Journal of Informatics and Web Engineering

Vol. 5 No. 1 (February 2026) eISSN: 2821-370X

Optimised Data Integration using Transformer Model and Resource Description Framework

Jerome Aondongu Achir[1*], Muhammad Abdulkarim[2], Mohammed Abdullahi[3]

[1]Department of Computer Science, Joseph Sarwuan Tarka University Makurdi, Makurdi 970101, Benue, Nigeria

[2,3]Department of Computer Science, Ahmadu Bello University, Zaria 810211, Kaduna, Nigeria

**corresponding author: (achir.jerome@uam.edu.ng; ORCiD: 0000-0002-5756-2544)*

Abstract - Organizations have become highly reliant on a range of data sources that span structured, semi-structured, and unstructured data types. These repositories allow large-scale storage for faster ingestion and analytics but pose tremendous challenges of integration owing to schema and contextual differences. Traditional data integration methods, such as the ontology-based Resource Description Framework (RDF), are often inadequate when dealing with these challenges. They specifically struggle with the dynamic evolution of the schema of data sources, context-aware interpretation, and achieving interoperability across heterogeneous data sources. This paper presents an integrated system that augments resource description knowledge with token embeddings using the attention mechanism of the transformer model with relative positional encoding to overcome these weaknesses. Data from unstructured sources are used to create an embedding, whereas structured data are mapped into the RDF. The embeddings were then integrated into the RDF using *hasEmbedding.* Virtual transformations are employed to handle schema alignment and cosine similarity merges similar entities to provide a unified data view. Thus, the model explicitly integrates contextual knowledge within resource description knowledge triples, thereby improving the semantic representation. The proposed system uses a Simple Protocol and Resource Description Knowledge Query Language for the efficient querying of resource description knowledge, thus enhancing interoperability across domains. The proposed model produces a result that attains a good schema mapping accuracy of 97.82%, thus enabling more accurate and meaningful linking of heterogeneous datasets. Empirical trials involving use cases across human activity analysis and flood risk management prove the system's robustness, scalability, and effectiveness for knowledge discovery while allowing cross-domain integration of heterogeneous types of data within intricate scenarios. The results show that incorporating embedding into RDF reduces dependence on strict, pre-defined ontologies, simplifies schema on-demand alignment, and allows unified querying without the need to curate the integrated data into a traditional data warehouse.

Keywords—Data Integration, Heterogeneous Data Sources, Resource Description Framework, Embeddings, Schema Alignment, Ontology Knowledge Graphs, Transformer Models, Context-Aware

Received: 12 May 2025; Accepted: 4 August 2025; Published: 16 February 2026

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2026.5.1.8
© Universiti Telekom Sdn Bhd.
Published by MMU Press. URL: journals.mmupress.com/jiwe

1. INTRODUCTION

In the age of big data, organizations rely on insights from multiple data repositories [1]. These repositories serve as decision-making data across various sectors and domains [2], [3]. While modern data repositories can store massive amounts of heterogeneous data, they also bring integration challenges owing to format differences, semantic inconsistencies, entity disambiguation, and unstructured data complexity [4], [5].

Traditional data integration models, especially those using RDF and ontology-based models, attempt to address these issues, but struggle with schema evolution, context awareness, and adaptability, especially in dynamic environments [6], [7], [8]. Therefore, we propose research that aims to address these limitations by proposing a hybrid RDF-embedding integration model that combines RDF with semantic vector representations from transformer-based models. The model supports automated schema alignment through virtual feature transformations and enriches RDF triples with contextual embeddings to enable entity linking. These embeddings capture semantic relationships; therefore, integration across heterogeneous sources is more flexible and accurate.

Schmidt [9] opined that schema alignment based on embedding reduces manual curation costs. In addition, querying such integrated systems through a Simple Protocol and Resource Description Query Language (SPARQL) benefits from semantically enriched and unified representations [7], [10], resulting in more efficient and context-aware retrieval [11], [12]. Therefore, the proposed model uses embeddings to bridge the gap between the symbolic knowledge representation of RDF and the sub symbolic learning techniques of embeddings, thus improving data integration.

This paper is structured as follows: Section 2 captures related studies carried out by other researchers; Section 3 describes the methodology used to achieve the aim of the research; Section 4 captures and discusses the results obtained; Section 5 discusses the drawn conclusion; and Section 6 lists all consulted research materials.

2. LITERATURE REVIEW

Sakr [13] investigated the challenges of incorporating RDF data in large environments. Their evaluation of distributed RDF repositories also emphasized the importance of the efficiency and scalability of SPARQL queries. However, they noted that frequent updates remain a significant hurdle to scalability. Zhang [14] proposed the integration of structured and unstructured data into knowledge graphs using deep embeddings. They combined RDF with BiLSTM-based entity embeddings to enhance the schema alignment and query expansion. The main limitation is the need to retrain embeddings every time the schema changes [15]. Chen [16] enabled a semantic search over heterogeneous biomedical datasets by developing an ontology-enhanced RDF model incorporated with BERT embeddings. Although this approach is effective for concept linking, it requires domain-specific customization and may have limitations in generalizability. Wang [17] sought to enhance linked data integration using temporal graph embeddings. Their methodology involved tracking evolving knowledge graphs, although the system was complex and computationally demanding. Nguyen [18] performed an ontology alignment study using neural embeddings by combining Graph Neural Networks (GNNs) with word embeddings to recognize similar concepts semantically. This method is particularly sensitive to vocabulary variations and is dependent on access to high-quality pre-aligned datasets. Ali and Mehmood [19] integrated multi-format government datasets using RDF and BERT. They used BERT-based entity recognition models and mapped the recognized entities into RDF triples. A major limitation was the finding that pre-trained models performed poorly in domain-specific governmental contexts. Shao [20] analysed RDF entity disambiguation using contextual embeddings. They integrated BERT with SPARQL query enrichment to disambiguate ambiguous entities. However, they noted that the fetching of embeddings caused significant latency in query execution. Nundloll [21] targeted the semantic integration of flood risk data by applying the RDF/Ontology Web Language (OWL) to structured datasets. Although suitable for public sector planning, their model lacks adaptability to real-time unstructured data. Ali [22] presented a semantic alignment analysis of multilingual RDF datasets by using multilingual BERT to link cross-lingual definitions. The method showed admirable effectiveness in language adaptation but was largely dependent on the availability of high-quality translation corpora. Similarly, Song [23] proposed a hybrid model that combined symbolic RDF reasoning with neural attention mechanisms in one model. Such unification allows the simultaneous performance of neural flexibility and logical reasoning [24], but with a price in performance caused by the two-layered design. Despite these efforts, no research has satisfactorily addressed the problems of schema evolution, data ambiguity, and context awareness; hence, there is a need for a comprehensive approach.

