Federated Architecture Based on Knowledge Graphs and Blockchain for Semantic Integration of Public Health Data

Nacles B. P. Gomes[1], **Laís do Nascimento Salvador**[1], **Paulo Caetano da Silva**[2]

[1]Programa de Pós-Graduação em Ciência da Computação (PGCOMP)
Universidade Federal da Bahia (UFBA) – Salvador – Brazil
[2]Universidade Salvador (UNIFACS) – Salvador – Brazil

nacles.gomes@ufba.br, laisns@ufba.br

and caetano.paulo@animaeducacao.com.br

Abstract. *The integration of heterogeneous data from the Brazilian Unified Health System (SUS) for epidemiological surveillance remains challenging. Legacy systems such as SIM (mortality) and SINASC (live births) operate independently, without shared identifiers. Existing approaches typically address these dimensions separately, lacking an integrated treatment of semantic mediation, decentralized custody, provenance auditing, and uncertain entity reconciliation over SUS legacy data. This paper presents the Distributed Federated Semantic Architecture (DFSA), combining knowledge graphs, permissioned blockchain, distributed encrypted storage, and fuzzy logic. The solution was validated with real-world data from the municipality of Camaçari, Bahia, Brazil, comprising 5.4 million RDF triples. It achieved a federated latency of less than 10 ms, with an estimated throughput of 1.6 million queries per hour, along with full auditability. The solution also received positive feedback from managers due to the automated generation of indicators.*

1. Introduction

Epidemiological surveillance constitutes one of the fundamental pillars of public health systems, being essential for disease monitoring, evidence-based policy formulation, and efficient resource allocation. In Brazil, the Unified Health System (SUS) provides coverage to over 210 million inhabitants, operating through dozens of information systems distributed across more than 5,500 municipalities. Among these systems, the most prominent are SIM (Mortality Information System), SINASC (Live Birth Information System), SINAN (Notifiable Diseases Information System), and e-SUS PEC (Electronic Citizen Health Record). Developed independently and at different historical periods, these systems employ distinct technologies and schemas — SIM, for instance, uses the Firebird[1] relational database for death certificates, while SINASC adopts a similar architecture with a different schema for live birth certificates. This structural heterogeneity imposes significant limitations on integrated health surveillance analyses: municipal managers frequently need to cross-reference mortality information with birth data to compute indicators such as the infant mortality rate; however, the absence of semantic integration between these systems renders this process manual, time-consuming, and prone to inconsistencies.

[1]Open-source relational DBMS; version 1.5 used by SUS legacy systems. `https://firebirdsql.org/`

The integration of heterogeneous data sources is a classical problem in computer science (Doan et al. 2012). Architectures based on mediators and wrappers (Wiederhold 1992), ETL pipelines (Halevy et al. 2006), and OBDA approaches (Calvanese et al. 2007; Calvanese et al. 2017) constitute well-established approaches, yet they exhibit limitations in federated scenarios involving distributed data and stringent governance requirements. More recently, knowledge graphs and blockchain technologies have been explored in a complementary fashion, although typically in isolation (Hogan et al. 2021; Hasselgren et al. 2020). In turn, systematic reviews indicate persistent gaps in the semantic interoperability of legacy health systems (Mello et al. 2022).

In the Brazilian context, initiatives such as the National Health Data Network (RNDS) seek to promote interoperability based on established standards, including HL7, FHIR, and SNOMED CT (Ministério da Saúde do Brasil 2020; da Silva Costa et al. 2025). However, these initiatives focus predominantly on the exchange of individual clinical data at the point of care, without addressing the semantic integration of data from health information systems — such as mortality and live birth records — required for the extraction of indicators from legacy systems such as SIM and SINASC.

Recent work in health-applied computing has advanced relevant subproblems, such as elastic human resource allocation in multi-hospital environments (Fischer et al. 2024), information extraction from unstructured texts into ontologies (Silva et al. 2022), the use of ontologies with inference for risk identification (Hoffmann et al. 2025), epidemiological data integration and visualization (Ribeiro et al. 2025), and AI-supported interoperability approaches (Ferreira et al. 2025). Despite these advances, none collectively describes a multi-domain and multi-organizational architecture that unifies semantic representation, decentralized governance, distributed storage, and formal uncertainty treatment.

