An Intelligent Data Integration Approach Using Schema Matching, Entity Resolution, and Semantic Technologies

DHANAVATH NAGARAJU,
Research scholar, Department of Computer Science and Engineering, Dr.A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh

sonydhanavath07@gmail.com

SUPERVISOR:
Dr. MANISH GAUR
PROFESSOR
COMPUTER SCIENCE AND ENGINEERING
Dr.A.P. J Abdul Kalam Technical University
Lucknow, Uttar Pradesh
drmanishgauraktu@gmail.com

CO-SUPERVISOR:
Dr.D. SRINIVAS REDDY
PROFESSOR
COMPUTER SCIENCE AND ENGINEERING
VAAGESWARI COLLEGE OF ENGINEERING
KARIMNAGAR, TELANGANA
srinivasreddydhava@gmail.com

***Abstract*— The rapid expansion of heterogeneous big data has created significant challenges for organizations attempting to integrate information generated from diverse data sources. Modern enterprise environments continuously produce structured, semi-structured, and unstructured data through relational databases, cloud platforms, Internet of Things (IoT) devices, enterprise applications, web services, and social media networks. Traditional data integration techniques primarily rely on manually defined transformation rules and static schema mappings, making them insufficient for handling the complexity and dynamic nature of heterogeneous big data. This paper presents a conceptual intelligent data integration approach that incorporates data preprocessing, schema matching, entity resolution, semantic integration, and metadata management to improve data quality, interoperability, and integration efficiency. The proposed approach emphasizes intelligent automation to reduce manual intervention while enhancing consistency among heterogeneous datasets. Data preprocessing improves information quality by eliminating duplicate records, correcting inconsistencies, handling missing values, and standardizing data formats. Schema matching and entity resolution establish relationships among heterogeneous datasets and identify duplicate entities, whereas semantic integration and metadata management resolve conceptual conflicts and maintain consistent descriptions of integrated information. The proposed approach provides a comprehensive conceptual methodology capable of supporting business intelligence, predictive analytics, decision support, healthcare, finance, scientific research, and other data-intensive applications. The conceptual analysis indicates that integrating intelligent techniques throughout the data integration lifecycle can significantly improve the quality, reliability, and usability of heterogeneous data in modern big data environments.**

***Keywords*— Data Integration, Schema Matching, Entity Resolution, Semantic Integration, Metadata Management, Data Preprocessing, Big Data**

I. INTRODUCTION

The continuous growth of digital transformation has significantly increased the amount of data generated by modern organizations. Every day, enterprise applications, relational databases, cloud computing platforms, Internet of Things (IoT) devices, mobile applications, web services, and social media networks generate enormous volumes of structured, semi-structured, and unstructured data. These heterogeneous datasets contain valuable information that can support strategic planning, predictive analytics, business intelligence, healthcare management, scientific research, and financial analysis. However, integrating these diverse datasets into a unified and meaningful repository remains one of the most challenging tasks in modern big data environments.

Heterogeneous data integration involves combining information originating from multiple independent systems that differ in structure, storage format, schema design, metadata, communication protocols, and semantic representation. Since organizations frequently utilize different technologies for storing and processing information, similar data may be represented using different attribute names, database structures, coding standards, or conceptual definitions. These differences create inconsistencies that reduce interoperability and complicate analytical processing.

Traditional data integration techniques generally rely on Extract–Transform–Load (ETL) processes, manually defined schema mappings, and predefined

transformation rules. Although these approaches have been widely adopted in conventional relational database environments, they often struggle to integrate rapidly evolving heterogeneous datasets generated from cloud platforms, IoT devices, distributed systems, and web services. As data sources continue to increase in both number and complexity, manual integration becomes increasingly expensive, time-consuming, and difficult to maintain.

To address these challenges, researchers have introduced intelligent techniques capable of automating different stages of the data integration process. Data preprocessing improves the quality of collected information by removing duplicate records, correcting inconsistencies, handling missing values, and standardizing heterogeneous formats. Schema matching automatically identifies corresponding attributes among independent databases, while entity resolution detects duplicate records representing identical real-world entities. Semantic integration establishes meaningful relationships among heterogeneous concepts, and metadata management maintains standardized descriptions that improve interoperability and data governance.

This paper proposes a conceptual intelligent data integration approach that combines these intelligent techniques into a unified methodology for improving heterogeneous data integration. Rather than focusing on software implementation, the paper presents the conceptual workflow and theoretical advantages of incorporating preprocessing, schema matching, entity resolution, semantic technologies, and metadata management into modern data integration systems. The proposed approach aims to improve data quality, interoperability, integration efficiency, and decision support capabilities across heterogeneous big data environments.

II. RELATED WORK

Rahm and Bernstein [1] presented one of the most comprehensive surveys on automatic schema matching techniques for heterogeneous database integration. Their research classified schema matching approaches into schema-level, instance-level, linguistic, structural, and hybrid methods for identifying corresponding attributes across heterogeneous databases. The authors emphasized that automatic schema matching significantly reduces manual effort while improving the consistency of integrated datasets. They also discussed various similarity measures used to compare attribute names, metadata, and database structures. Although the survey provided a strong theoretical foundation for schema matching, it mainly focused on traditional database environments and did not consider modern heterogeneous big data generated from cloud platforms, IoT devices, and distributed systems. Their work remains one of the fundamental references for intelligent schema matching in heterogeneous data integration.

Elmagarmid et al. [2] investigated duplicate record detection and entity resolution techniques for improving data quality in heterogeneous databases. The study reviewed deterministic, probabilistic, and machine learning-based approaches for identifying duplicate records representing the same real-world entity. The authors demonstrated that effective entity resolution significantly improves the accuracy and consistency of integrated datasets by eliminating redundant information before analytical processing. They also highlighted several challenges associated with noisy, incomplete, and inconsistent data collected from multiple independent sources. The research concluded that intelligent entity resolution plays an essential role in modern data integration systems by improving data reliability and supporting accurate organizational decision-making.

Getoor and Machanavajjhala [3] presented an extensive review of entity resolution methodologies and discussed both theoretical foundations and practical applications. Their research examined various similarity measures, probabilistic models, clustering techniques, and machine learning algorithms used for identifying duplicate entities across heterogeneous information systems. The authors emphasized that intelligent entity resolution improves interoperability by merging records representing identical real-world objects while preserving data consistency. The study also identified several open research challenges, including scalability, uncertainty management, and automated learning for heterogeneous environments. Their findings demonstrated that entity resolution remains a critical component of intelligent data integration systems.

Portisch et al. [4] conducted a comprehensive survey on ontology-based schema matching and semantic integration for heterogeneous information systems. Their research analyzed the use of background knowledge, ontologies, and semantic web technologies to improve automatic schema matching. The authors found that semantic information significantly improves matching accuracy by identifying conceptual relationships that cannot be detected through structural similarity alone. The survey also discussed the limitations of existing ontology-based approaches, including ontology construction complexity and domain dependency. Their work highlighted the importance of semantic technologies for improving

interoperability and resolving semantic conflicts during heterogeneous data integration.

Masmoudi et al. [5] presented a comprehensive survey on semantic data integration and querying techniques for heterogeneous data environments. The authors reviewed materialized, virtual, and hybrid integration architectures while examining ontology-based integration, schema mapping, metadata management, and semantic query processing. The study concluded that metadata management and semantic technologies are fundamental for improving interoperability among heterogeneous information systems. Furthermore, the survey identified scalability, semantic consistency, and intelligent automation as important research challenges in modern big data environments. The authors suggested that future intelligent integration systems should combine semantic technologies, metadata management, and automated processing techniques to improve integration efficiency and support advanced analytical applications.

III. DATASET DETAILS

The proposed intelligent data integration approach is designed to process heterogeneous datasets collected from multiple independent information systems. Rather than relying on a single dataset, the conceptual methodology supports the integration of structured, semi-structured, and unstructured data commonly generated in modern enterprise environments. These datasets differ in schema design, storage format, attribute naming conventions, metadata descriptions, and semantic representations, making intelligent preprocessing and integration essential for achieving high-quality integrated information.

Structured datasets are primarily obtained from relational database management systems such as MySQL, PostgreSQL, Oracle Database, and Microsoft SQL Server. These datasets consist of well-defined tables containing rows and columns organized according to predefined schemas. Examples include customer information, employee records, financial transactions, inventory management systems, healthcare records, banking applications, and enterprise resource planning (ERP) databases. Although structured data provide high consistency, integrating multiple relational databases often requires intelligent schema matching because different organizations may represent similar information using different table structures and attribute names.

Semi-structured datasets represent another important source of heterogeneous information considered in the proposed approach. These datasets commonly appear in XML, JSON, CSV, and YAML formats generated through web services, REST APIs, cloud applications, and mobile platforms. Unlike relational databases, semi-structured data do not follow fixed schemas but instead use hierarchical structures that allow greater flexibility for representing diverse information. However, differences in document organization, metadata definitions, and hierarchical relationships create additional challenges during schema matching and semantic integration.

Unstructured datasets form the largest category of information in modern big data environments. These datasets include documents, emails, PDF files, web pages, social media posts, images, videos, audio recordings, and system log files. Since unstructured data lack predefined schemas, intelligent preprocessing techniques such as text normalization, duplicate detection, metadata extraction, and semantic analysis become essential before meaningful integration can be achieved. The proposed approach conceptually supports preprocessing and semantic interpretation of unstructured information to improve integration quality and interoperability.

The conceptual approach also considers datasets generated from Internet of Things (IoT) devices and cloud computing platforms. IoT sensors continuously generate high-velocity streams of environmental measurements, healthcare monitoring data, industrial sensor readings, transportation information, and smart city observations. Cloud platforms simultaneously store large volumes of distributed enterprise data, application logs, multimedia files, and transactional information. These heterogeneous datasets require intelligent preprocessing and metadata management to ensure consistent integration across distributed computing environments.

For conceptual evaluation, the proposed methodology assumes that heterogeneous datasets contain common data quality issues encountered in real-world applications, including duplicate records, missing values, inconsistent attribute names, varying data formats, semantic conflicts, and incomplete metadata. The intelligent integration techniques described in this paper—including data preprocessing, schema matching, entity resolution, semantic integration, and metadata management—are conceptually applied to address these challenges before integrating the data into a unified repository. This approach is expected to improve data quality, interoperability, and integration efficiency while providing reliable information for business intelligence, predictive analytics, and organizational decision-making.

Dataset Type	Typical Sources	Common Integration Challenges	Intelligent Technique Applied
Structured Data	MySQL, Oracle, PostgreSQL, SQL Server	Schema differences, duplicate records	Schema Matching, Entity Resolution
Semi-Structured Data	XML, JSON, REST APIs, CSV	Variable structure, metadata inconsistency	Schema Matching, Metadata Management
Unstructured Data	Documents, Emails, Images, Social Media	No fixed schema, semantic ambiguity	Preprocessing, Semantic Integration
IoT Data	Smart Sensors, Wearable Devices	High-velocity data, inconsistent formats	Data Preprocessing, Metadata Management
Cloud Data	AWS, Azure, Google Cloud	Distributed storage, interoperability	Semantic Integration, Metadata Management
Enterprise Data	ERP, CRM, Banking Systems	Data redundancy, inconsistent terminology	Entity Resolution, Schema Matching

Table 1. Characteristics of Heterogeneous Datasets

IV. PROPOSED METHODOLOGY

4.1 Overview

The proposed methodology presents a conceptual intelligent data integration approach designed to improve the quality, consistency, and interoperability of heterogeneous datasets. Modern organizations generate structured, semi-structured, and unstructured data from multiple independent sources, including relational databases, cloud platforms, Internet of Things (IoT) devices, enterprise applications, web services, and social media. Since these datasets differ in structure, format, schema definitions, and semantic representation, integrating them using conventional methods often results in inconsistencies, duplicate records, and semantic conflicts.

To address these challenges, the proposed methodology incorporates five intelligent integration techniques: data preprocessing, schema matching, entity resolution, semantic integration, and metadata management. These techniques work together in a sequential workflow to improve data quality before storing the integrated information in a unified repository. The methodology minimizes manual intervention, enhances interoperability among heterogeneous systems, and provides reliable data for business intelligence, analytics, and decision support.

4.2 Intelligent Data Preprocessing

Data preprocessing is the first stage of the proposed methodology and focuses on improving the quality of heterogeneous datasets before integration. Data collected from multiple sources frequently contain missing values, duplicate records, inconsistent attribute formats, invalid entries, and noisy information. If these issues are not addressed during the initial stage, they can significantly reduce the effectiveness of subsequent integration processes.

The preprocessing stage conceptually performs data cleaning, normalization, validation, duplicate detection, missing value handling, and standardization of data formats. Attribute names, date formats, numerical values, and textual information are transformed into consistent representations to improve compatibility across heterogeneous systems. By eliminating inconsistencies at an early stage, preprocessing establishes a reliable foundation for schema matching and semantic integration while improving the overall quality of integrated data.

4.3 Intelligent Schema Matching

After preprocessing, the methodology performs intelligent schema matching to identify relationships among heterogeneous database schemas and document structures. Since different organizations often represent similar information using different attribute names and structural designs, automatic identification of corresponding attributes is essential for successful data integration.

The conceptual schema matching process compares attribute names, data types, metadata descriptions, structural relationships, and semantic similarities to determine equivalent fields across independent datasets. Intelligent matching techniques reduce the need for manual mapping and improve integration accuracy by identifying meaningful correspondences among heterogeneous schemas. This process enables datasets originating from different platforms to be integrated into a consistent conceptual model.

4.4 Entity Resolution

Entity resolution is performed after schema matching to identify duplicate records that represent the same real-world object or entity. Multiple heterogeneous datasets frequently contain redundant records because identical entities may be stored in different systems using slightly different names, identifiers, addresses, or descriptive information.

The proposed methodology conceptually applies similarity comparison techniques to detect duplicate entities and merge related records into unified representations. Entity resolution improves data consistency by eliminating redundancy while preserving complete information from multiple sources. This process reduces storage duplication, enhances analytical accuracy, and provides more reliable information for organizational decision-making.

4.5 Semantic Integration

Semantic integration resolves conceptual differences that cannot be identified through structural comparisons alone. Even when datasets contain similar information, differences in terminology, business rules, attribute definitions, and domain-specific vocabulary may prevent successful integration.

The proposed methodology conceptually incorporates semantic technologies to establish meaningful relationships among heterogeneous concepts. Ontology-based knowledge representation and semantic mapping techniques enable the identification of equivalent concepts despite differences in naming conventions or organizational standards. Semantic integration therefore improves interoperability by ensuring that integrated datasets preserve consistent meanings across multiple heterogeneous sources.

4.6 Metadata Management

Metadata management provides standardized descriptions for all integrated datasets throughout the integration lifecycle. Metadata include information about data sources, schema definitions, attribute descriptions, ownership, data quality indicators, timestamps, and transformation history.

The proposed methodology conceptually organizes metadata within a centralized metadata repository that supports data governance, consistency, traceability, and efficient information retrieval. Metadata management simplifies future integration activities by maintaining complete documentation of integrated datasets and enabling users to understand the origin, structure, and meaning of stored information. Effective metadata management also improves interoperability and facilitates long-term maintenance of heterogeneous data repositories.

4.7 Integrated Workflow

The intelligent data integration process begins by collecting heterogeneous datasets from multiple independent sources. These datasets first undergo intelligent preprocessing to improve their quality through cleaning, normalization, validation, and duplicate removal. The cleaned datasets are then processed using intelligent schema matching to identify corresponding attributes among different schemas. Entity resolution follows by detecting duplicate records representing identical real-world entities and merging them into unified representations. Semantic integration resolves conceptual conflicts and establishes meaningful relationships among heterogeneous information using semantic mapping techniques. Finally, metadata management records the structural and semantic descriptions of the integrated datasets before storing them in a centralized repository. The integrated repository supports business intelligence, predictive analytics, reporting, visualization, and organizational decision-making by providing accurate, consistent, and interoperable information.

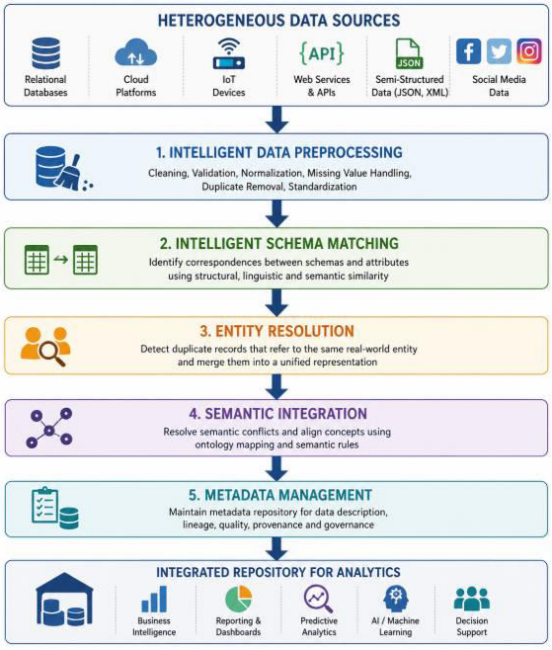

Figure [1]: Proposed Intelligent Data Integration Methodology

V. RESULT AND DISCUSSION

The proposed intelligent data integration approach was conceptually evaluated based on its ability to improve data quality, interoperability, and integration efficiency in heterogeneous big data environments. The evaluation focuses on the effectiveness of the intelligent techniques

incorporated into the methodology, including data preprocessing, schema matching, entity resolution, semantic integration, and metadata management. These techniques are expected to address the common challenges associated with integrating heterogeneous datasets generated from multiple independent sources such as relational databases, cloud platforms, IoT devices, web services, and enterprise applications.

The first stage of the proposed methodology, intelligent data preprocessing, is expected to significantly improve the quality of heterogeneous datasets before integration. Real-world datasets often contain missing values, duplicate records, inconsistent formats, invalid entries, and noisy information that reduce integration accuracy. By performing data cleaning, normalization, validation, and standardization, the preprocessing stage conceptually produces more reliable and consistent datasets. This improvement in data quality provides a strong foundation for subsequent integration processes and minimizes the propagation of errors throughout the workflow.

The schema matching component conceptually enhances interoperability by automatically identifying corresponding attributes among heterogeneous schemas. Different information systems frequently represent similar data using different attribute names, structures, or formats. Intelligent schema matching reduces manual mapping efforts by establishing meaningful relationships between heterogeneous schemas based on structural, linguistic, and semantic similarities. As a result, data originating from multiple independent systems can be integrated more efficiently into a unified representation.

The proposed methodology also incorporates entity resolution to improve the consistency of integrated data. Duplicate records representing the same real-world entity are common in heterogeneous environments where information is collected from multiple databases. Through intelligent similarity comparison and record merging, entity resolution conceptually eliminates redundant information while preserving complete and accurate entity descriptions. This process improves data consistency, reduces redundancy, and supports more reliable analytical processing.

Another significant contribution of the proposed methodology is the application of semantic integration to resolve conceptual conflicts among heterogeneous datasets. Structural matching alone is often insufficient because different organizations may use different terminology to describe identical concepts. Semantic integration establishes meaningful relationships among heterogeneous concepts using ontology-based mapping and semantic rules. This conceptual process improves interoperability by ensuring that integrated datasets maintain consistent meanings regardless of their original representations.

Metadata management further enhances the effectiveness of the proposed methodology by maintaining standardized descriptions of integrated datasets. Metadata provide information regarding data origin, schema definitions, attribute descriptions, transformation history, ownership, and quality indicators. Centralized metadata management improves traceability, governance, and future maintenance while supporting efficient information retrieval and integration.

Overall, the conceptual analysis indicates that combining these intelligent techniques within a unified methodology provides considerable advantages over conventional data integration approaches. The methodology is expected to improve data quality, reduce manual intervention, enhance interoperability, eliminate redundant records, resolve semantic inconsistencies, and support scalable integration of heterogeneous datasets. These improvements provide a reliable foundation for business intelligence, predictive analytics, artificial intelligence applications, and organizational decision-making.

Intelligent Technique	Purpose	Expected Outcome
Data Preprocessing	Clean and standardize heterogeneous datasets	Improved data quality and consistency
Schema Matching	Identify corresponding attributes across schemas	Better interoperability and reduced manual mapping
Entity Resolution	Detect and merge duplicate records	Reduced redundancy and improved data accuracy
Semantic Integration	Resolve semantic conflicts among heterogeneous data	Enhanced semantic consistency and interoperability
Metadata Management	Maintain standardized dataset descriptions	Improved governance, traceability, and data management

Table [1]: Conceptual Evaluation of Intelligent Integration Techniques

DISCUSSION

The proposed intelligent data integration approach demonstrates the importance of incorporating automation throughout the data integration lifecycle. Unlike conventional integration methods that rely heavily on manual schema mapping and predefined transformation rules, the proposed methodology combines intelligent preprocessing, schema matching, entity resolution, semantic integration, and metadata management into a coordinated workflow. Each component contributes to improving the overall quality and usability of integrated information.

Among the five intelligent techniques, data preprocessing serves as the foundation by ensuring that heterogeneous datasets are accurate and consistent before integration. Schema matching and entity resolution automate the identification of corresponding attributes and duplicate entities, reducing manual effort and improving integration accuracy. Semantic integration addresses conceptual inconsistencies that cannot be resolved through structural comparisons alone, while metadata management supports long-term governance, traceability, and interoperability.

Although this study presents a conceptual methodology without experimental implementation, the theoretical evaluation suggests that integrating intelligent techniques throughout the data integration process can significantly improve integration efficiency and information quality. The proposed approach is applicable to healthcare, finance, manufacturing, education, smart cities, e-commerce, cloud computing, and other data-intensive domains where heterogeneous information must be integrated to support analytics and decision-making.

VI. CONCLUSION

The increasing volume and diversity of data generated from modern information systems have made intelligent data integration an essential requirement for organizations operating in heterogeneous big data environments. Data collected from relational databases, cloud platforms, Internet of Things (IoT) devices, enterprise applications, web services, and social media differ in structure, format, schema design, and semantic representation, making their integration a complex task. Traditional integration approaches primarily rely on manual schema mapping and predefined transformation rules, which are often insufficient for handling the complexity, scalability, and dynamic nature of heterogeneous datasets. Therefore, incorporating intelligent techniques into the data integration process has become increasingly important for improving data quality, interoperability, and integration efficiency.

This paper presented a conceptual intelligent data integration approach that integrates five key intelligent techniques: data preprocessing, schema matching, entity resolution, semantic integration, and metadata management. These techniques collectively address major challenges encountered during heterogeneous data integration, including inconsistent data formats, duplicate records, schema differences, semantic conflicts, and incomplete metadata. By organizing these techniques into a unified conceptual methodology, the proposed approach aims to automate critical integration tasks while reducing manual intervention and improving the overall reliability of integrated information.

Among the proposed techniques, intelligent data preprocessing provides the foundation for successful integration by improving the quality of heterogeneous datasets through data cleaning, normalization, validation, missing value handling, duplicate removal, and standardization. Schema matching establishes meaningful correspondences between heterogeneous database schemas, while entity resolution identifies and merges duplicate records representing the same real-world entities. Furthermore, semantic integration enhances interoperability by resolving conceptual differences through semantic mapping, and metadata management ensures standardized documentation, governance, and traceability throughout the integration process.

The conceptual analysis indicates that combining these intelligent techniques can significantly improve the effectiveness of heterogeneous data integration. The proposed methodology is expected to enhance data quality, reduce redundancy, improve semantic consistency, strengthen interoperability, and support efficient information sharing across multiple heterogeneous systems. These improvements provide a reliable foundation for advanced applications such as business intelligence, predictive analytics, artificial intelligence, reporting, and organizational decision-making.

Although the proposed approach has been presented as a conceptual methodology without experimental implementation, it establishes a comprehensive theoretical foundation for intelligent heterogeneous data integration. The methodology is flexible enough to support various application domains, including healthcare, finance, education, manufacturing, smart cities, scientific research, cloud computing, and e-commerce, where heterogeneous information must be integrated efficiently for analytical purposes.

In conclusion, the proposed Intelligent Data Integration Approach Using Schema Matching,

Entity Resolution, and Semantic Technologies provides a systematic and scalable methodology for improving heterogeneous data integration in modern big data environments. By incorporating intelligent preprocessing, automated schema matching, entity resolution, semantic integration, and metadata management into a unified workflow, the approach contributes to improving data quality, interoperability, and integration efficiency. Future research can extend this conceptual work by implementing the proposed methodology on real heterogeneous datasets, evaluating its performance using distributed computing platforms such as Apache Hadoop and Apache Spark, and integrating advanced artificial intelligence and machine learning techniques to further enhance automation, scalability, and intelligent decision support.

REFERENCES

[1] E. Rahm and P. A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," The VLDB Journal, vol. 10, no. 4, pp. 334–350, 2001.

[2] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, Jan. 2007.

[3] L. Getoor and A. Machanavajjhala, "Entity Resolution: Theory, Practice and Open Challenges," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2018–2019, Aug. 2012.

[4] J. Portisch, M. Hladik, and H. Paulheim, "Background Knowledge in Ontology Matching: A Survey," Semantic Web, vol. 13, no. 1, pp. 117–158, 2022.

[5] M. Masmoudi, S. Ben Abdallah Ben Lamine, M. H. Karray, B. Archimede, and H. Ben Ghezala, "Semantic Data Integration and Querying: A Survey and Challenges," ACM Computing Surveys, vol. 56, no. 8, Article 204, 2024.

[6] E. Rahm and P. A. Bernstein, “A Survey of Approaches to Automatic Schema Matching,” The VLDB Journal, vol. 10, no. 4, pp. 334–350, Dec. 2001.

[7] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate Record Detection: A Survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, Jan. 2007.

[8] L. Getoor and A. Machanavajjhala, “Entity Resolution: Theory, Practice and Open Challenges,” Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2018–2019, Aug. 2012.

[9] J. Euzenat and P. Shvaiko, Ontology Matching, 2nd ed. Berlin, Germany: Springer, 2013.

[10] J. Portisch, M. Hladik, and H. Paulheim, “Background Knowledge in Ontology Matching: A Survey,” Semantic Web, vol. 13, no. 1, pp. 117–158, 2022.

[11] M. Masmoudi, S. Ben Abdallah Ben Lamine, M. H. Karray, B. Archimede, and H. Ben Ghezala, “Semantic Data Integration and Querying: A Survey and Challenges,” ACM Computing Surveys, vol. 56, no. 8, Article 204, 2024.

[12] X. L. Dong and D. Srivastava, “Big Data Integration,” Proceedings of the VLDB Endowment, vol. 6, no. 11, pp. 1188–1189, Aug. 2013.

[13] A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration. Burlington, MA, USA: Morgan Kaufmann, 2012.

[14] A. Y. Halevy, A. Rajaraman, and J. Ordille, “Data Integration: The Teenage Years,” Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 9–16, 2006.

[15] P. A. Bernstein and S. Melnik, “Model Management 2.0: Manipulating Richer Mappings,” Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1–12, 2007.

[16] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 1st ed. San Rafael, CA, USA: Morgan & Claypool, 2011.

[17] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data – The Story So Far,” International Journal on Semantic Web and Information Systems, vol. 5, no. 3, pp. 1–22, 2009.

[18] N. F. Noy, “Semantic Integration: A Survey of Ontology-Based Approaches,” ACM SIGMOD Record, vol. 33, no. 4, pp. 65–70, Dec. 2004.

[19] M. Lenzerini, “Data Integration: A Theoretical Perspective,” Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), pp. 233–246, 2002.

[20] D. K. W. Chiu and P. C. K. Hung, “A Survey of Metadata Management in Distributed Information Systems,” Information Systems Frontiers, vol. 7, no. 2, pp. 145–158, 2005.

[21] A. Gangemi and V. Presutti, “Ontology Design Patterns,” in Handbook on Ontologies, 2nd ed., Berlin, Germany: Springer, 2009, pp. 221–243.

[22] P. