{FAIR}$ to $KG$ and demonstrate that SemLytics enhances it. In [Castro et al. 2025b], FLAMI achieved an overall score of 73.4% and an interoperability score of 55.6%. After integrating SemLytics, interoperability is fully satisfied through RDF triple materialization and reuse of standard vocabularies, enabling semantic links between data and metadata. This increases FLAMI interoperability from 55.6% to 100%, while overall FAIR compliance rises from 73.4% to 84.5%. Since both metrics ($q_{map}$ and $q_{FAIR}$) satisfy their thresholds, Stage 5 is skipped and the process concludes in Stage 6.

## 5. Conclusion

We proposed SemLytics, a Semantic Enrichment Layer for integration into data warehouse and lakehouse-based analytics architectures. We introduced an algorithm to generate a knowledge graph grounded in multidimensional concepts and evaluated using FAIR and mapping quality metrics. An implementation of SemLytics based on FLAMI demonstrated the feasibility of the approach, increasing interoperability from 55.6% to 100% and overall FAIR compliance from 73.4% to 84.5%, while materializing 306,460 RDF triples without data loss. Future work includes exploring quality thresholds across different relational-to-RDF mapping approaches and FAIR compliance frameworks.

**Acknowledgments and Use of GenAI**. This work was supported by the project “IntegraSanca: Conectando Dados, Transformando São Carlos”, funded by the São Paulo Research Foundation (FAPESP, Grant No. 2025/07160-0), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), Finance Code 001. The first author was also supported by FAPESP (Grant No. 2025/09288-4). The authors declare that generative AI tools (ChatGPT) were used exclusively for linguistic revision of this manuscript. All technical content, scientific contributions, and conclusions presented are the sole responsibility of the authors.

## References

Baader, F. and Nutt, W. (2003). Basic description logics. In Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. F., editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, chapter 2, pages 43–95.

Bimonte, S., Gallinucci, E., Marcel, P., and Rizzi, S. (2023). Logical design of multi-model data warehouses. *Knowl. Inf. Syst.*, 65(3):1067–1103.

Castro, J. P. C., Romero, L. M., Carniel, A. C., and Aguiar, C. D. (2022). Open science in the cloud: The CloudFAIR architecture for FAIR-compliant repositories. In *Proc. ADBIS*, pages 56–66.

Castro, J. P. C., Romero, L. M. F., Chaves, A. C., and Aguiar, C. D. (2025a). The BigFAIR architecture: Enabling big data analytics in FAIR-compliant repositories. *J. Data Inf. Qual.*, 17(4).

Castro, J. P. C., Vasconcelos, G. F. X., Vargas-Solar, G., and Aguiar, C. D. (2025b). Building FAIR-compliant Lakehouses with FLAMI. In *Proc. CAiSE*, pages 113–129.

Chen, Z., Chamorro-Padial, J., et al. (2025). Tab2KGWiz: Interactive assistant for mapping tabular data to knowledge graphs. *SoftwareX*, 31:102221.

D'Aquin, M., Kirstein, F., Oliveira, D., Schimmler, S., and Urbanek, S. (2022). FAIREST: A framework for assessing research repositories. *Data Intell.*, 5:1–48.

Etcheverry, L. and Vaisman, A. (2012). QB4OLAP: A new vocabulary for OLAP cubes on the semantic web. In *CEUR Workshop Proceedings*, volume 905.

Janssen, N., Ilayperuma, T., et al. (2024). The evolution of data storage architectures: Examining the secure value of the data lakehouse. *J. Data Inf. Manag.*, 6:309–334.

Kimball, R. and Ross, M. (2013). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 3rd edition.

Mathis, C. (2017). Data lakes. *Datenbank-Spektrum*, 17(3):289–293.

Medeiros, C. B., Darboux, B. R., et al. (2020). *IAP input into the UNESCO Open Science Recommendation*. Available at https://www.interacademies.org/sites/default/files/2020-07/Open_Science_0.pdf.

Nath, R. P. D., Hose, K., Pedersen, T. B., and Romero, O. (2017). SETL: A programmable semantic extract-transform-load framework for semantic data warehouses. *Inf. Syst.*, 68:17–43.

Stvilia, B., Pang, Y., Lee, D. J., and Gunaydin, F. (2024). Data quality assurance practices in research data repositories: A systematic literature review. An annual review of information science and technology (arist) paper. *J. Assoc. Inf. Sci. Technol.*, 76:238–261.

Tarasowa, D., Lange, C., and Auer, S. (2015). Measuring the quality of relational-to-rdf mappings. In *Proc. KESW*, pages 210–224.

Vazquez, P., Hirayama-Shoji, K., Novik, S., Krauss, S., and Rayner, S. (2022). Globally accessible distributed data sharing (GADDS): A decentralized FAIR platform to facilitate data sharing in the life sciences. *Bioinformatics*, 38:3812–3817.

Wilkinson, M. D., Dumontier, M., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1):1–9.