Given this scenario, this paper proposes the Distributed Federated Semantic Architecture (DFSA), a solution that integrates knowledge graphs, permissioned blockchain, IPFS, and fuzzy logic to enable federated semantic integration of public health data. The remainder of the paper is organized as follows: Section 2 presents the theoretical background; Section 3 describes the research methodology; Section 4 details the proposed architecture; Section 5 presents the implementation; Section 6 discusses the experimental results; Section 7 presents the discussion; and Section 8 provides the conclusions.

2. Theoretical Background

This section presents the conceptual foundations underpinning the Distributed Federated Semantic Architecture (DFSA), the central contribution of this work. The exposition follows a problem-oriented logic, starting from the concrete challenge of data integration within the SUS context — exemplified by the SIM and SINASC systems —, and subsequently introducing the four pillars that complementarily address the identified dimensions: flexible semantic representation, decentralized governance and auditability, resilient distributed storage, and uncertainty treatment in entity reconciliation.

2.1. The Problem: Data Integration in SIM and SINASC

The challenge of integrating heterogeneous data in complex ecosystems such as SUS manifests across multiple dimensions: structural (distinct database schemas), syntactic (different formats and encodings), semantic (divergent meanings for similar terms), and organizational (distributed governance among municipalities, states, and the federal government). The SIM and SINASC systems embody these challenges in an emblematic manner: developed independently over legacy relational databases with periodic synchronization, both operate without unique identifiers that would facilitate interoperability. This characteristic transforms the correlation of records belonging to the same individual — for example, linking a death certificate (SIM) to a birth certificate (SINASC) for survival analyses — into a non-trivial entity reconciliation problem. Within the DFSA, the central question that emerges is: **what set of technologies can, in an integrated manner, address the multiple facets of this problem?**

2.2. Knowledge Graphs: Flexible Semantic Representation

The first pillar of the DFSA addresses the semantic and structural dimensions of the problem. Knowledge Graphs offer schema flexibility, native inference, and incremental integration (Hogan et al. 2021), characteristics particularly suited to the distributed SUS ecosystem. By representing data as RDF[2] triples (subject-predicate-object), it becomes possible to connect entities from different domains without imposing a rigid schema, thus facilitating the gradual evolution of the data model. The SPARQL[3] language, in turn, enables federated queries over multiple graphs, supporting cross-domain analyses without physical data centralization — a fundamental requirement for respecting the autonomy of federated entities.

In the context of SIM and SINASC, Knowledge Graphs allow the representation, within a unified model, of the heterogeneous structures of both systems, creating semantic bridges between concepts such as “death certificate” and “birth certificate”. Nevertheless, semantic representation alone does not resolve the challenges of governance and data provenance in a multi-institutional environment.

2.3. Permissioned Blockchain: Governance and Auditability

The second pillar addresses the organizational and legal dimensions. Permissioned blockchains are distributed immutable ledger networks in which only authorized participants can perform transactions and validate blocks. Unlike public blockchains (such as Bitcoin), permissioned ones are designed for corporate and governmental environments requiring access control, verified identity, and consensus-based governance among participants. In the healthcare context, this type of ledger provides verifiable and immutable provenance for metadata, being particularly relevant for meeting LGPD[4] requirements regarding the traceability of sensitive personal data processing (Androulaki et al. 2018; Hasselgren et al. 2020).

[2] W3C standard model for data representation on the Semantic Web, based on subject-predicate-object triples.

[3] W3C standard query language for RDF graphs, supporting federated queries over multiple endpoints.

[4] Lei Geral de Proteção de Dados (Law No. 13,709/2018): Brazil’s data protection law establishing principles of purpose limitation, data minimization, and traceability.

In the DFSA, the blockchain does not store raw health data (which reside in distributed repositories), but rather provenance metadata: who accessed what, when, for what purpose, and which transformations were applied. This creates an audit layer that operates in parallel with the semantic layer, ensuring that the flexibility of Knowledge Graphs does not compromise legal accountability. However, blockchain is not designed for large-volume data storage, which requires a third component.

