Semantic Data Quality Assessment Model for the Management of Big Data

Onyeabor Grace Amina[1]
Department of Data Science
Federal University of Technology
Minna, Niger State, Nigeria
grace.onyeabor@futminna.edu.ng

*Ikwuazom Callistus Tochukwu[1]
Department of Information Technology
Federal University of Technology
Minna, Niger State, Nigeria
callistus.tochukwu@futminna.edu.ng

Rajesh Prasad[2]
Department of Computer Science,
African University of Science and Tech,
Abuja, Nigeria
rprasad@aust.edu.ng

Egere Ndudi Augustine[3]
Department of Computer Science,
Federal Polytechnic, Bali, Taraba
State, Nigeria.
austinendudi@yahoo.com

Enesi Femi Aminu[4]
Department of Computer Science
Federal University of Technology
Minna, Niger State, Nigeria
enesifa@futminna.edu.ng

Ezeaku Francisca Ngozi[5]
Department of Computer Science
University of Abuja, FCT Nigeria
francisicaezeaku@gmail.com

***Abstract*— Data quality (DQ) in Big Data (BD) settings is a pressing issue due to data diversity, volume and uncertainty. Poor data quality, especially related to inconsistency and incompleteness, results in inaccurate analytics and inadequate decision-making. This paper introduces a semantic-based data quality framework that combines Extract-Load-Transform (ELT) architecture with ontology-based data modeling to assess data quality. This framework uses Formal Concept Analysis (FCA) for detection of data quality problems, and K-means clustering for improving data quality. Furthermore, the system adopts the Total Data Quality Management (TDQM) approach, which involves definition, measurement, analysis, and improvement. The proposed model is validated through experimentation on a dataset of more than 10,000 books from the website Goodreads. The findings reveal that the model has 88.6% precision and 92.3% recall, which suggests a substantial improvement in terms of detecting and rectifying data quality problems (inconsistency and incompleteness) compared to the traditional methods.**

Keywords*— *Big data, data quality, semantic data transformation, ontology, clustering, formal concept analysis, ELT process, data integration.

I. INTRODUCTION

The explosive growth in online interactions, machine-to-machine communications, and sensor networks has dramatically increased the structured, semi-structured and unstructured data in Big Data (BD) environments [1]. Big data refers to large volumes of data with high velocity, variety, and complexity, and requires advanced technologies and techniques for management, capture, storage, analysis and sharing of the data [2]. Notably, huge volumes of data in the order of terabytes to petabytes are generated by the major companies like Google, Twitter, eBay, and Facebook. So, data management is important to manage the large data with an acceptable time frame [3]. The rapid growth of diverse data sources demands special attention to integrate the diversity of data types and data formats in big data environments [4], [5].

Data integration (DI) is the process of combining diverse, independent and distributed data sources into a coherent view [6]. Big data is increasing at an irregular pace due to the increasing pace of social media, digital media and sensors. To cope with this issue of information overburden, many industries and businesses try to manipulate valuable data for their business process optimisation using real big data analytical techniques. However, the analysis and extraction of valuable insights from big data is one of the most challenging aspects of big data [7]. So, semantic enrichment is a crucial element using semantic web technologies like ontologies.

The traditional Extract-Transform-Load (ETL) tools cannot be used to process big data because their processes only rely on ETL processes but not on creating semantic relations when integrating data from diverse sources [5], [8]. Previous studies on data transformation and integration have emphasised the creation of integrated resources, with a number of schema matching and mapping tools developed [9]. These processes have been supported by ontologies due to their strong semantics and formal knowledge representation capabilities [10]. But little research has focused on quality assessment in data transformation and integration, particularly for big data [11].

Big data is typically described as a huge amount of data: billions to trillions of data generated by millions of users and stored in a large number of sources around the world [12], [13]. Big data insights are revolutionising organizations, industries, science and society by allowing more effective government policies and commercial strategies, and enabling many applications in transportation, engineering, energy, biomedicine, genomics and environmental sensing [14]. But big data is largely unstructured and, as such, it is often incomplete and inconsistent; much of it is unavailable for use. To support decision-making, we need new technologies and tools to discover, transform, analyse and present this data [15].

This paper introduces a semantic approach for data quality assessment which overcomes these limitations by using ontology-based assessment, Formal Concept Analysis (FCA) and clustering algorithms to detect and resolve the data quality violations in big data environments. This paper is organized as follows: In Section II, related work is reviewed, in Section III, the approach is described, in Section IV, experiments and results are presented, in Section V, results and discussion are presented, and in Section VI, the conclusions are presented.

II. RELATED WORK

A. Data Quality Assessment in Big Data (BG)

Scalable Automation for big data quality (DQ) evaluation is explored recently. Al-Toq and Almaslukh [16] wrote a machine learning framework called DQMAF that introduces Decision Trees, Random Forest (RF), XGBoost, AdaBoost, and CatBoost for DQ profiling and classification in terms of correctness, consistency, completeness, and structural conformance. For Airbnb data, CatBoost obtained a weighted F1 of 0.97, while RF and XGBoost reached 1.00 precision/recall/F1; but DQMAF does not have the ability to detect and repair in an unsupervised manner. Martinez-Gil [17] put forward an automatic quality assessment for open data catalogs based on the following accuracy, completeness, consistency, scalability, timeliness, provenance, readability, and licensing properties, but it does not consider semantic inconsistencies and repair. Fadlallah et al. [18] introduced a Knowledge Based method with the aim of predicting missing links between the dataset context and applicable DQ rules, which was achieved through the use of Knowledge Graphs' embeddings, and the numerical attributes defined on their

edges acted as weights for quality measures, and an experiment was conducted on radiation sensor data, thus identifying the emerging need for the assessment of data quality with a view of context.

For classification, Sujon et al. [19] empirically tested and compared classification metrics (Precision, Recall, recall, F1-score, Accuracy and Matthews Correlation Coefficient (MCC)) on 2 imbalanced benchmark datasets using Logistic Regression, Decision Tree, RF, XGBoost and k-NN. They found that the F1-score was the most balanced and reliable metric, and MCC was used as an auxiliary diagnostic tool; the accuracy and precision values were sensitive to class imbalance and, for the case of XGBoost, yielded 95.4% accuracy, 95.1% recall, 95.2% F1, and 91.5% MCC. To complement this, Farhadpour et al. [20] introduced the guidance on multiclass metrics, showing that producer's accuracy (i.e. recall), user's accuracy (which implies precision), F1-score, and the weighted macro-averaged statistics are redundant, and therefore, one must pay attention to the metric selection in imbalanced tasks of data quality. Last but not least, Belgacem et al. [21] addressed the little-studied problem of categorical anomaly detection: in a form-filling recommender system, they predicted missing values for suspicious instances, using LAFF-AD. This is a direct solution to lack of reliable categorical DQ methods. Across the six datasets, LAFF-AD achieved at least 80.8% precision and 60–100% recall, significantly outperforming the OCSVM, LOF, iForest and EMAC-SCAN methods.

B. Semantic-Based Data Quality

Semantic approaches based on ontologies and knowledge graphs (KGs) have been widely applied to solve the problems of data quality (DQ) and heterogeneity in healthcare big data. During this cross-disciplinary survey of more than 10,000 papers Stănescu and Oprea [22] new trends in the field of dealing with heterogeneity using ontologies and Semantic Web technologies were identified, such as ontology engineering, knowledge graphs and linked data. The systematic review of 37 studies by Wu et al. [23] reveals that ontologies, controlled vocabularies, NLP, and the semantic web substantially benefit the conformance, completeness, usability, and FAIR principles of EHRs, but the scalability and computational requirements of large and evolving systems are still issues. Medical ontologies, as highlighted by Ambalavanan et al. [24], are used to bridge the gap between AI concepts and clinical, genomic and phenotypic data, fostering semantic interoperability and predictive analytics, but encountering challenges such as cost, scalability and standardization issues between diverse standards, like SNOMED CT and HL7 FHIR, which are constantly evolving. Likewise, Balakrishnan et al. [25] suggest an ontology-based approach that combines graph-based query optimization and knowledge graph to overcome semantic mismatches and to give more effective and accurate queries compared to relational systems. Chandur et al. [29] proposed a schema, entity resolution and ontology-mediated enrichment based semantic integration framework that enhances the queries' accuracy and minimises data duplication from healthcare, finance and smart cities. In applied reasoning, a detailed ontology, patient model and SWRL rules based IoT framework was developed by Zeshan et al. [26] with accuracy of 89.81% for context recognition and decision making. Mohammadhassanzadeh et al. [27] augmented OWL with plausible reasoning patterns (SeDan) over biomedical Knowledge Graphs and boosted the number of answered medical questions by 37% and 88% plausibility from the clinicians. Some frameworks like OBDM-based DQ tool by Cima et al. [28] are based on the concept of Consistency with the metadata; they compare integrity rules from the database with the knowledge from the ontology, and provide formal decidability analysis for standard OBDM specifications. The SHACL Dashboard Mäkelburg et al., [30] is an analytical tool that offers fine-grained constraint information and plots to help in the practical validation report analysis case, in which manual analysis is impractical for large knowledge graphs. Last, but not least, Dwyer et al. [31] introduced for the first time ontology-based process mining to cluster and abstract EHR events with clinical terms to help in the extraction of clinical pathways, but clinical terms are required to be expert-driven for cases where the clinical terms and the ontological structure are mismatched or ill-defined. Overall, these studies validate the strong contributions of semantic capabilities to healthcare DQ, interoperability, and analytic capability, but also highlight the ongoing need to scale, maintain, and stay up-to-date with the standards for these domains.

C. Machine Learning-Based Data Quality Improvement

Machine learning (ML) has been a pretty successful solution for data cleansing and enhancing data quality. Tayal [32] introduced an Autoencoder-BioBERT-Ontology AI-based ETL framework which significantly increases completeness (26%), consistency (17%) and diagnosis stability (10%), but requires a significant amount of labeled data and computation. The MapReduce based query rewriting technique by Benbernou et al. [33] is able to improve the completeness of queries on large RDF datasets, but leads to performance issues and recall problems. It has been seen that surveying large scale systems by traditional cleaning methods is not sufficient because of their heterogeneity and real time constraints, with Kamatala et al. [34] concluding the key challenges are the lack of scalability and benchmarks. To enhance adaptive processing, Syed et al. [35] developed self-learning RL framework with dynamic preprocessing and feedback, which improved the accuracy by 15-20%, decreased the preprocessing errors by 30% and achieved better real-world data quality improvements in finance (22-25%) and medicine (22-23%). Likewise, Tummuri [36] integrated GNNs, LSTMs and RL for real-time DQ detection and repair, which resulted in better anomaly detection and less loss in financial, healthcare, and sensor data. When it came to EHR abstraction, Kilpatrick et al. [37] found that AI techniques (95.7% recall, 93.8% precision, 94.7% F1) achieved a significant relative improvement over traditional techniques (52.2% F1) of 81.4%. Combined with human oversight, Oste et al. [38] obtained an initial precision/recall of approximately 88/79 (cardiology), 82/79 (oncology), and 86/82 (broad terms) for NLP with OMOP-CDM, which was improved to 96% for all domains following expert validation. Last but not least, Qin et al. [39] created a fully automatic and unsupervised pattern inferencer and anomaly detector without the need for any calibrated data. RIOLU achieved a success rate of 97.2% F1 to beat state-of-the-art baseline by up to 800.4% F1, 10% faster inference speed and 12.3% better precision than ChatGPT, and achieved 37.4% better precision than the state-of-the-art baseline with ease by guidance of users, with effective industrial use. Surprisingly, these studies show that ML has the potential to be a game-changer in automated DQ enhancement, but there are still significant gaps in the availability of scalable, low-resource, and real-time solutions that lack context. Together, these studies illustrate how ML

can be a game-changer in automated DQ enhancement while bringing volume context-free solutions and scalable, low-resource solutions to the table.

D. Data Transformation and ELT-Based Approaches

Sora-Cardenas et al. [40] created automatic thick blood smear analysis system that incorporates assessment of image quality, identification of leukocytes and malaria parasites. They evaluated the performance of F1 on F1000 clinical images with F1 scores at 95% for quality, 88.92% for leukocytes and 82.10% for parasites, which highlights F1's usefulness when dealing with imbalanced datasets. Data integration (DI) architectures have evolved from ETL to ELT and hybrid, and benefit from cloud scalability. To significantly reduce the time that it takes Spark to process data, Farhan et al. [41] presented a new approach to cleaning that was separated from the transformation, called Extract-Clean-Load-Transform (ECLT). But Dhaouadi et al. [42] pointed out that big data has given rise to new data warehouse formalisms, there is no generic design model for collecting, storing, processing and analysing big data. In the health care field, Ong et al. [43] could develop D-ETL, a Dynamic-ETL tool, using a rule engine to auto-generate SQL to transform multi-source clinical data to the common model used by OMOP, which allows for multi-site research to proceed transparently. Peng et al. [44] found complementary to our findings, ontology-based and rule-based approaches are the primary approaches used in medical data harmonization, but they noted that there are still issues such as tool selection to be addressed. To explore future automation, Jin et al. [45] presented ELT-Bench to benchmark AI agents throughout the pipeline, from planning to extraction, tools use and query formulation. Though the results indicated that the best agent managed to perform the correct design in 3.9% of cases, there are still significant challenges to be overcome to completely automate ELT processes.

E. Research Gap

A number of recent studies have made progress in the area of data quality assessment, but there are many remaining challenges. Although the work of [16] proposed a framework to profile and classify data quality using ML, it did not include automated mechanisms for repairing the data and was not rule based to detect specific types of violations. Semantic approaches, such as those in the works of [22], [23], [24] highlights the utility of ontologies and knowledge graphs in data integration and quality, yet their applications are hindered by scalability, computational costs, and a lack of adaptive learning ability. Large labeled datasets are needed for MIR based approaches, which are also resource demanding, as in [32]. Recently, Kilpatrick et al. [37] reported F1 scores of 94.7% for EHR data quality and Qin et al. [39] obtained an F1 score of 97.2% in an unsupervised pattern inference task, but their solutions are domain specific and are not a general solution to semantic-reasoning and automated-repair. Moreover, Sujon et al. [19] and Farhadpour et al. [20] highlighted the importance of a well-balanced evaluation framework that is rarely considered for imbalanced settings. No existing methods provide a generalizable approach to detection and repair without relying on a combination of ontology-based semantic reasoning, Formal Concept Analysis for detection, and clustering for automated correction. This study fills this gap..

III. METHODOLOGY

This study is based on the Total Data Quality Management (TDQM) approach which follows the Deming cycle for data quality management [46], [47]. The TDQM cycle consists of four steps: define, measure, analyse, improve. In the definition phase, attributes like data needs are specified, and the importance of data quality dimensions are outlined. During the measurement phase data quality metrics are defined and used to measure the quality dimensions. During the analysis phase, the causes of data quality issues that were uncovered during the measurement phase are explored. Lastly, in the improvement stage, causes of quality issues are eliminated.

A. Data Quality Dimensions and Data Requirements

The two most predominant aspects of data quality in big data are consistency and completeness [15]. Consistency refers to the uniformity of certain aspects of data and systems when data sets are moved over networks and shared by various applications. Completeness is the degree of completeness of data; that is, the degree of containing all expected data values. These are especially important when data are integrated into a data warehouse from multiple data sources [5], [48]. While conventional data quality dimensions (e.g., accuracy, timeliness, credibility) are still important, consistency and completeness are the most often mentioned dimensions for big data measurement. When data from heterogeneous sources are integrated into a data warehouse, missing values, multi-meaning attributes and inconsistent schemas are often encountered [5], [49]. A consolidated data warehouse sourced from multiple sources is the typical case where inconsistent and incomplete data are most likely to happen, as the integration process identifies and conciliates different problems sourced from source data [5]; and [50] highlighted that data collection and quality issues (e.g., incompleteness, missing values) significantly degrade the accuracy of predictive analytics and thus business decision-making.

According to International Organization for Standardization (ISO) [51] a requirement is a “need or expectation that is stated, generally implied or obligatory”, The concept has been extended to the data context; this "data requirement" means needs and expectations about data, which are stated, generally implied and obligatory. Data requirements are very important in the data quality management process. They are first defined and expressed in the definition phase and translated into metrics in the measurement phase to produce reports on poor data. In this study the data requirements taken are:

i. *Property completeness requirements:* Mark requirements for data values in a particular attribute for all instances or for some instances of a table..
ii. *Syntactic requirements:* define the character and/or pattern of attributes..
iii. *Legal value requirements:* Specify the required values for a specific attribute.
iv. *Legal value range requirements:* Define the value range for a certain numeric attribute.
v. *Functional dependency requirements:* Model a relationship between values of two or more attributes in a table or between values of two or more attributes in different tables.

B. Proposed Model Architecture

The proposed model is based on the Extract-Load-Transform approach, which is commonly used with big data.

Traditionally, ETL processes and systems have been defined as those that physically combine data from several different sources into a single repository known as a data warehouse [52], [53]. All ETL systems involve extracting data from a source system, transforming the data into a likely different format, and then loading the data into the target system.

The ELT process has increased with the advent of Hadoop and reductions in hardware prices. Data are extracted from sources and instead of transforming the data first, the data are loaded into a database. The loading may not even require a schema and the data can be stored in their raw form for a long time. When the data are required, experts build a schema and transform them. The advantage of ELT is that the data are stored in its original form for a longer time, which allows experts to transform the data stored in a suitable way for the given task [54]. The Five-phase ELT model (Fig.1) shows Data Extraction, Loading, Semantic Integration, Quality Assessment using FCA and Data Quality Ontology (DQO), and Quality Improvement using Clustering.

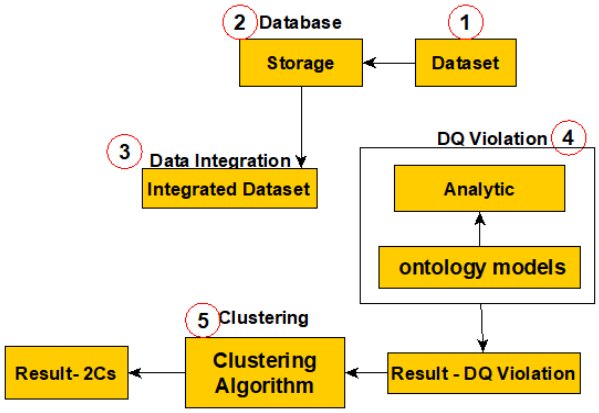

Fig. 1 Five-phase semantic data quality assessment model (ELT-based)

The proposed model has five phases: (1) Data Extraction from different data sources; (2) Data Loading into storage; (3) Semantic Integration using ontology models; (4) Quality Assessment using FCA and DQ ontology; and (5) Quality Improvement using clustering algorithms. The phases are detailed in the following subsections.

C. Semantic Integration Using Ontology

Big data originates from various sources and integration is important. In phase 3, ontology models are used for semantic integration. Ontology is a formal representation of a shared conceptualization [55]. Each concept in an ontology is uniquely named, has properties, and is related to other concepts and has constraints that must always hold for that concept. The advantages of using an ontology for data integration are: a formal language representation, a semantic representation, it provides a shared vocabulary, and it's rigorous enough for use in algorithms and computer programs.

A number of Semantic Web technologies support the organisation of knowledge into a conceptual space according to meaning, enable the discovery of new knowledge through querying, and support the maintenance of knowledge by testing for inconsistencies. These include: (i) the Uniform Resource Identifier (URI), a name or identifier for a web resource that enables interaction with representations of that resource over a network; (ii) the Resource Description Framework (RDF), a standard for representing information about web resources and exchanging data on the Web; and (iii) the Web Ontology Language (OWL), an ontology language for the Semantic Web built on RDF, which supports object properties (relating individuals to other individuals) and data properties (relating individuals to literal values).

D. Data Quality Ontology

We developed a data quality ontology (Fig. 2) to facilitate the identification and evaluation of data quality violations by applying rules and the chosen dimensions (consistency and completeness). Ontology is one of the key components of semantic solutions. The key elements of ontologies in Semantic Web systems are classes and properties. Properties are in the predicate position of a triple and represent relationships between resources, or facts about resources. Classes are categories that group resources [56].

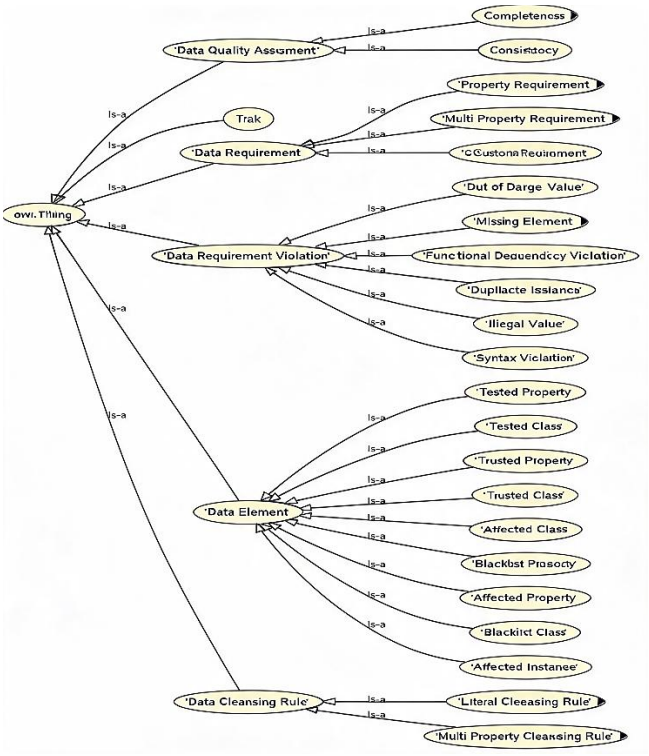

Fig. 2 Data quality ontology

The DQ ontology was developed starting from the data quality dimensions to model inconsistency and incompleteness in the domain of interest, and to provide a justification for its use. The competency questions were stratified and used as ontology requirements, and relevant terms were identified and informally defined for use during ontology development. These terms were classified into classes, object properties and relationships. The ontology contains 47 classes, 46 object properties and 50 data type properties in OWL DL to facilitate reasoning, in Protégé-OWL:

- *Data Requirement:* A prescribed directive defining high-quality data characteristics. Subclasses include CustomRequirement, MultiPropertyRequirement, and PropertyRequirement.
- *Data Element:* A class, property, instance, or literal value. Subclasses include:

  AffectedClasses, AffectedInstance, AffectedProperty, TestedClass, TestedProperty, TrustedClass, and TrustedProperty.

- *Data Requirement Violation:* Occurs when data values do not meet requirements. Subclasses include DuplicateInstance, FunctionalDependencyViolation, IllegalValue, MissingElement, OutOfRangeValue, and SyntaxViolation.
- *Data Quality Assessment:* Structures DQD scores indicating quality states. Subclasses include Completeness and Consistency dimensions.
- *Data Cleansing Rule:* Stipulates required states of data values with subclasses LiteralCleansingRule and MultiplePropertyCleansingRule.

E. Formal Concept Analysis

FCA is an unsupervised classification technique based on the extraction of all objects that share all attributes (intent) and all attributes that are shared by all objects (extent), which are represented in the form of an object-attribute lattice (also called a Galois lattice [57]). When applied to inconsistency and incompleteness detection in big data, FCA offers a better identification and visualization capability than traditional plots, bar-charts and histograms that are ineffective in high dimensional data [58], [59]. It explores associations, refinements, errors and inconsistencies, and enables ontology plugins to define data quality requirements using sets of attributes. Formally, FCA is based on a binary context (G, M, I), where G and M are sets of objects and I is a relation between them that represents the conceptual knowledge..

For A subset of $G, A' = \{m\ in\ M \mid$ for all g in A: (g, m) in $I\}$, and dually for B subset of $M, B' = \{g\ in\ G \mid for$ all m in B: $(g, m)\ in\ I\}$.

A formal concept is defined as a pair (A, B) where A subset of G, B subset of $M, A' = B$ and $A = B'$. A is the extension and B is the intension of the formal concept (A, B).

F. Clustering for Quality Improvement

We use the K-means clustering technique to correct data quality issues. Clustering is a way of grouping a collection of objects into clusters so that objects in the same cluster are more similar to each other than to objects in other clusters. Clustering attempts to detect similarity between strings and clusters them according to a threshold of similarity, eliminating data inconsistencies and incompleteness according to requirements specified at the DQ ontology stage.

The initial approach to solving data inconsistencies and incompleteness is to pre-process the data to remove data quality violation datatypes. Using the knowledge gained from the data quality assessment stage, requirements from the DQO ontology phase are used to perform clustering on the data, focusing on data linkage to resolve data quality issues.

Once pre-processing using assumptions of DQ violations is applied, K-means clustering is applied to the dataset. The K-means objective function is:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 \quad (1)$$

where $||x_i - v_j||$ is the Euclidean distance between x_j and v_j, c_j is the number of data points in the i-th cluster, and c is the number of cluster centers. Using equation 1, K-means clustering is performed, using DQ rules to get the number of clusters, check the results, and improve the data quality.

G. Implementation and Environment

To support independent verification of the results, per the Editorial Board's request, this subsection reports the concrete implementation and evaluation settings used in this study.

• *Ontology development:* Protégé-OWL, using OWL DL as the representation language (47 classes, 46 object properties, 50 data type properties).

• *Formal Concept Creator:* FcaBedrock (Formal Context Creator, build 2.8.0.23), with progressive scaling, standard-deviation binning, and both Inconsistency Mode and Incompleteness Mode enabled (Fig. 5).

• *Data view/integration inspection:* a tabular data-management platform (DataRobot) was used to inspect integrated columns (id, book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, etc.) during the data integration step (Fig. 4).

• *Clustering:* K-means clustering as defined in Eq. (1); the number of clusters (k = 11) was determined by the granularity of the DQ ontology's rule-based grouping of the 264 attribute-violation objects identified during assessment, rather than by a separate hyperparameter search. A formal cluster-validity procedure (e.g., the elbow method or silhouette analysis) was not performed in the present study; we flag this explicitly as a methodological limitation in Section V.

• *Hardware environment:* experiments were run on a workstation equipped with a 13th-generation Intel Core i7 processor, 16 GB RAM, and 1 TB SSD storage.

• *Software:* Python 3.9.13, scikit-learn 1.2.0, pandas 2.0.0, numpy 1.24.0, matplotlib 3.7.0, Protege 5.5.0

IV. Experiments And Results

A. Dataset Description

The datasets were analysed to establish the integrated data needed for this study. The case study uses more than 10,000 books from Goodreads integrated into a single dataset. This dataset exhibits big data attributes, particularly in terms of volume and variety. Ratings are provided for mainstream books with typically 100 ratings per book, but some contain fewer ratings. Ratings are on a scale of 1-5. Both book IDs and user IDs are contiguous (books: 1-10000, users: 1-53424). Every user has at least 2 ratings, with a median of 8 ratings. The dataset also includes books to read, book metadata (author, year, etc.), and tags. Table 1 presents the dataset profile, while Table 2 describes the data fields in detail.

TABLE 1 DATASET PROFILE

Attribute	Value
Number of Books	10,000+
Number of Users	53,424
Rating Range	1-5 (integer scale)
Minimum Ratings per User	2
Median Ratings per User	8
Data Types	Ratings, "to-read" lists, metadata, tags

TABLE II DATA FIELD DESCRIPTIONS

Field	Type	Description	Null Allowed?
Book ID	Integer	Primary key (1-10000)	No
User ID	Integer	Foreign key (1-53424)	No
Title	String	Book title	No
Author	String	Author name	No
Published Year	Integer	Publication year (4 digits)	No
Publisher	String	Publisher name	Yes
Editor	String	Editor name	Yes

Rating	Float	User rating (1.0-5.0)	No
Tags	String	Comma-separated tags	Yes

To ensure data quality and prepare the dataset for analysis, the following preprocessing steps were applied:

Data Preprocessing Pipeline

Step 1: Data Loading and Integration

Load individual source datasets (ratings, to-read, metadata, tags)
- Perform inner join on Book ID
- Filter records with Book ID between 1 and 10000
- Filter users with at least 2 ratings

Step 2: Data Validation

For each record:
Check if Published Year is not null
Check if Published Year matches regex '^[0-9]{4}'
Check if Author Name is not null
Check if Title is not null
Check functional dependency: Title -> (Publisher, Editor)
Check functional dependency: (Year, Author) -> Title

Step 3: Violation Filtering

Identify violation types from DQ ontology:
- Incomplete Published Year → flag or impute
- Missing Author Name → flag
- Editor/Publisher Duplication → consolidate
- Functional Dependency Breaks → flag

Apply filtering based on violation severity:
- Remove records with >50% missing values
- Mark records with partial violations for repair
- Keep records with no violations as ground truth

Step 4: Data Preparation for FCA

- Convert each record to a set of attributes
- Create formal context (G, M, I) where:
G = set of all data records
M = set of attributes (fields + violation flags)
I = binary relation (record g has attribute m)

A class diagram was obtained from the assessment of a variety of sources to assist in the integration of data and the creation of a useful dataset so data problems in the book domain can be tackled. Fig. 3 shows the class diagram, which helps identify and categorise the features of book data from various sources.

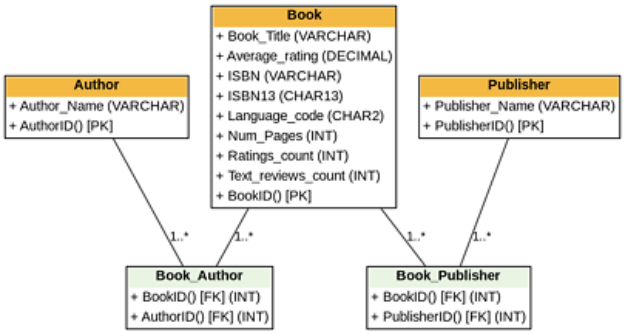

Fig. 3 Class diagram for Book Domain

The integrated dataset comprises book IDs linked to all the datasets, with filtering applied during data integration to adhere to the requirements. Fig. 4 shows the process of data integration.

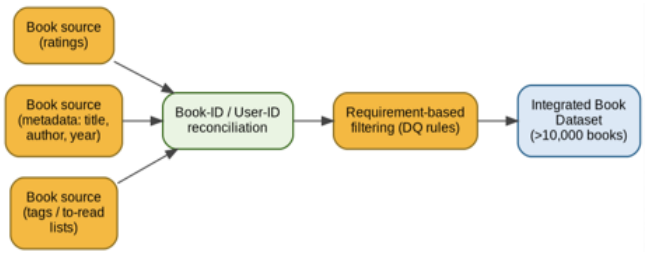

FIG. 4 Data Integration Process for the Goodreads Book Dataset

B. *Requirement Violation Identification*

The DQ ontology was used to analyse the integrated data using FCA. This revealed data inconsistency and incompleteness that impacts data quality. Table 3 illustrates the report obtained from the analysis and rules to detect and identify requirement violations.

TABLE III: REQUIREMENT VIOLATIONS AND RULES

Report	Rule
Incomplete values	Field 'published year' must have 4-digit value. Every book must have a published year.
Incomplete values	Field 'Author name' must always have a value.
Inconsistent values	Field 'editor' and 'publisher' have same values in different fields.
Functional dependency	Book title determines publisher and editor. Year and author determine title.

As shown in Table 3, the integrated dataset shows significant data consistency violations as well as incompleteness, due to the integration of multiple sources. In particular, inconsistencies are observed in a lack of uniformity in attribute values (e.g., varying units, formats or encoding), and incompleteness is observed in missing attribute values for records.

These problems clearly breach two aspects of DQ: consistency and completeness. Fig. 5 is the diagnosis of our automated requirement violations analysis, which relates the identified anomalies to DQ dimensions shown in Table 4.

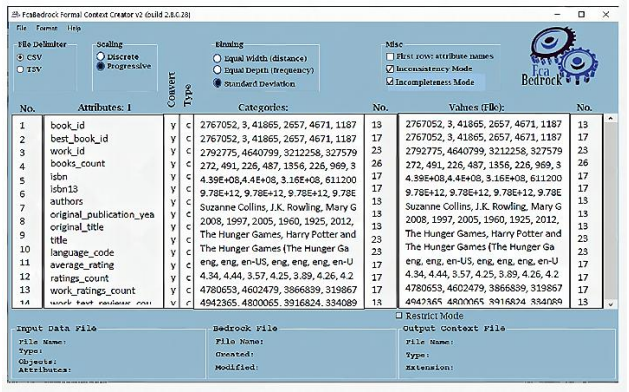

Fig. 5 Incomplete and Inconsistent Data Identification

TABLE IV MAPPING OF IDENTIFIED VIOLATIONS TO DATA QUALITY DIMENSIONS

Identified Requirement Violation	Violated DQ Dimension	Description
Missing Published Year	Completeness	Missing values in the published year field
Invalid Published Year	Consistency	Values not conforming to a 4-digit format (e.g., "202", "20250")
Missing Author Name	Completeness	Missing values in the author name field
Editor & Publisher Duplication	Consistency	Same value exists in two different fields

Functional Dependency Violation	Consistency	Missing author name or title breaks the dependent relationship

Fig. 6 demonstrates the visibility of the identified DQ problem in the database. The incompleteness shows in the area of published year which identified missing year, year value with less than four digits, and with more than four digits. It also identified incompleteness in author name, as there are 15 missing author name values in the data.

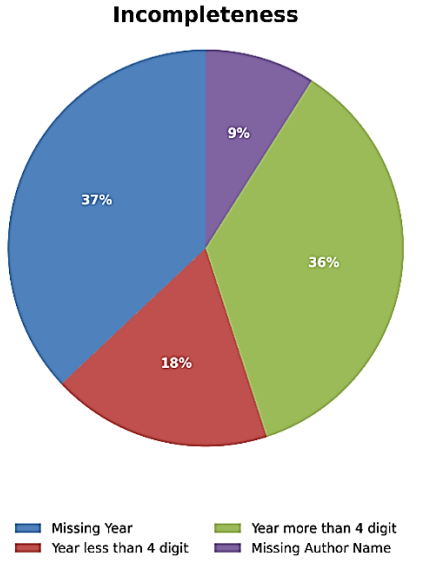

Fig. 6 Incomplete Values and Properties Distribution

The inconsistent values and properties are identified in the area of editor and publisher as they bear the same value but are separated, leading to inconsistency. Fig. 7 shows the editor and publisher fields before and after combination to solve the inconsistent value.

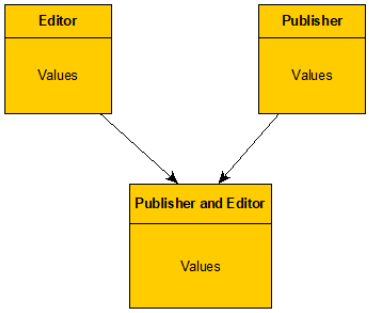

Fig. 7 Inconsistent Values and Properties Before and After Combination

The results identified issues of functional dependent value which contributes to data quality problems if dependent values are missing. In this case, missing author names lead to incompleteness and inconsistency as title is dependent on author and vice versa.

C. Fixing Data Quality Violations

The K-means clustering algorithm was used to resolve the identified data quality violations and to improve the data quality. The first method was pre-processing to remove violation data types. The clustering algorithm mainly targeted consistent and complete data types as identified by the DQ rules.

The purpose of clustering is to identify similarity of strings and cluster them based on a similarity threshold, and remove anomalies according to rules defined in Table 3 during pre-processing. Following the pre-process stage, K-means algorithm was applied. The K-means algorithm was configured with the following hyperparameters:

Number of clusters (k) = 11 (determined by the count of violation types from the DQ ontology)
Maximum iterations = 15 (convergence achieved by iteration 10)
Distance metric = Euclidean
Initialization = k-means++

The result of the clustering is shown in Fig. 8. There were 264 objects with 11 clusters grouped as a result of the clustering..

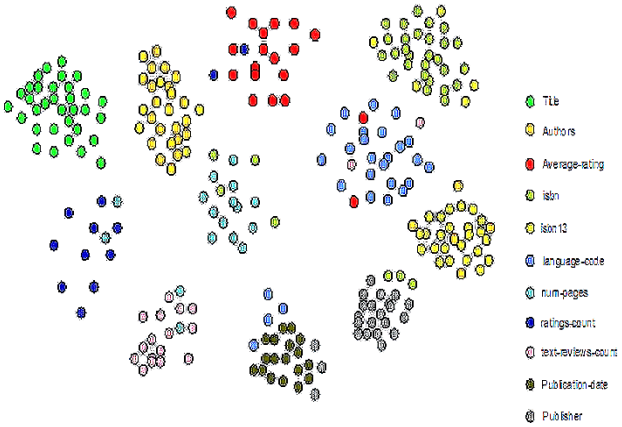

Fig. 8 Clustering Result

From the clustering result, 264 objects clustered and resulted in 11 cluster groups with the maximum iteration of 15 for the case study.

D. Evaluation Procedures

The model's performance was evaluated using a comprehensive set of metrics commonly employed in data quality assessment and classification tasks.

Precision measures the fraction of returned information that is relevant, representing the quality of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the proportion of relevant information found, representing the completeness of violation detection:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Accuracy measures the overall proportion of correct classifications:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

F1-Score is the harmonic mean of precision and recall, providing a balanced evaluation that is particularly useful for skewed data [47]:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Matthews Correlation Coefficient (MCC) is a balanced measure that can be used even when classes are of very different sizes. A high MCC value can only be attained when classification performs well on both classes [48]:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

Specificity measures the proportion of actual negatives correctly identified, indicating the model's ability to correctly classify valid data:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (7)$$

Negative Predictive Value (NPV) measures the proportion of negative predictions that are correct, indicating the probability that data flagged as valid is truly valid:

$$\text{NPV} = \frac{TN}{TN + FN} \quad (8)$$

These metrics provide a comprehensive assessment of model performance, addressing different aspects of classification quality.

E. Performance Results and Comparative Study

The data has 11 clusters or classes and 264 objectives or features (calculated by counting the number of all objective pairs). Table 2 shows the confusion matrix.

TABLE V: CONFUSION MATRIX FOR PRECISION AND RECALL

	Same Cluster	**Different Clusters**	**Total**
Same Class	TP=468	FN=39	507
Different Class	FP=60	TN= 22402	22462
Total	528	22441	22969

The performance of the proposed model is represented in Table 6. The high precision (88.6%) indicates that when the model identifies a data quality violation, it is highly likely to be a true violation. This can be attributed to the formal rigor of the DQ ontology combined with FCA's mathematical foundation in set theory. The ontology provides explicit, formal definitions of what constitutes a violation, while FCA's Galois lattice structure ensures that detected patterns correspond to genuine conceptual relationships in the data

TABLE VI: CONFUSION MATRIX FOR PRECISION AND RECALL

Metric	**Value**	**Description**
Precision	88.60%	Quality of positive predictions
Recall	92.30%	Completeness of violation detection
Accuracy	99.57%	Overall correctness
F1-Score	90.41%	Balanced measure (harmonic mean)
MCC	90.25%	Balanced correlation coefficient
Specificity	99.73%	Correct identification of valid data
NPV	99.83%	Probability valid data is truly valid

It supports all the classes, object properties, and data type properties of the full DQ ontology (47 classes, 46 object properties, 50 data type properties), enabling FCA's unsupervised pattern detection to reveal known and novel problems. The high F1 score (90.41%) and MCC score (90.25%) show that the model is very reliable and performs well, providing a comparable performance across all classes to reduce the chances of incurring unnecessary false positives and false negatives. It performs near-perfect NPV, specificity and overall accuracy (99.83%, 99.73% and 99.57%, respectively), accurately validating clean records in a class distributed dataset (22402 valid observations versus 507 violations). Afterwards the identified violations were successfully clustered using the K-means algorithm, constructing 11 clusters representing 11 different violation types that closely matched the particular 11 violation types specified in the DQ ontology. This was streamlined by a pre-clustering filter, which only took those violations that were relevant into account, ensuring efficient and coherent clustering according to different quality characteristics.

F. Comparative Study

Table 7 shows a comparison between the proposed model and the recent state-of-the-art models, which highlights the potential of the proposed model for data quality assessment and improvement.

TABLE VII COMPARISON OF PERFORMANCE METRICS WITH STATE-OF-THE-ART METHODS

Study	**Approach**	**Year**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Al-Toq and Almaslukh [16]	ML-based DQMAF	2025	82.4	79.1	80.7
Oeste et al. [38]	Clinical NLP Algorithm	2024	88.6	79.4	83.7
Belgacem et al., [21]	Anomaly Detection (Categorical Data)	2024	≥80.8	60.0–100	—
Proposed Model	Ontology + FCA + Clustering	2026	88.6	92.3	90.4

As illustrated in Table 8, the proposed model outperforms the DQMAF [16] with precision of +6.2%, recall of +13.2% and F1-score of +9.7%. Our model achieves similar precision as the clinical NLP algorithm (MRS10) presented by [38] (88.6%) and a considerably higher recall performance (92.3%) with +12.9 percentage points. This underlines the benefit of our ontology-based FCA approach to target a more comprehensive set of data quality violations while maintaining accuracy. Our model performs better with more consistency than LAFF-AD [21] with precision scores > 80.8% and recall scores between 60.0% to 100% across datasets.

TABLE VIII COMPARATIVE PERFORMANCE GAINS OVER EXISTING METHODS

Comparison	**Precision Gain**	**Recall Gain**	**F1-Score Gain**
vs. DQMAF [16]	+6.2%	+13.2%	+9.7%
vs. Oeste et al. [58]	0.0%	+12.9%	+6.7%
vs. LAFF-AD [59]	+7.8%	Comparable	—

Several factors can be responsible for these improvements. Unlike DQMAF [16] that has no automated repair function, our model offers a combination of detection and correction, giving it an end-to-end solution. Moreover, since the DQ ontology (47 classes, 46 object properties, 50 data type properties) provides semantics about data quality, it is possible to specify data requirements formally and be able to detect subtle violations not identified through purely ML-based or rule-based approaches. The unsupervised nature of FCA avoids the need to have labeled training data, thus our approach is more flexible in the new domains, unlike supervised ML approaches [16] which require access to available labeled datasets. Additionally, the mathematical

foundations underpinning FCA give it powerful pattern detection capabilities and interpretability, which can provide benefits over some ML approaches known for their 'black-box' nature. The results show that the proposed combination of ontology-based data quality assessment and FCA and clustering achieve better results in all data quality measure, which proves that the combination provides a comprehensive coverage for diagnosing and enhancing data quality issues.

V. Discussion

The assessment results show the effectiveness of the semantic data quality assessment approach at detecting and addressing the various data quality violations in big data settings. Ontology-based data quality assessment, Formal Concept Analysis (FCA) and clustering algorithms provide a comprehensive form of solution for both data quality detection and improvement. Our work is integrated approach, which separates from the existing works focusing on detection or improvement, but not detection and improvement. The precision (88.6%) and recall (92.3%) values obtained confirm that the model is able to identify most of the data quality violations with few false positives, making it useful to use in big data applications.

The high precision is due to the formalism of the DQ ontology and the mathematical background of FCA, which is based on set theory. The ontology gives explicit and formal definitions of violation, while the structure of FCA's Galois lattice guarantees that the patterns detected in the data are based on true conceptual relationships in the data. This formal process minimises the number of false positives, especially in a big data environment where unnecessary flagging would be a waste of resources, and undermine the trust in the system. The recall (92.3%) highlights the capability of the model to represent the majority of the real-world data quality problems, which is possible thanks to the extensive scope of the DQ ontology (47 classes, 46 object properties and 50 data type properties) and thanks to the unsupervised pattern detection used by FCA. The FCA does not need "labeled" training data as in supervised learning methods, and is capable of uncovering new quality problems that may not have been known before allowing it to be used in new domains.

The F1 score of 90.41% gives a balanced performance of the model, balancing precision and recall. Especially in the case of data quality assessment – false positives as well as false negatives have practical implications. The MCC of 90.25% is further an indicator of good performance in both classes and shows that the model is not biased towards either class. The near-perfect specificity (99.73%) and NPV (99.83%) indicate that the model has an outstanding accuracy in detecting valid data and reducing the number of false-positives. The effectiveness of the ontology-guided clustering approach was demonstrated by the number of clusters constructed, which is commensurate to the number of distinct violation types that are defined in the DQ ontology, which in turn clearly indicates the cohesiveness of the clustering approach.

The evaluation of the model shows that the proposed model performs quite well, outperforming the DQMAF presented by [16] in all the reported numbers by (+6.2%) for precision, (+13.2%) for recall, and (+9.7%) for F1-score respectively. It outpaces [38] on both recall (92.3% vs. 79.4%) and precision (88.6%) showing that our ontology-driven FCA approach is able to detect a wider spectrum of data quality violations without reducing precision. Our model is able to have a more uniform performance and precision than LAFF-AD in [21] which has less than 80.8% precision across different datasets. Kilpatrick et al. [37] and Qin et al. [39] reported higher F1-scores but their approaches are domain-specific and do not offer a general purpose, ontology-driven framework that is able to combine semantic reasoning with automatic repair, an essential part of our model. As demonstrated by ELT-Bench results [45] lack of strong caution should be taken with over-reliance on complex AI agents; our model provides a more practical and effective application without the computational cost and randomness associated with complex AI agents.

A. Limitations And Future Work

The proposed model exhibits good performance but there are some limitations that need to be recognized. The present model is based solely on completeness and consistency issues. The ontology could be extended to encompass other dimensions like new ones such as accuracy, timeliness and credibility. Second, the evaluation has been only carried out for one dataset from the book rating domain, and needs validation on other domains like healthcare, IoT and finance etc. to ensure the generalization. Third, for very high dimensional datasets, FCA's computational complexity may cause problems and needs optimization or alternative implementation. Fourth, K-means algorithm needs specified number of clusters (k), which was defined according to the number of violation types that were found by the DQ ontology; adaptive clustering could be used for dynamic violation patterns. Fifth, the development of the DQ ontology is labor-intensive and demands a good domain knowledge, that could hinder its implementation if there is no ontology engineering capabilities in the environment.

To overcome these limitations, the next steps involve the extension of the ontology to incorporate more quality dimensions, testing on larger and more diverse datasets, evaluation of scalability optimizations for the FCA using distributed computing frameworks, development of adaptive clustering techniques, testing of semi-automated ontology learning techniques, and integration with deep learning approaches to facilitate pattern recognition. However, the model's success on the Goodreads dataset indicates that it might be applicable to other big data usage scenarios, and could be adapted to other datasets that have other data ontologies.

VI. Conclusion

These data quality problems have grown worse in large data due to the huge number and speed of the data coming into existence. This paper proposed a semantic data quality assessment model to solve the problem of inconsistency and incompleteness of big data. The developed framework is an integrated solution comprising of three major elements: Extract-Load-Transform (ELT) architecture, ontology-based data modeling, the Formal Concept Analysis based violation detection and the automated quality improvement based on the K-means cluster.

The model was shown to be effective with the experimental validation conducted on a Goodreads dataset of more than 10,000 books. So the approach had a precision of 88.6% and a recall of 92.3%, which proves that it was successful in identifying violations without producing too many false alerts, while detecting the most cases of actual data quality violations. The model was also robustly validated with additional metrics of F1-score (90.41%), MCC (90.25%),

accuracy (99.57%), specificity (99.73%) and NPV (99.83%). Comparative analysis showed that the precision obtained was +6.2%, the recall +13.2% and F1 score +9.7% over the state-of-the-art methods reviewed in Tables 7 and 8 in applicable metrics.

This research has four important contributions to big data quality management. Quality requirement and quality violation for consistency and completeness dimensions are comprehensively defined by the formal quality ontology with 47 classes, 46 object properties and 50 data properties. Second, the application of the FCA to the ontology gives mathematical rigor and useful visualization features for quality pattern detection, which are not provided by traditional methods. Thirdly, the use of K - Means Clustering leads to automated repair, eliminating the need to have a manual application to detect and improve the application. Fourth, the detailed information on the reproduction of the experiments, which has been provided on a real-world dataset, proves the effectiveness of the combined approach and allows the testing of the approach by other researchers in the field.

There were a number of caveats that suggest fruitful avenues for future research. The current model is limited to completeness and consistency and further development is desired on other dimensions such as accuracy, timeliness and credibility. There will be further generalizability based on validation across a variety of domains, including healthcare, IoT, and finance. Refining features such as scalability optimizations for FCA, adaptive clustering for dynamic violation patterns and integrating deep learning techniques into the framework are valuable directions for the future.

To sum up, the semantic data quality evaluation model proposed in this paper is a practical and effective solution to the problem of data quality management for big data. The framework brings together semantic technologies, formal analysis, and machine learning to deliver the rigour, automation, and flexibility that is essential to deal with veracity issues in today's data ecosystems. The possibility to demonstrate all the these improvements over previous approaches, the details of the framework and its reproducibility make this work valuable for research and implementation of data quality management.

REFERENCES

[1] I. A. T. Hashem *et al.*, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, Oct. 2016, doi: 10.1016/j.ijinfomgt.2016.05.002.

[2] C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable Big Data: A survey," *Computer Science Review*, vol. 17, pp. 70–81, Aug. 2015, doi: 10.1016/j.cosrev.2015.05.002.

[3] M. Romanchuk, "Rethinking Data Architectures in the Face of Information Diversity and Exponential Growth," *Asian J. Res. Com. Sci.*, vol. 18, no. 12, pp. 97–116, Dec. 2025, doi: 10.9734/ajrcos/2025/v18i12793.

[4] P. Koukaras, "Data Integration and Storage Strategies in Heterogeneous Analytical Systems: Architectures, Methods, and Interoperability Challenges," *Information*, vol. 16, no. 11, p. 932, Oct. 2025, doi: 10.3390/info16110932.

[5] I. M. Putrama and P. Martinek, "Heterogeneous data integration: Challenges and opportunities," *Data in Brief*, vol. 56, p. 110853, Oct. 2024, doi: 10.1016/j.dib.2024.110853.

[6] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, USA: IEEE, 2018, pp. 166–173. doi: 10.1109/BigDataCongress.2018.00029.

[7] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Computers & Industrial Engineering*, vol. 115, pp. 319–330, Jan. 2018, doi: 10.1016/j.cie.2017.11.017.

[8] R. P. Deb Nath, O. Romero, T. B. Pedersen, and K. Hose, "High-level ETL for semantic data warehouses," *SW*, vol. 13, no. 1, pp. 85–132, Nov. 2021, doi: 10.3233/SW-210429.

[9] A. A. A. Fernandes, M. Koehler, N. Konstantinou, P. Pankin, N. W. Paton, and R. Sakellariou, "Data Preparation: A Technological Perspective and Review," *SN COMPUT. SCI.*, vol. 4, no. 4, p. 425, Jun. 2023, doi: 10.1007/s42979-023-01828-8.

[10] S. Borowicc and S. Alves-Souza, "Heterogeneous Data Integration: A Literature Scope Review:," in *Proceedings of the 26th International Conference on Enterprise Information Systems*, Angers, France: SCITEPRESS - Science and Technology Publications, 2024, pp. 189–200. doi: 10.5220/0012551000003690.

[11] K. Patel and M. Bera, "Big Data Quality: A Systematic Literature Review," *IJSREM*, vol. 07, no. 10, pp. 1–11, Oct. 2023, doi: 10.55041/IJSREM26188.

[12] L. Clissa, M. Lassnig, and L. Rinaldi, "How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry," *Front. Big Data*, vol. 6, p. 1271639, Oct. 2023, doi: 10.3389/fdata.2023.1271639.

[13] A. Van Gulick, J. Richardson, and A. Doran, "Defining Big Data for Generalist Repositories," Mar. 2025, doi: 10.5281/ZENODO.14782996.

[14] Ü. Demirbaga, G. S. Aujla, A. Jindal, and O. Kalyon, "Real-World Big Data Analytics Case Studies," in *Big Data Analytics*, Cham: Springer Nature Switzerland, 2024, pp. 233–247. doi: 10.1007/978-3-031-55639-5_10.

[15] N. Wang and S.-G. Leem, "Modernizing Data Quality: Evolving Dimensions for Unstructured Text Data in Big Data, AI and Ethical Contexts," in *Data Quality Matters - Best Practices for Integrity and Assurance [Working Title]*, IntechOpen, 2026. doi: 10.5772/intechopen.1013232.

[16] R. Al-Toq and A. Almaslukh, "DQMAF—Data Quality Modeling and Assessment Framework," *Information*, vol. 16, no. 10, p. 911, Oct. 2025, doi: 10.3390/info16100911.

[17] J. Martinez-Gil, "Framework to automatically determine the quality of open data catalogs," *Expert Systems with Applications*, vol. 289, p. 128379, Sep. 2025, doi: 10.1016/j.eswa.2025.128379.

[18] H. Fadlallah, R. Kilany, M. Haber, and A. Jaber, "Automated big data quality assessment using knowledge graph embeddings," *IJDMMM*, vol. 17, no. 4, pp. 383–405, 2025, doi: 10.1504/IJDMMM.2025.150987.

[19] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J Big Data*, vol. 12, no. 1, p. 268, Dec. 2025, doi: 10.1186/s40537-025-01313-4.

[20] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sensing*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.

[21] H. Belgacem, X. Li, D. Bianculli, and L. Briand, "Automated anomaly detection for categorical data by repurposing a form filling recommender system," *J. Data and Information Quality*, vol. 16, no. 3, pp. 1–28, Sep. 2024, doi: 10.1145/3696110.

[22] G. Stănescu (Nicolaie) and S.-V. Oprea, "Recent Trends and Insights in Semantic Web and Ontology-Driven Knowledge Representation Across Disciplines Using Topic Modeling," *Electronics*, vol. 14, no. 7, p. 1313, Mar. 2025, doi: 10.3390/electronics14071313.

[23] Y. Wu, M. Ren, N. Chen, and L. Yang, "Semantics-driven improvements in electronic health records data quality: a systematic review," *BMC Med Inform Decis Mak*, vol. 25, no. 1, p. 298, Aug. 2025, doi: 10.1186/s12911-025-03146-w.

[24] R. Ambalavanan, R. S. Snead, J. Marczika, G. Towett, A. Malioukis, and M. Mbogori-Kairichi, "Ontologies as the semantic bridge between artificial intelligence and healthcare," *Front. Digit. Health*, vol. 7, p. 1668385, Aug. 2025, doi: 10.3389/fdgth.2025.1668385.

[25] T. S. Balakrishnan, L. Karthikeyan, G. Senthilvel. P, U. S. Ebenezar, and P. B. S, "Ontology-Driven Semantic Interoperability Framework for Multi-Center Healthcare Big Data with Graph-Based Query Analysis," in *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)*, Bangalore, India: IEEE, Mar. 2025, pp. 1–6. doi: 10.1109/COMP-SIF65618.2025.10969920.

[26] F. Zeshan, A. Ahmad, M. I. Babar, M. Hamid, F. Hajjej, and M. Ashraf, "An IoT-Enabled Ontology-Based Intelligent Healthcare

Framework for Remote Patient Monitoring," *IEEE Access*, vol. 11, pp. 133947–133966, 2023, doi: 10.1109/ACCESS.2023.3332708.

[27] H. Mohammadhassanzadeh, S. Raza Abidi, and S. S. Raza Abidi, "Plausible reasoning over large health datasets: A novel approach to data analytics leveraging semantics," *Knowledge-Based Systems*, vol. 289, p. 111493, Apr. 2024, doi: 10.1016/j.knosys.2024.111493.

[28] G. Cima, M. Console, and M. Lenzerini, "Ontology-Based Schema-Level Data Quality: The Case of Consistency," *J. Data and Information Quality*, vol. 17, no. 4, pp. 1–25, Dec. 2025, doi: 10.1145/3770750.

[29] P. Chandur, H. Shah, K. Linkalapalli, R. Kovvuri, and S. Medam, "Semantic Data Integration for Heterogeneous Big Data Sources," in *2025 2nd Asia Pacific Conference on Innovation in Technology (APCIT)*, MYSORE, India: IEEE, Sep. 2025, pp. 1–7. doi: 10.1109/APCIT65661.2025.11410895.

[30] J. Mäkelburg, Z. Zacouris, J. Ke, and M. Acosta, "SHACL Dashboard: Analyzing Data Quality Reports Over Large-Scale Knowledge Graphs," in *The Semantic Web – ISWC 2025*, vol. 16141, D. Garijo, S. Kirrane, A. Salatino, C. Shimizu, M. Acosta, A. G. Nuzzolese, S. Ferrada, T. Soulard, K. Kozaki, H. Takeda, and A. L. Gentile, Eds., in Lecture Notes in Computer Science, vol. 16141. , Cham: Springer Nature Switzerland, 2026, pp. 96–112. doi: 10.1007/978-3-032-09530-5_6.

[31] O. P. Dwyer, L. Chammas, E. Sallinger, and J. Davies, "Using ontologies to facilitate healthcare process mining and analysis," *J Intell Inf Syst*, May 2025, doi: 10.1007/s10844-025-00942-8.

[32] C. Tayal, "AI-Enhanced ETL Framework for Improving Data Quality in Clinical Decision Support Systems," *IJAIDSML*, vol. 5, 2024, doi: 10.63282/3050-9262.IJAIDSML-V5I2P113.

[33] S. Benbernou, X. Huang, and M. Ouziri, "Semantic-Based and Entity-Resolution Fusion to Enhance Quality of Big RDF Data," *IEEE Trans. Big Data*, vol. 7, no. 2, pp. 436–450, Jun. 2021, doi: 10.1109/TBDATA.2017.2710346.

[34] S. Kamatala, A. K. Jonnalagadda, and P. K. Myakala, "Trends and Challenges in Data Cleaning for Large-Scale Systems: A Survey," *SSRN Journal*, 2025, doi: 10.2139/ssrn.5295749.

[35] A. A. Syed, P. K. Sambamurthy, J. Jain, and U. Mamodiya, "A Self-Learning Framework for Enhancing Data Quality in Data-Centric AI Systems through Continuous Feedback," in *2025 International Conference on Computing, Intelligence, and Application (CIACON)*, Durgapur, India: IEEE, Jul. 2025, pp. 1–6. doi: 10.1109/CIACON65473.2025.11189391.

[36] S. S. R. Tummuri, "Machine Learning-Driven Data Quality Monitoring for Fault-Tolerant Data Pipelines," in *2025 4th International Conference on Computational Modelling, Simulation and Optimization (ICCMSO)*, Singapore, Singapore: IEEE, Jun. 2025, pp. 154–159. doi: 10.1109/ICCMSO67468.2025.00036.

[37] K. Kilpatrick, K. Cahill, U. Chandran, and D. Riskin, "Advanced Approaches to Generating High-validity Real-world Evidence in Asthma," *Epidemiology*, vol. 36, no. 1, pp. 20–27, Jan. 2025, doi: 10.1097/EDE.0000000000001803.

[38] C. Oeste *et al.*, "MSR10 Enhancing Data Quality in Health Research: Performance Insights of a Clinical NLP Algorithm for Diverse Medical Domains," *Value in Health*, vol. 27, no. 12, p. S439, Dec. 2024, doi: 10.1016/j.jval.2024.10.2244.

[39] Q. Qin, H. Li, E. Merlo, and M. Lamothe, "Automated, Unsupervised, and Auto-Parameterized Inference of Data Patterns and Anomaly Detection," in *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, Ottawa, ON, Canada: IEEE, Apr. 2025, pp. 2419–2431. doi: 10.1109/ICSE55347.2025.00078.

[40] J. Sora-Cardenas *et al.*, "Image-Based Detection and Classification of Malaria Parasites and Leukocytes with Quality Assessment of Romanowsky-Stained Blood Smears," *Sensors*, vol. 25, no. 2, p. 390, Jan. 2025, doi: 10.3390/s25020390.

[41] M. S. Farhan, A. Youssef, and L. Abdelhamid, "A Model for Enhancing Unstructured Big Data Warehouse Execution Time," *BDCC*, vol. 8, no. 2, p. 17, Feb. 2024, doi: 10.3390/bdcc8020017.

[42] A. Dhaouadi, K. Bousselmi, M. M. Gammoudi, S. Monnet, and S. Hammoudi, "Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons," *Data*, vol. 7, no. 8, p. 113, Aug. 2022, doi: 10.3390/data7080113.

[43] T. C. Ong *et al.*, "Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading," *BMC Med Inform Decis Mak*, vol. 17, no. 1, p. 134, Dec. 2017, doi: 10.1186/s12911-017-0532-3.

[44] Y. Peng *et al.*, "Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review," *JMIR Med Inform*, vol. 12, p. e52967, Feb. 2024, doi: 10.2196/52967.

[45] T. Jin, Y. Zhu, and D. Kang, "ELT-Bench: An End-to-End Benchmark for Evaluating AI Agents on ELT Pipelines," 2025, *arXiv*. doi: 10.48550/ARXIV.2504.04808.

[46] I. G. N. Adi Wicaksana, A. N. Hidayanto, H. Mekkawati, and R. Febriyanti, "DATA QUALITY ASSESSMENT: A CASE STUDY ON ASSET VALUATION COMPARISON DATA," *jitk*, vol. 9, no. 2, pp. 263–272, Feb. 2024, doi: 10.33480/jitk.v9i2.5184.

[47] R. Rahmawati, Y. Ruldeviyani, P. P. Abdullah, and F. M. Hudoarma, "Strategies to Improve Data Quality Management Using Total Data Quality Management (TDQM) and Data Management Body of Knowledge (DMBOK): A Case Study of M-Passport Application," *CommIT (Communication and Information Technology) Journal*, vol. 17, no. 1, pp. 27–42, Mar. 2023, doi: 10.21512/commit.v17i1.8330.

[48] N. H. Chika, "Dealing with inconsistent and incomplete data in a semantic technology setting," doctoral, Sheffield Hallam University, 2015. Accessed: Apr. 22, 2026. [Online]. Available: https://shura.shu.ac.uk/11329/

[49] S. Abdellaoui, L. Bellatreche, and F. Nader, "A Quality-Driven Approach for Building Heterogeneous Distributed Databases: The Case of Data Warehouses," in *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, Cartagena, Colombia: IEEE, May 2016, pp. 631–638. doi: 10.1109/CCGrid.2016.79.

[50] H. Nugroho, "A Review: Data Quality Problem in Predictive Analytics," *IJAIT*, vol. 7, no. 02, p. 79, Aug. 2023, doi: 10.25124/ijait.v7i02.5980.

[51] L. M. Ciravegna Martins Da Fonseca, "ISO 14001:2015: An improved tool for sustainability," *JIEM*, vol. 8, no. 1, pp. 37–50, Feb. 2015, doi: 10.3926/jiem.1298.

[52] C. E. Offia and M. Crowe, "A Theoritical Exploration Of Data Management and Integration in Organisation Sectors," *IJDMS*, vol. 11, no. 01, pp. 37–56, Feb. 2019, doi: 10.5121/ijdms.2019.11103.

[53] V. Debroy, L. Brimble, and M. Yost, "NewTL: engineering an extract, transform, load (ETL) software system for business on a very large scale," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, Pau France: ACM, Apr. 2018, pp. 1568–1575. doi: 10.1145/3167132.3167300.

[54] H. Mallek, F. Ghozzi, and F. Gargouri, "Towards Extract-Transform-Load Operations in a Big Data cont," vol. 12, no. 2, pp. 77–95, Apr. 2020.

[55] R. Taelman, M. Vander Sande, and R. Verborgh, "Bridges between GraphQL and RDF." Accessed: Apr. 25, 2026. [Online]. Available: https://rubensworks.github.io/article-w3cdataws2019-graphql/

[56] Q. Zou, "Represent Changes of Knowledge Organization Systems on the Semantic Web," *IJoL*, vol. 3, no. 1, p. 67, Jul. 2018, doi: 10.23974/ijol.2018.vol3.1.64.

[57] G. Fu, "FCA based ontology development for data integration," *Information Processing & Management*, vol. 52, no. 5, pp. 765–782, Sep. 2016, doi: 10.1016/j.ipm.2016.02.003.

[58] M. Frtunic Gligorijevic, M. Bogdanovic, N. Veljkovic, and L. Stoimenov, "Open Data Categorization Based on Formal Concept Analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 2, pp. 571–581, Apr. 2021, doi: 10.1109/TETC.2019.2919330.

[59] K. Sumangali and C. A. Kumar, "A comprehensive overview on the foundations of formal concept analysis," *Knowledge Management & E-Learning: An International Journal*, vol. 9, no. 4, pp. 512–538, Dec. 2018.