Although our research work focuses on RDF-based integration and transformer embeddings, it aligns with the broader aim of applying artificial intelligence (AI) and data integration to complex, real-world domains, as also explored by Kalid et al. [25]. The integration of culinary data for recommendation systems [26] and personalized healthcare solutions using mobile AI [27] illustrate the practical applications of data integration and AI. Furthermore, weather-based arthritis tracking [28], [29] and migraine management have demonstrated the importance of linking heterogeneity.

3. RESEARCH METHODOLOGY

This involves the development of ontology using Protégé software, creation of embedding vectors, and integration of data from heterogeneous sources into the linked model using RDF with the additional property of *hasEmbedding* for unstructured data sources. This model assumes that semi-structured data sources contain non-hierarchical data, such as those produced from Internet of Things (IoT) devices and social media. This is accomplished by transforming unstructured data into an embedding by (1) performing part-of-speech tagging of tokens, (2) classifying entities into various classes, and (3) classifying the relationship between entities to capture semantic relationships between entities and convert unstructured data into a subject-predicate-object format. A *hasEmbedding* property was then created and added to the RDF.

Structured data is by design relational in nature, and in the required format of subject-predicate-object, it is ready for transformation into the RDF format.

The structure of the proposed RDF/Embedding system is captured in Figure 1.

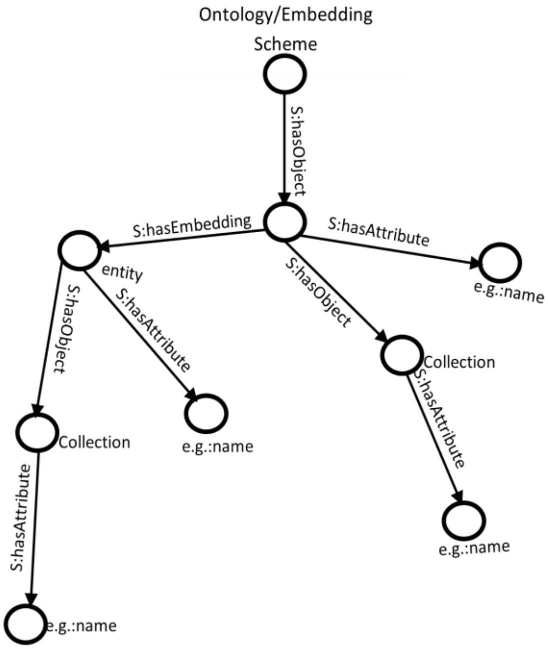


Figure 1. Architecture of The Proposed Model

The root node, *Ontology/Embedding Scheme*, serves as the top-level concept, defining the structural framework for entities and relationships. The primary relationships include *S:hasEmbedding*, which links the ontology to the entities in the unstructured dataset; *S:hasObject*, which defines relationships where an entity contains or is associated with another object (sub-entity); and *S:hasAttribute*, which connects an entity or object to its attributes, such as a name property. Entities and their attributes include names and other descriptive properties, and they can exist as collections, such as compound names, which also possess attributes. The hierarchical structure follows a tree-like format, in which entities contain objects, which in turn can have collections and further attributes. This hierarchy is particularly useful for semantic data representation in RDF models, allowing for the efficient organization and retrieval of information.

Because data are from different sources, situations may sometimes arise for the transformation of attributes for proper feature matching. These features from different sources are then virtually transformed to create entity links to avoid duplicate structures.

3.1 Transformation of Data

This method uses ontology to define related concepts to stitch entities together from heterogeneous data sources. In structured data, some fields may require virtual composition or decomposition of attributes before proper transformation occurs. The transformation is virtual, to protect the original nature of the data source.

3.2 Virtual Transformation of Source Attributes

To preserve the structure of the raw data stored in the sources, we adopt two virtual transformation operations, *Composition* μ and *Decomposition* γ, to work on the schema of the raw data rather than on the data itself [30], [31]. This allows the proposed research to virtually map an attribute in a source schema to a property in the global ontology. Because an entity may span more than a single field, for example, surname, firstname may be different fields in a structured data source, there is a need for composition and decomposition.

3.2.1 Composition μ

Given a set of source attributes A, $A= (a_1, a_2, a_3 ..., a_k)$, $A \subseteq S_i$, $S_i \in S$, $1 < k \leq n$, $n=|S_i|$, the composition operator $\mu A, a\mu$ comprises A into a single virtual attribute $a\mu$.

$A = (a_1,a_2,a_3,..,a_k)$ is a subset of attributes selected from a larger schema S_i, meaning that it comes from a specific schema or data source. S_i belongs to a broader set of schemas, S, which is mapped to G.

$1<k\leq n$, where k is the number of attributes in A, and $n=|S_i|$ is the total number of attributes in S_i. This implies that A consists of multiple attributes but does not cover all the attributes of S_i.

Operator $\boldsymbol{\mu A, a_\mu}$ combines the attributes in A into a single virtual attribute a_μ. This implies that the composed attribute does not exist physically but is created dynamically through a transformation process.

3.2.2 Decomposition

Given an attribute $a_y \in S, S \in G$, the decomposition operator $\gamma_{a\ y}$,A_γ decomposes the attributes a_γ into a set of virtual attributes A_γ, where $A_\gamma = a_{\gamma 1}, a_{\gamma 2}, .., a_{\gamma k}\}$, k >1.

$a_\gamma \in S_i$: a_γ is an attribute that belongs to a specific source S_i.

S ∈ G: The source S is part of a larger schema G.

The operator $\boldsymbol{\gamma a_y}, A_{\gamma,}$ acts on attribute $\boldsymbol{a_y}$ and decomposes it into several parts.

A_γ=$\boldsymbol{\{a_{y1}, a_{y2}, .., a_{yk}\}}$: The result of decomposition is a set of virtual attributes A_γ, where $k>1$. This indicates that a_y can be broken down into more than one attribute. This is captured in the composition and decomposition algorithm (Algorithm 1) presented in Appendix A.

3.3 System Design Pattern

The proposed system comprises the following components:

a) Data Preprocessing:

- POS - Transformer models classify tokens using relative positional encoding [32].

- NER Tagging - Transformer models classify entities using relative positional encoding.
- Relationship Extraction - Identifies entity-relationship-entity triples.

b) Embedding Generation: Token embeddings from transformers encode semantic context.

c) RDF Integration:

- Structured data is directly mapped to RDF: Structured data is directly mapped to RDF by preserving the existing table relationships and simply expressing them as triples since tables, columns, rows, and relationships already exist clearly. In the mapping process, each table becomes an RDF class, each row an RDF entity, and each column an RDF predicate. The data are now a graph of triples, easily queried with SPARQL, and linked across systems.
- Unstructured/semi-structured data were converted to RDF triples with embedded vectors, as shown in Figure 1.

3.4 Entity Linking

To achieve proper integration, similar entities extracted from multiple sources must be linked to determine if two entities refer to the same real-world object. We compared their embeddings using cosine similarity. The similarity score ranges between -1 and 1, where 1 indicates identical entities, 0 indicates no similarity, and -1 indicates completely dissimilar [33]. A predefined threshold is set to 0.7 and higher [34], [35], [36] to determine if entities are sufficiently like be linked, and their entity-relationship-entity schema is extracted into the RDF for integration.

3.5 Dataset

The dataset employed was obtained from the Groningen Meaning Bank (GMB), which is a text corpus gathered from news articles. Geospatial information from the United Kingdom Postcode, Heterogeneity Human Activity Recognition (HHAR) sensor-based activity recognition datasets, Food Hygiene Rating Scheme (FHRS) datasets, and the Flood Warning System (FWS) dataset were employed and integrated into an RDF store. The choice of these datasets was based on availability and was randomly but logically selected to spread across structured, semi-structured, and unstructured data sources.

3.6 Model Evaluation

Mapping Accuracy (MA) with a gold-standard mapping dataset remains the best metric for data integration, but owing to the non-availability of a gold-standard mapping dataset on the used datasets, a rule-based approach was employed for the system evaluation. It was evaluated by creating validation rules that cross-referenced the integrated data with the original datasets to verify accuracy.

Eleven (11) evaluation rules were developed to test the accuracy of the integration. The rules were classified into different evaluation metrics, as shown in Table 1.

The rules are applied and used to calculate the correctness of the sample for each rule, error on each rule, model accuracy, and model errors, as captured in Equations (1), (2), (3), and (4), respectively.

rule_accuracy = (Number of compliant samples / Total applicable samples) × 100

$$metric\ accuracy = \left(\frac{number\ of\ compliant\ samples}{total\ applicable\ samples}\right) \times 100 \tag{1}$$

$$model\ accuracy = \frac{\sum metric\ accuracy}{no\ of\ metric\ (9)} \tag{2}$$

$$metric\ error = \left(\frac{number\ of\ noncompliant\ samples}{total\ applicable\ samples}\right) \times 100 \tag{3}$$

$$model\ error = \frac{\sum metric\ error}{no\ of\ metrics\ (9)} \tag{4}$$

Table 1. Model Evaluation Metrics

Metric	Rules Covered	Evaluation Approach
Postcode Uniqueness	Rule i	Verify HHAR postcodes exist in reference datasets
Postcode City	Rule ii	Match city names against authoritative geographic databases
Business Postcode	Rules iii and iv	Check business existence in FHRS + postcode consistency (both must pass)
Embedding Consistency	Rules v and xi	Validate geopolitical entities in HHAR + cosine similarity ≥0.7 (both must pass)
Flood Warnings	Rule vi	Confirm flood-warning postcodes exist in hhar_postcode.csv
Sensor Data	Rule vii	Ensure sensor values (x/y/z) fall within [-20,20] and [0,20] ranges
Activity Location	Rule viii	Check activity coordinates within the attributed location bounds
Temporal Consistency	Rule ix	Validate that flood warning dates precede business inspection dates
FHRS Completeness	Rule x	Confirm no missing RatingValue or Postcode fields in business records

4. RESULTS AND DISCUSSIONS

4.1 Experiments

Several experiments were conducted to ascertain how well the heterogeneous data sources were integrated using both RDF and token embeddings to create an on-demand schema to demonstrate the effectiveness of the integration.

Considering a smart city environment scenario, we attempt to understand human activity across different locations to improve public safety, urban planning, and infrastructure management. The choice of scenario is based on the nature of the data available in the integrated RDF. To achieve this, we conducted three experiments by querying the RDF that cuts across different data sources. The idea is to ascertain how well the integrated data provides meaningful information in a unified view of demand.

4.1.1 Experiment I: Query Data for Movement Patterns

These data help analyse movement patterns within an area by detecting anomalous behaviours and potential traffic congestion. The query statement to perform the retrieval is captured in Algorithm 2 (Appendix A), and Table 2 captures the result from the query.

Table 2 shows the results obtained from the SPARQL query, providing insights into human activities across different locations, linking named entities (from the GMB), geospatial information (from the UK Postcode dataset), and human activity recognition data (from the HHAR dataset).

Table 2. Insights Into Human Activities Across Different Locations

Entity	Embedding	Postcode	Latitude	Longitude	User
London	City	SW1A 1AA	51.501	-0.141	a
Manchester	City	M1 1AE	53.479	-2.245	b
Google	Organization	SW1P 3AT	51.495	-0.135	c

Table 2. Insights Into Human Activities Across Different Locations (continued)

Activity	Device	Sensor	X	Y	Z
Walking	nexus4_1	Accelerometer	0.12	-0.08	9.81
Biking	s3mini_2	Gyroscope	1.23	-0.45	8.76
Sitting	samsungold_2	Accelerometer	0.01	0.00	0.02

Each row in the result represents a specific activity performed by a user in a certain location, captured via the sensors of a smart device (accelerometer or gyroscope).

For instance, the user "a" is walking in London at postcode SW1A 1AA, and the accelerometer sensor in a Nexus 4 smartphone records movement. The acceleration values of X, Y, and Z indicate a moderate motion. That is, X (0.12) and Y (-0.08) indicate minor sideways or forward movement, and Z (9.81) aligns with Earth's gravitational pull, indicating a normal walking motion.

The unstructured GMB Dataset provides information about named entities and their embeddings, contributing to the *Entity* and *Embedding* columns in the results. The Postcode Dataset is a structured dataset and is responsible for mapping real-world locations to entities, supplying the *Postcode*, *Latitude*, and *Longitude* columns. The structured Heterogeneous Human Activity Recognition (HHAR) dataset captures user activities along with sensor readings, contributing to the *User*, *Activity*, *Device*, *Sensor*, *X, Y*, and *Z* columns.

By integrating these datasets, the query effectively linked entity recognition, geographical information, and human activity data into a unified representation.

4.1.2 Experiment II: Human activities related to food establishments

The SPARQL query retrieves human activities, food establishments, hygiene ratings, and location details by integrating the HHAR dataset (S2), the UK Postcode dataset (S2), the FHRS dataset, and the Groningen Meaning Bank (GMB) dataset (S1). The goal is to analyse how human activities are distributed across different locations, particularly near food establishments, while also assessing their hygiene ratings, as captured in Algorithm 3 (Appendix A), while Table 3 captures the result of the query.

Table 3. Human Activities Related to Food Establishments

User	Activity	Food Establishment	Hygiene Rating	Postcode	Latitude
A	Sitting	Joe's Café	5	SW1A 1AA	51.501
B	Walking	Pizza Express	3	M1 1AE	53.479

In Table 3, the goal of the results is to analyse how human activities are distributed across different locations, particularly near food establishments, while also assessing their hygiene ratings. User *A* was recorded as sitting in an area associated with Joe's café, a food establishment located at postcode SW1A 1AA, which corresponds to central London.

Joe's Café has a hygiene rating of 5, which indicates excellent food hygiene standards. This means that the café maintains good cleanliness, food safety practices, and compliance with health regulations; people sitting or dining in this area are less likely to be exposed to foodborne illnesses, which is in a high-traffic area, making its good hygiene rating crucial for public health.

Because User A was sitting, it suggests that this area may have cafés, seating areas, and places where people take breaks. In addition, if many users exhibit similar sitting behaviour, city planners may consider expanding public seating in this location.

User B recorded walking in an area linked to Pizza Express, a food establishment located at postcode M1 1AE, which corresponds to Central Manchester.

The Pizza Express has a hygiene rating of 3, which indicates adequate but not excellent food hygiene standards.

This suggests that the establishment meets minimum legal requirements but may have some hygiene issues, regular inspections and improvements in cleanliness, food handling, or staff training might be needed, and the location is frequented by pedestrians, meaning food safety is important for public health.

The *User* and *Activity* columns were derived from the HHAR Dataset, which captures human activities such as sitting and walking. The *Food Establishment* and *Hygiene Rating* columns come from the Food Hygiene Ratings Dataset and provide information on various food establishments and their hygiene ratings. The *Postcode*, *Latitude*, and *Longitude* columns originate from the Postcode Dataset, which maps locations to real-world geographic coordinates. By integrating these datasets, the system effectively links human activities to food establishments, while incorporating spatial information for a comprehensive analysis.

4.1.3 Experiment III: Identifying Safe Food Establishments in Flood-Prone Areas

This query returns food establishments in flood-prone areas, as shown in Table 4.

As shown in Table 4, businesses in flood-prone areas may need to take proactive measures to ensure business continuity, considering the potential flooding risks. This could involve securing perishable food items, implementing protective measures to safeguard their premises, and ensuring that customers and employees remain safe during floods. Despite restaurants' excellent hygiene standards, floodwater contamination poses a significant threat to food safety, making it essential to have good protocols in place to prevent any potential health hazards.

Table 4. Food Establishments in Flood-Prone Areas

Business Name	Postcode	Rating	Inspection Date	Warning Type	Warning Date	Latitude	Longitude	Authority
McDonald's	SW1A 1AA	5	2024-02-15	Flood Watch	2024-03-01	51.501	-0.141	Westminster
Greggs	M1 1AE	4	2024-02-10	Flood Watch	2024-03-02	53.479	-2.245	Manchester

The FHRS dataset provides key details regarding food establishments, including the *business name, hygiene rating, inspection date*, and *authority* responsible for food safety oversight. This dataset ensures that information about food businesses and their compliance with hygiene standards is well-documented.

The *HHAR Postcode dataset* contributes essential geographical data by mapping *postcodes* to their corresponding *latitude* and *longitude*. This allows precise location tracking of food establishments and their surrounding environments.

The FWS dataset contains records of past flood warnings, specifying the *type of warning* issued and the *date* it was announced. This captures the effectiveness of the data-integration system fares. Furthermore, machine learning algorithms can be used to analyse these data for future prediction.

This system considers privacy concerns because it does not process or store personally identifiable information (PII). The activity and movement patterns used in the scenarios were derived purely from aggregated sensor readings and location postcodes, without linking to personal identities, such as names, phone numbers, or device IDs. This ensures that movement trends are analysed while preserving individual privacy, as anonymous patterns are used.

4.2 Evaluation

To evaluate how well our RDF integration algorithm correctly links entities across the datasets, a sample dataset of 5000 was randomly selected from the integration. A rule-based evaluation approach was employed, and the rules listed in Table 1 were applied to obtain the results for each metric, as described in Section 3. The results are presented in Table 5.

Table 5 presents each metrics accuracies and errors percentages for all the applied rules. The model accuracy and error are then taken as an average of the metrics, which is 97.82% and 2.18 respectively.

These metrics are visually captured in Figure 2.

Table 5. Data Integration Validation Report

Metric	Rules Covered	Score (%)	Error (%)
Postcode Uniqueness	Rule i	98.50	1.50
Postcode City	Rule ii	97.20	2.80
Business Postcode	Rules iii and iv	99.10	0.90
Embedding Consistency	Rules v and xi	98.00	2.00
Flood Warnings	Rule vi	98.43	1.57
Sensor Data	Rule vii	95.03	4.97
Activity Location	Rule viii	99.61	0.39
Temporal Consistency	Rule ix	96.52	3.49
FHRS Completeness	Rule x	98.01	1.99

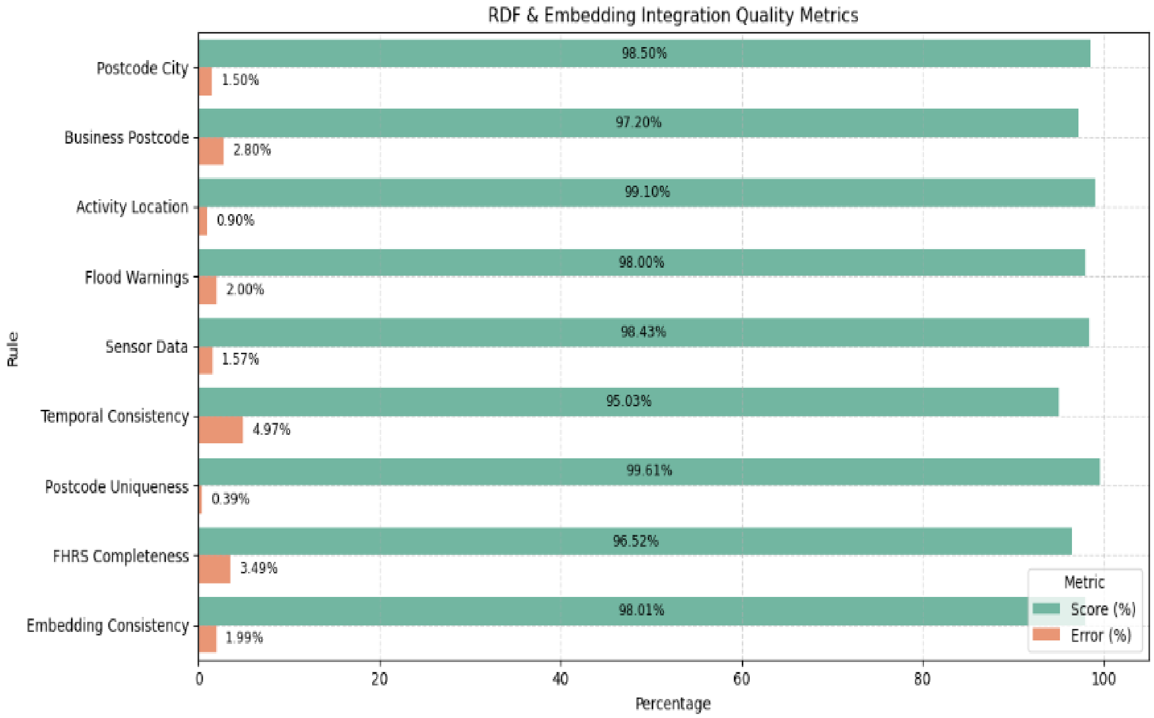


Figure 2. RDF and Embedding Integration Metrics

Figure 2 shows the RDF and embedding integration quality metrics across the nine (9) rules. The rule accuracy ranged from 95.03% (Temporal Consistency) to 99.61% (Postcode Uniqueness), while the error percentage ranges from 0.39 (Postcode) 4.97 (Temporal).

4.3 Comparison with Existing Models

The proposed model has improved the work of Nundloll [21], as embedding has increased the capability of the model in data generalization. Table 6 compares the results from the RDF integration with and without embeddings and demonstrates how embeddings influence data generalization and performance when the system is queried.

Table 6. Data Generalization With and Without Embeddings

Aspect	With Embeddings	Without Embeddings
Context Awareness	High (Embeddings link semantically similar entities)	Limited (Only direct relationships used)
Data Abstraction	Strong (Abstract concepts can be derived)	Weak (Only explicit matches retrieved)
Handling Synonyms	Yes (Embeddings capture synonyms and similar words)	No (Requires exact text matches)
Linking Implicit Data	Yes (Embeddings infer missing links)	No (Strict matching required)

With embeddings, querying for "fast-food chains" returns McDonald's, Burger King, and KFC because embeddings recognize semantic similarity. Without embedding, the query retrieves exact text matches and misses logically related entities.

Scenario I

Finding all fast-food chains returns "McDonald's", "KFC", "Burger King", while without Embeddings only returns "McDonald's" as explicitly labelled as "fast food chain". This demonstrates a higher generalization with embeddings by allowing smarter queries.

Scenario II

To retrieve average hygiene ratings for all dining places: Embeddings group "restaurants", "cafés", and "diners" together under "dining establishments", but without embeddings, only businesses explicitly labelled as "restaurant" will be retrieved.

Scenario III

The user finds relevant results, even when different words are used. For instance, "London", "Greater London", and "City of London" are semantically linked to postcode *SW1A 1AA*.

Scenario IV

Using the embedding, we can infer that "SW1A 1AA" relates to the entity "London", even if the word "London" is never explicitly stated in the postcode dataset.

This also implies that with embeddings, words with similar meanings are automatically linked; however, without embeddings, queries fail unless all possible variations are explicitly included.

The results from the SPARQL queries demonstrate how embeddings influence data integration and knowledge extraction. It has been shown that incorporating embeddings into RDF enhances data generalization by improving context awareness, entity linking, and abstraction in knowledge graph queries.

5. CONCLUSION

In this study, we propose an integrated data framework that combines transformer-based embeddings and RDF data based on a relational graph to enhance ontology-driven semantic solutions. It introduces a *hasEmbedding* property into RDF, thereby enabling dynamic schema alignment, enhanced context-aware entity resolution, and improved semantic interoperability. It proposes virtual schema transformations using composition/decomposition operators to reduce the reliance on rigid ontologies. The system merges semantic context with RDF triples and self-aligning benchmarks, entity reconciliation, and cross-domain interoperation on structured and unstructured data. An evaluation using 11 rule-based metrics demonstrated an accuracy of 97.82% and confirmed that the approach was robust and flexible. Real-life use cases conducted using smart city scenarios, such as activity recognition and flood risk analysis, demonstrate that the approach is practically usable. The enhanced framework is more powerful than prior RDF models because it enhances the generalizability and context-sensitive semantics of data integration in complex environments where RDF runs short.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their valuable comments.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Jerome Aondongu Achir: Conceptualization, Data Curation, Methodology, Validation, Writing;
Muhammad Abdulkarim: Project Administration, Supervision, Review and Editing;
Mohammed Abdullahi: Supervision, Review and Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] S. Azzabi, Z. Alfughi, and A. Ouda, "Data Lakes: A Survey of Concepts and Architectures," *Computers*, vol. 13, no. 7, pp. 183, 2024, doi:10.3390/computers13070183.

[2] R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data Lakes: A Survey of Functions and Systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12571–12590, 2023, doi: 10.1109/TKDE.2023.3270101.

[3] S. Ranatunga, R. S. Ødegård, K. Jetlund, and E. Onstein, "Use of Semantic Web Technologies to Enhance the Integration and Interoperability of Environmental Geospatial Data: A Framework Based on Ontology-Based Data Access," *ISPRS Int. J. Geo-Inf.*, vol. 14, no. 2, pp. 52, 2025, doi:10.3390/ijgi14020052.

[4] E. Gilman, F. Bugiotti, A. Khalid, H. Mehmood, P. Kostakos, L. Tuovinen, J. Ylipulli, X. Su, and D. Ferreira, "Addressing Data Challenges to Drive the Transformation of Smart Cities," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 5, Art. no. 88, pp. 1–65, 2024, doi: 10.1145/3663482.

[5] Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang, "A Novel Cascade Binary Tagging Framework for Relational Triple Extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2020, pp. 1476–1488, doi: 10.18653/v1/2020.acl-main.136.

[6] C. Lu, H. Zhou, and H. Su, "Persona and Contextual Semantic Embeddings for Entity Alignment," in *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, Aug. 2023, doi: 10.1109/ICIEA58696.2023.10241455.

[7] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word re presentations," in *Proc. NAACL-HLT*, 2018, pp. 2227–2237, doi:10.18653/v1/N18-1202

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi:10.18653/v1/N19-1423

[9] Y. Liu, E. Pena, A. Santos, E. Wu, and J. Freire, "Magneto: Combining Small and Large Language Models for Schema Matching," *arXiv:2412.08194v1* [cs.DB], Dec. 2024.

[10] N. Fanourakis, V. Efthymiou, D. Kotzinos, and V. Christophides, "Knowledge graph embedding methods for entity alignment: Experimental review," *Data Min Knowl Disc*, vol. 37, pp. 2070–2137, 2023, doi: 10.1007/s10618-023-00941-9.

[11] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, S. Ben Yahia, "Data quality in ETL process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676–687, Jan. 2019, doi: 10.1016/j.procs.2019.09.223.

[12] M. Farber and A. Rettinger, "A systematic approach to evaluating knowledge graph quality," *Semantic Web Journal*, vol. 11, no. 2, pp. 393–420, 2020, doi:10.3233/SW-190362.

[13] S. Sakr, M. Wylot, and P. Cudré-Mauroux, "RDF data management: A survey of systems," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–84, 2019, doi: 10.1145/3342190.

[14] S. Zhang, J. Li, and Z. Liu, "Embedding-enhanced RDF for hybrid knowledge graphs," *Information Sciences*, vol. 534, pp. 186–203, 2020, doi:10.1016/j.ins.2020.03.075.

[15] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, F. S. Alviano, M. N. Maleshkova, A. N. Ngomo, V. Tamma, and A. Zimmermann, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021, doi:10.1145/3447772.

[16] Q. Chen, A. Allot, and Z. Lu, "LitCovid: An open database for COVID-19 research," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1534–D1540, 2020, doi:10.1093/nar/gkaa807.

[17] T. Wang, Y. Zhang, and L. Guo, "Temporal knowledge graph embeddings for evolving RDF data," *Knowledge-Based Systems*, vol. 194, p. 105532, 2020, doi:10.1016/j.knosys.2019.105532.

[18] D. Q. Nguyen, T. Vu, and A. Nguyen, "Ontology matching with GNNs," *Semantic Web Journal*, vol. 12, no. 5, pp. 887–905, 2021, doi:10.3233/SW-210436.

[19] M. Ali, and R. Mehmood, "A semantic model for public administration data," *Government Information Quarterly*, vol. 38, no. 3, pp. 101592, 2021, doi:10.1016/j.giq.2021.101592.

[20] Y. Shao, B. Liu, and M. Zhang, "Context-aware RDF disambiguation," *Journal of Web Semantics*, vol. 67, pp. 100663, 2021, doi:10.1016/j.websem.2021.100663.

[21] V. Nundloll, A. Oloke, P. Smart, and N. Shadbolt, "Semantic integration of flood risk data using OWL and RDF," *Environmental Challenges*, vol. 4, pp. 100064, 2021, doi:10.1016/j.envc.2021.100064.

[22] W. Ali, M. Khan, A. Shams, A. Ullah, and M. M. Rathore, "Multilingual RDF integration using transformer embeddings," *Data & Knowledge Engineering*, vol. 145, pp. 102116, 2023, doi: 10.1016/j.datak.2022.102116.

[23] Y. Song, L. Zhang, Q. Wang, and S. Lin, "Hybrid reasoning over RDF with neural attention," *Neurocomputing*, vol. 553, pp. 126837, 2024, doi:10.1016/j.neucom.2023.126837.

[24] W. Li, R. Peng, and Z. Li, "Improving knowledge graph completion via increasing embedding interactions," *Applied Intelligence*, vol. 52, pp. 9289–9307, 2022, doi : 10.1007/s10489-021-02554-2.

[25] C. M. Chituru, S.-B. Ho, and I. Chai, "Diabetes risk prediction using Shapley additive explanations for feature engineering," *Journal of Informatics and Web Engineering*, vol. 4, no. 2, pp. 18–35, 2025, doi: 10.33093/jiwe.2025.4.2.2.

[26] M. T.T. Yong, S.-B. Ho, and C.-H. Tan, "Migraine generative artificial intelligence based on mobile personalized healthcare," *Journal of Informatics and Web Engineering*, vol. 4, no. 1, pp. 275–291, 2025, doi: 10.33093/jiwe.2025.4.1.20.

[27] J.L. Goh, S.-B. Ho, and C.-H. Tan, "Weather-based arthritis tracking: A mobile mechanism for preventive strategies," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 210–225, 2024, doi: 10.33093/jiwe.2024.3.1.14.

[28] J. C. Couto and D. D. Ruiz, "An overview about data integration in data lakes," *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, Madrid, Spain, 2022, pp. 1-7, doi: 10.23919/CISTI54924.2022.9820576.

[29] W.X. Ong, S.-B. Ho, and C.-H. Tan, “Enhancing migraine management system through weather forecasting for a better daily life,” *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 201–217, 2023, doi: 10.33093/jiwe.2023.2.2.15.

[30] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, “Big-data science in porous materials: Materials genomics and machine learning,” *Chem. Rev.*, vol. 120, no. 16, pp. 8066–8129, 2020, doi: 10.1021/acs.chemrev.0c00004.

[31] T. Abgrall, “Schema Decomposition via Transformation Patterns,” in *Proc. 42nd ACM SIGMOD-SIGACT-SIGAI Symp. Principles of Database Systems (PODS '23)*, Seattle, WA, USA, Jun. 2023, pp. 1–13.

[32] M, Abdulkarim, M. Abdullahi, and J.A. Achir, “Improving Part-of-Speech Tagging with Relative Positional Encoding in Transformer Models and Basic Rules”, *Indonesian Journal of Data and Science*, *6*(1), pp. 10-19, https://doi.org/10.56705/ijodas.v6i1.184

[33] Z. Lin, D. Yang, and X. Yin, “Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions,” *IEEE Access*, vol. 8, pp. 156663–156676, 2020, doi:10.1109/ACCESS.2020.3002977

[34] J. C. Couto, and D. D. Ruiz, “An overview about data integration in data lakes,” *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, Madrid, Spain, 2022, pp. 1-7, doi: 10.23919/CISTI54924.2022.9820576.

[35] Q. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading Wikipedia to answer open-domain questions,” in *Proc. ACL*, 2017, pp. 1870–1879, doi:10.18653/v1/P17-1178.

[36] J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva, “The Groningen Meaning Bank,” in *Handbook of Linguistic Annotation*, N. Ide and J. Pustejovsky, Eds. Dordrecht, Netherlands: Springer, 2017, pp. 463–496, doi: https://doi.org/10.1007/978-94-024-0881-2_20.

APPENDIX A

A mapping algorithm based on the composition and decomposition principles is presented in Algorithm 1.

Algorithm 1: entity composition and decomposition

```
Function ComputeMap(g, sij, t)

For each element p in collection g
    Set result as the outcome of calling Marcher with p, g, $s_{ij}$, and t
    If result(A) is not empty
        If result(A) has exactly one element and result(p) has exactly one element
            Add mapping from p to the single element in result(A) to $s_{ij}$
        Otherwise, if result(A) has more than one element and result(p) has exactly one element
            Set parent as the parent of the first element in result(A)
            Create newNode by joining parent and the label of p
            If the type of the first element in result(A) is embedding
                Add type embedding for newNode to $s_{ij}$
                Add relation from parent to newNode as embedding
            Otherwise
                Add type attribute for newNode to $s_{ij}$
                Add relation from parent to newNode as attribute
            Add mapping from p to newNode in $s_{ij}$
            For each element h in result(A)
                Remove relation between parent and result(A)[h]
                Add composition relation from result(A)[h] to newNode in $s_{ij}$

        Otherwise
            If the type of the first element in result(A) is not object
```

```
                Update its type to object in $s_{ij}$
            For each element h in result(p)
                Create newNode by combining the first element of result(A) with the label of result(p)[h]
                If the type of result(p)[h] is embedding
                    Add type embedding for newNode to $s_{ij}$
                    Link newNode as embedding of the first element in result(A)
                Otherwise
                    Add type attribute for newNode to $s_{ij}$
                    Link newNode as attribute of the first element in result(A)
                Add decomposition relation from newNode to the first element of result(A)
                Add mapping from result(p)[h] to newNode in $s_{ij}$
            Skip further handling of result(p)

After all iterations, add $S_i$ to S
```

Algorithm 2: Movement pattern query

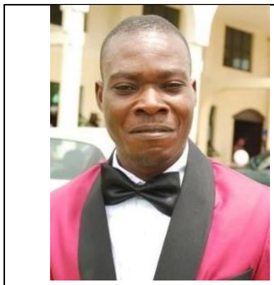
```
SELECT ?entity ?embedding ?postcode ?latitude ?longitude ?user ?activity ?device ?sensor ?x ?y ?z
WHERE {
   # Retrieve Named Entity Embeddings (GMB Dataset - S1)
   ?entity rdf:type :Entity .
   ?entity :hasEmbedding ?embedding .
   ?entity owl:sameAs ?postcode .  # Entity mapped to real-world postcodes

   # Retrieve Postcode Information (Postcode Dataset - S2)
   ?postcode rdf:type :Entity .
   ?postcode :latitude ?latitude .
   ?postcode :longitude ?longitude .

   # Retrieve User Activities (HHAR Dataset - S2)
   ?user rdf:type :Entity .
   ?user :performedActivity ?activity .
   ?user :occursAt ?postcode .  # Activity mapped to postcode

   # Retrieve Sensor Data
   ?user :hasSensorReading ?sensorReading .
   ?sensorReading :measuredBy ?sensor .
   ?sensorReading :x ?x .
   ?sensorReading :y ?y .
   ?sensorReading :z ?z .

   # Retrieve Device Used
   ?user :usedDevice ?device .
}
ORDER BY ?postcode ?user
```

Algorithm 3: Human activities related to food establishments

```
SELECT ?user ?activity ?foodEstablishment ?hygieneRating ?postcode ?latitude ?longitude
WHERE {
   ?user rdf:type :Entity .
   ?user :performedActivity ?activity .
```

```
    ?user :occursAt ?postcode .
    ?user :visited ?foodEstablishment .
    ?foodEstablishment rdf:type :FoodEstablishment .
    ?foodEstablishment :hasHygieneRating ?hygieneRating .
    ?foodEstablishment :hasPostcode ?postcode .
    ?postcode :isLocatedAt [ :latitude ?latitude ; :longitude ?longitude ] .
}
ORDER BY ?postcode
```

BIOGRAPHIES OF AUTHORS

	**Jerome Aondongu Achir** is a researcher at Joseph Sarwuan Tarka University Makurdi. He specializes in Data Mining and Natural Language Processing. His work explores advanced language models for natural language processing and computational linguistics. He continues to push boundaries in text understanding and model interpretability. He can be contacted at email: achir.jerome@uam.edu.ng.
	Abdulkarim Muhammad is a Professor in the Computer Science Department at Ahmadu Bello University, Zaria. He has served as Head of Department in the Department of Computer Science. His publications include collaborative work on wireless networks and 5G technologies and is known for integrating both teaching and research, contributing to the development of computer science education in Nigeria. He can be contacted at email: mmmhammmad@gmail.com.
	Abdullahi Mohammed is a Professor in the Department of Computer Science at Ahmadu Bello University, Zaria. His research spans artificial intelligence, optimization, and heuristic algorithms, and he has published widely in these areas. He contributes to mentoring students, shaping curricula, and leading research collaborations in Nigeria and abroad. He can be contacted at email: moham08@gmail.com.