2.4. IPFS: Resilient Distributed Storage

The third pillar addresses the infrastructural dimension. The **InterPlanetary File System (IPFS)** is a peer-to-peer distributed file system that employs content-based addressing, in contrast to the location-based addressing of the traditional web (Benet 2014). In IPFS, files are identified by their cryptographic content identifiers (CIDs), which ensures immutability and integrity. This characteristic makes it particularly suitable for storing health data in multi-institutional scenarios. In the DFSA, IPFS serves as the persistence layer for semantically represented data, creating an immutable link between the semantic representation and the raw content while keeping blockchain storage lean.

2.5. Fuzzy Logic: Entity Reconciliation under Uncertainty

The fourth and final pillar addresses the most critical operational challenge for the effective integration of SIM and SINASC: entity reconciliation under uncertainty. Fuzzy logic (Zadeh 1965) is an approximate reasoning approach that enables the treatment of uncertainty inherent in record matching across distinct sources (Christen 2012; Todorov et al. 2014). Unlike deterministic methods (which operate with binary matching thresholds), fuzzy logic expresses degrees of similarity within the interval $[0, 1]$, propagating confidence throughout the integration process. This is essential in the SIM–SINASC scenario, where records belonging to the same individual may exhibit variations in names (e.g., “Maria da Silva” vs. “M. Silva”) or addresses. In the DFSA, fuzzy reconciliation results feed into the Knowledge Graphs, which then represent not only entities but also probabilistic relationships among them.

2.6. The Complementarity of the Four Pillars

In summary, each pillar of the DFSA addresses a specific facet of the integration problem within SUS: **Knowledge Graphs** provide the semantic flexibility to represent heterogeneous data; **Permissioned Blockchain** ensures the decentralized governance and auditability required by the LGPD; **IPFS** provides resilient and immutable storage without central points of failure; and fuzzy logic enables precise entity reconciliation under conditions of uncertainty.

3. Research Methodology

This research adopted the **Design Science Research (DSR)** approach (Hevner et al. 2004; Peffers et al. 2007), organized into six stages. The first four stages were completed in this iteration: **(i) Problem Identification** through a Systematic Literature Review (543 studies, 23 selected), which revealed gaps in security/privacy (13%), formal uncertainty treatment (21.7%), and the absence of multi-domain validation; **(ii) Objective Definition**, translating the identified gaps into requirements for semantic integration, decentralized governance, auditability, and scalability; **(iii) Design and Development** of the DFSA

(Section 4) and the prototype (Section 5); and **(iv) Demonstration** with real-world SIM and SINASC data (Section 6). The stages **(v) Evaluation** and **(vi) Communication** are ongoing, including extension to SINAN/e-SUS Notifica and source code availability. During the research, LLMs supported auxiliary activities such as code generation, documentation, and language editing.

4. Distributed Federated Semantic Architecture (DFSA)

The DFSA is a conceptual architecture organized into four layers, as shown in Figure 1, separating the responsibilities of data ingestion, semantic transformation, distributed storage with governance, and semantic query orchestration. In contrast to classical live-source federation (Wiederhold 1992), the DFSA adopts a materialized semantic federation model: data are collected from legacy systems, transformed into RDF graphs, stored in distributed infrastructure, and queried through a semantic orchestration layer. This choice reflects the operational reality of SUS legacy systems — and other domains with heterogeneous legacy sources lacking query endpoints — and is grounded on four principles: **Federation without centralization**; **Consensus-based governance**; **Verifiable provenance**; and **Uncertainty tolerance**.

Conceptually, **Layer 1 (Ingestion)** defines distributed agents for collecting data from heterogeneous sources and producing standardized pre-graphs. **Layer 2 (Semantics)** transforms these pre-graphs into RDF graphs grounded in Source Ontologies (SO) and reconciles entities across distinct graphs. **Layer 3 (Storage and Governance)** combines content-addressable distributed storage (IPFS) with immutable metadata recording on blockchain. **Layer 4 (Federation)** executes distributed SPARQL queries over the graphs, consolidating results from multiple sources.

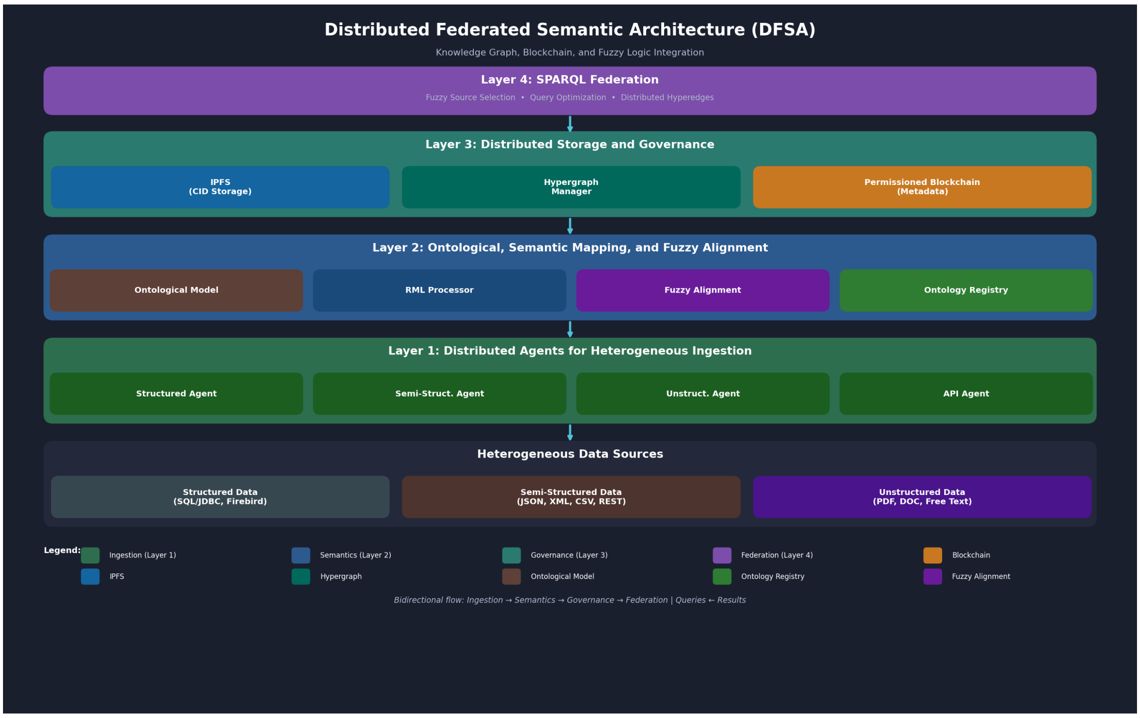

Figure 1. Overview of the DFSA with four layers and cross-cutting components of blockchain, IPFS, and fuzzy logic. Source: Authors

5. Implementation of the DFSA Solution

The current prototype implements and evaluates the DFSA primarily for structured SIM and SINASC data accessed through JDBC-based agents. Layers 1–3 are operational in this scope, covering ingestion, RDF transformation, encrypted IPFS storage, and blockchain metadata registration. Layer 4 is implemented as a custom semantic orchestration mechanism. Agent families for semi-structured, file-based, and unstructured data remain planned extensions.

5.1. Layer 1: Distributed Agents for Data Ingestion

Agents are autonomous and specialized software components, each responsible for accessing, collecting, and normalizing data from a specific source, implementing a standardized pipeline: collection → normalization → semantic mapping → publication. They operate asynchronously following reactive principles (Bainomugisha et al. 2013) and the Template Method pattern (Gamma et al. 1995), comprising five components: (i) source access mechanism; (ii) parsing and normalization; (iii) filtering and metadata enrichment; (iv) asynchronous parallelism manager; and (v) entity publisher with provenance.

The layer defines four agent families: (a) structured data, based on JDBC introspection; (b) semi-structured data, such as JSON, XML, CSV, and REST APIs; (c) file-based data; and (d) unstructured data, using ontology-guided NLP[5]. In the current validation, only structured JDBC-based agents were implemented and tested experimentally.

5.2. Layer 2: Ontological Layer and Mapping

This layer transforms pre-graphs into graphs grounded in Source Ontologies (SO) for SIM and SINASC, extended from prior work (Gomes et al. 2022; Franco et al. 2025). The SIM SO comprises 9 classes, 8 object properties, and 42 data properties; the SINASC SO comprises 10 classes, 9 object properties, and 33 data properties. Rather than a monolithic global ontology, the prototype uses source ontologies connected through alignment mappings acting as a lightweight mediated semantic layer.

The transformation process is supported by four components: (i) **Internal Ontological Model**; (ii) **Ontology Registry**; (iii) **Automated Mapping Engine**; and (iv) **Reconciliation Engine**. The mapping engine operates in two stages: lexico-semantic alignment between relational attributes and ontology properties (combining nominal similarity, synonym dictionary, content compatibility, and SQL–OWL type compatibility), followed by domain heuristics for confidence adjustment of candidate pairs.

The reconciliation engine identifies potentially equivalent entities across graphs through a composite score based on the classical weighted similarity framework (Christen 2012):

$$\mathrm{score}(e_i, e_j) = w_{name}\, s_{name} + w_{prop}\, s_{prop} + w_{ctx}\, s_{ctx} + w_{data}\, s_{data}, \quad w_k \in [0,1],\ \textstyle\sum_k w_k = 1.$$

[5]Natural Language Processing: enables extraction of structured information from free-text fields.

Each $s_k \in [0, 1]$ denotes a partial similarity (name, comparable attributes, contextual coherence, and key-field consistency, respectively), and each w_k is a domain-configurable weight. For the SIM–SINASC validation, weights were set to w_{name}=0.5, w_{prop}=0.2, w_{ctx}=0.2, w_{data}=0.1, reflecting the discriminative power of name-based matching in Brazilian civil registry records. The final decision employs fuzzy rules and configurable thresholds, with borderline cases forwarded for assisted validation.

5.3. Layer 3: Distributed Storage and Governance

This layer implements a data plane/control plane[6] separation (Tanenbaum and van Steen 2015): RDF graphs (1–10 MB/batch) are stored on IPFS (Benet 2014) with unique CIDs[7] ensuring integrity and deduplication; provenance metadata (CID, author, timestamp, access and retention policies) are recorded on the MultiChain blockchain (<1 KB/record). Graph encryption employs AES-256-GCM with a random initialization vector. MultiChain was chosen over platforms such as Hyperledger Fabric (Androulaki et al. 2018) due to its lower deployment overhead and computational cost for a controlled validation environment.

5.4. Layer 4: SPARQL Federation

The federation layer executes distributed queries over multiple RDF graphs, conceptually aligned with FedX[8] (Schwarte et al. 2011) and ANAPSID[9] (Acosta et al. 2011) but implemented as custom orchestration integrating IPFS, in-memory cache, and blockchain-based auditing. It supports two strategies: (1) per-source execution (parallel/sequential) and (2) in-memory join. The current implementation should be understood as a custom orchestration layer inspired by these engines, rather than a direct extension. A federated query involves source selection based on metadata and semantic mappings; retrieval of authorized graph identifiers from the blockchain; loading of RDF graphs from IPFS or cache; and execution of subqueries with in-memory joins, followed by audit logging.

6. Experimental Results

Experiments were conducted on a machine with Intel Core i9-ULTRA 185H (16 cores), 32 GB RAM, 1 TB NVMe SSD, and Windows 11. The dataset comprises 5,444,201 RDF triples (1,406,778 from SIM; 4,037,423 from SINASC). The protocol consisted of 100 executions (10 warm-up, 90 measurement) with a 95% confidence interval, following benchmark methodology from prior work (Bizer and Schultz 2009; Han et al. 2018).

Table 1 summarizes the obtained metrics. The DFSA exhibits bimodal behavior: simple queries (2.2 ms) and federated queries (4.5 ms) fall within the interactive range, with a 2.3 ms increment attributable to orchestration overhead — an acceptable cost for transparent querying across multiple sources. The throughput of 1.66M queries/hour reflects scalability for massive indicator generation. Inference latency (67.7 s) is suitable for analytical but not real-time scenarios. Benchmark queries covered three

[6]Architectural pattern separating effective data storage (data plane) from governance metadata and access policies (control plane).

[7]Content Identifier: cryptographic hash ensuring content integrity and deduplication.

[8]SPARQL federation engine using triple grouping and parallel execution across endpoints.

[9]Federated SPARQL engine supporting real-time adaptation to endpoint latency and availability.

classes: single-source lookup, two-source SIM–SINASC indicator generation, and offline inference for derived knowledge materialization.

Table 1. DFSA performance metrics.

Metric	Mean	σ	P_{95}
Simple Latency (ms)	2.2	3.8	13
Federated Latency (ms)	4.5	2.9	9
Inference Latency (ms)	67,673	7,129	80,962
Throughput (queries/hour)[a]	1,661,538	–	–
Overhead vs. Centralized (%)[b]	106.2	216.1	415.4

[a] $\hat{T}_{hour} = 3{,}600{,}000/\bar{t}_{ms}$ queries/hour (benchmark `LIMIT 1`, cache enabled, sequential).

[b] $(t_{f,i} - t_{1,i})/t_1 \times 100$ (single-source vs. 2 sources). Source: Authors.

6.1. Municipal Indicators

Municipal epidemiological indicators are quantitative measures for monitoring morbidity and mortality, supporting health surveillance and resource allocation. All data from the municipality of Camaçari used in the validation were anonymized and handled under institutional ethical approval, with no exposure of individually identifiable information.

Table 2 presents indicators validated by the local Health Department. The underlying integration mechanism is illustrated in Figure 2, which details: (a) a federated SPARQL query correlating SIM and SINASC to compute the death-to-live-birth ratio per 100 live births (2023); (b) a fragment of the query result; and (c) the resulting SIM dashboard, highlighting end-to-end pipeline transparency.

Table 2. Epidemiological indicators obtained via federated SPARQL queries.

Indicator	Value	Period
Total Deaths (SIM)	10,505	2019–2023
Total Live Births (SINASC)	22,350	2019–2023
Infant Mortality Rate (per thousand)	12.67	2023
Firearm Deaths (ICD X93.4)[a]	147	2023
Preterm Births (%)	11.0	2023

Data extracted via federated SPARQL queries over SIM and SINASC RDF graphs. Source: Authors.

7. Discussion

The validation of the DFSA with over 5.4 million RDF triples demonstrated technical feasibility across multiple dimensions. The federated query latency of 4.5 ms ($P_{95} = 9$ ms) falls below the 100 ms interactive threshold, indicating suitability for operational use by health managers. The estimated throughput of 1.66M queries/hour confirms scalability for massive indicator generation. Immutable auditability via blockchain with AES-256-GCM encryption ensures LGPD compliance.

Compared with recent work in health-applied computing (Ferreira et al. 2025; Ribeiro et al. 2025; Hoffmann et al. 2025; Silva et al. 2022; Fischer et al. 2024), the DFSA distinguishes itself by constituting an integrated, multi-domain, and multi-organizational architectural proposal. While those works address relevant subproblems

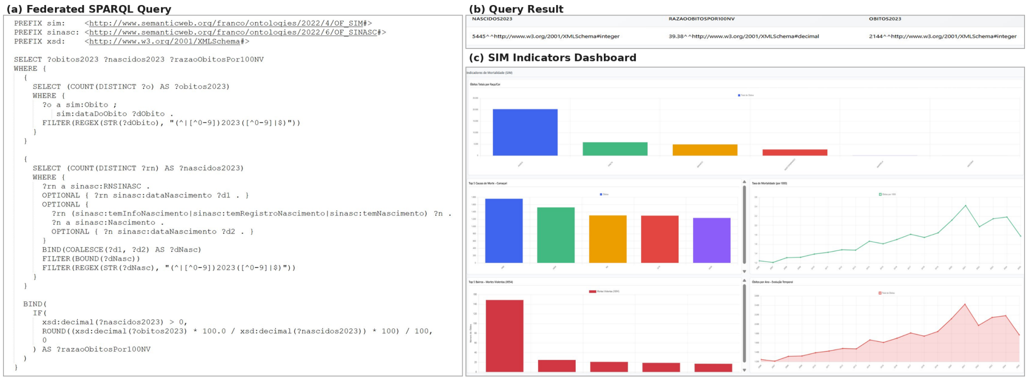

Figure 2. SIM × SINASC query for computing the death-to-live-birth ratio per 100 inhabitants — with SIM Dashboard — municipality of Camaçari, Bahia. Source: Authors

— such as domain-specific interoperability, clinical record extraction, or indicator visualization — none proposes a unified architecture combining knowledge graph-based semantic representation, decentralized blockchain governance, distributed storage, and formal uncertainty treatment within a single federated platform targeting SUS data integration.

7.1. Limitations

The implementation presents four limitations: **(i) Workload representativeness**: metrics were obtained under warm/hot conditions with low-cardinality queries, not fully capturing large-scale concurrent access; **(ii) Semantic inference cost**: the derived knowledge generation stage exhibited a mean latency of 67.7 s, suitable for analytical and offline enrichment processes but inadequate for real-time interactive use, as it does not lie on the critical path of federated queries; **(iii) Distributed storage security**: although the prototype implements AES-256-GCM encryption, it still lacks consolidated key management and rotation mechanisms for multi-institutional scenarios; **(iv) Operational and adoption challenges**: the current implementation does not evaluate the long-term costs of maintaining blockchain nodes across multiple municipalities, nor does it address the logistical complexity of ensuring data persistence in IPFS through formal replication agreements. These aspects are planned for future iterations.

Regarding the LGPD right to erasure, the architecture avoids storing raw health data on-chain; deletion is handled at the encrypted data and key-management layers, while the blockchain preserves only immutable audit metadata. To mitigate blockchain scalability issues, only compact governance metadata, CIDs, timestamps, and audit records are stored on-chain, with RDF graphs remaining off-chain in encrypted distributed storage.

8. Conclusion and Future Work

This paper presented the Distributed Federated Semantic Architecture (DFSA), combining knowledge graphs, permissioned blockchain, IPFS, and fuzzy logic to address the integration of heterogeneous SUS data. The hypothesis that the synergy of these technologies enables a truly federated, scalable, and auditable semantic integration platform was evaluated through a prototype applied to real-world data from Camaçari, Bahia.

The main contributions include: (i) transparent integration between systems via federated queries correlating mortality and birth data; (ii) complete auditability via permissioned blockchain with LGPD-compliant encryption; (iii) preservation of decentralized data custody; and (iv) extensibility to new sources through specific agents and ontologies. The positive evaluation by the Municipal Health Department corroborates practical applicability.

This work corresponds to the first DSR iteration, with partial evaluations validating technical feasibility. Upcoming iterations include: (i) extension to SINAN and e-SUS Notifica; (ii) cache optimization and incremental inference; and (iii) the evolution of cryptographic management for multi-institutional environments, encompassing asynchronous inference, the evaluation of partial materialization strategies, and distributed parallelization, with the aim of reducing computational costs without compromising semantic expressiveness.

References

[Acosta et al. 2011] Acosta, M., Vidal, M., Lampo, T., Castillo, J., and Ruckhaus, E. (2011). ANAPSID: an adaptive query processing engine for SPARQL endpoints. *The Semantic Web ISWC 2011*, 7031:18–34.

[Androulaki et al. 2018] Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., Caro, A. D., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., Cocco, S. W., and Yellick, J. (2018). Hyperledger fabric: A distributed operating system for permissioned blockchains. In *Proceedings of the Thirteenth EuroSys Conference*, EuroSys '18, pages 30:1–30:15. ACM.

[Bainomugisha et al. 2013] Bainomugisha, E., Carreton, A. L., van Cutsem, T., Mostinckx, S., and Meuter, W. D. (2013). A survey on reactive programming. *ACM Computing Surveys*, 45(4):52:1–52:34.

[Benet 2014] Benet, J. (2014). IPFS - content addressed, versioned, P2P file system. *CoRR*, abs/1407.3561.

[Bizer and Schultz 2009] Bizer, C. and Schultz, A. (2009). The berlin SPARQL benchmark. *International Journal on Semantic Web and Information Systems*, 5(2):1–24.

[Calvanese et al. 2017] Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodríguez-Muro, M., and Xiao, G. (2017). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 8(3):471–487.

[Calvanese et al. 2007] Calvanese, D., Giacomo, G. D., Lembo, D., Lenzerini, M., and Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, 39(3):385–429.

[Christen 2012] Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer.

[da Silva Costa et al. 2025] da Silva Costa, M. V., Camargos, M. C. S., Viana, S. M. N., and de Souza Mendes, U. V. (2025). Avanços e desafios da interoperabilidade no Sistema Único de Saúde. *Journal of Health Informatics*, 17(1):1112.

[Doan et al. 2012] Doan, A., Halevy, A. Y., and Ives, Z. G. (2012). *Principles of Data Integration*. Morgan Kaufmann, 1st edition.

[Ferreira et al. 2025] Ferreira, A., Redaelli, E., Maldaner, L., and Rigo, S. (2025). Inovação e transformação no ecossistema de saúde auditiva: Inteligência artificial e interoperabilidade. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 985–990, Porto Alegre, RS, Brasil. SBC.

[Fischer et al. 2024] Fischer, G., Righi, R., Costa, C., and Roehrs, A. (2024). Combinando elasticidade proativa e reativa para gestão multi-hospitalar de recursos humanos no contexto das cidades inteligentes. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 249–260, Porto Alegre, RS, Brasil. SBC.

[Franco et al. 2025] Franco, S. d. C., Gomes, N. B. P., and Salvador, L. d. N. (2025). Integração de dados de saúde pública para apoio à tomada de decisão: Uma abordagem baseada em grafo de conhecimento semântico. In *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025)*, volume 4177 of *CEUR Workshop Proceedings*, pages 63–76. CEUR-WS.org.

[Gamma et al. 1995] Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

[Gomes et al. 2022] Gomes, N. B. P., Franco, S. d. C., and Salvador, L. d. N. (2022). ONTOVID – uma abordagem para construção de grafos de conhecimento semântico com enfoque em notificações e óbitos relacionados ao novo coronavírus (COVID-19). In *Anais do XXII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 425–436, Porto Alegre, RS, Brasil. SBC.

[Halevy et al. 2006] Halevy, A., Rajaraman, A., and Ordille, J. (2006). Data integration: The teenage years. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB)*, pages 9–16. VLDB Endowment.

[Han et al. 2018] Han, R., John, L. K., and Zhan, J. (2018). Benchmarking big data systems: A review. *IEEE Transactions on Services Computing*, 11(3):580–597.

[Hasselgren et al. 2020] Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S. A., and Faxvaag, A. (2020). Blockchain in healthcare and health sciences—A scoping review. *International Journal of Medical Informatics*, 134:104040.

[Hevner et al. 2004] Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1):75–105.

[Hoffmann et al. 2025] Hoffmann, J., Renner, J., Real, R., and Barbosa, J. (2025). Ontodemer: Uma ontologia para identificação de riscos de doenças com base nos comportamentos de cadeirantes. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 164–175, Porto Alegre, RS, Brasil. SBC.

[Hogan et al. 2021] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4):71:1–71:37.

[Mello et al. 2022] Mello, B. H., Rigo, S. J., da Costa, C. A., da Rosa Righi, R., Aguiar, B. D., Bez, M. R., and Schunke, L. C. (2022). Semantic interoperability in health records standards: a systematic literature review. *Health and Technology*, 12:255–272.

[Ministério da Saúde do Brasil 2020] Ministério da Saúde do Brasil (2020). Portaria GM/MS nº 2.049, de 24 de setembro de 2020. institui a rede nacional de dados em saúde (rnds). Accessed: 2025-03-13.

[Peffers et al. 2007] Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77.

[Ribeiro et al. 2025] Ribeiro, F., Sinigaglia, M., Ávila, A., and Silva, I. (2025). Investigação de padrões epidemiológicos em oncologia pediátrica a partir de integração e visualização de dados. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 677–688, Porto Alegre, RS, Brasil. SBC.

[Schwarte et al. 2011] Schwarte, A., Haase, P., Hose, K., Schenkel, R., and Schmidt, M. (2011). FedX: optimization techniques for federated query processing on linked data. In *The Semantic Web - ISWC 2011*, volume 7031 of *Lecture Notes in Computer Science*, pages 601–616. Springer.

[Silva et al. 2022] Silva, D., Mello, B., Fröhlich, W., Rigo, S., Schwertner, M., and Costa, C. (2022). Avaliação de modelos para extração de dados não estruturados de um sistema ehr para atender a estrutura final de uma ontologia. In *Anais do XXII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 437–448, Porto Alegre, RS, Brasil. SBC.

[Tanenbaum and van Steen 2015] Tanenbaum, A. S. and van Steen, M. (2015). *Distributed Systems: Principles and Paradigms*. Pearson, 3rd edition.

[Todorov et al. 2014] Todorov, K., Hudelot, C., and Geibel, P. (2014). Fuzzy ontology alignment using background knowledge. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 134–145. AAAI Press.

[Wiederhold 1992] Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer*, 25(3):38–49.

[Zadeh 1965] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.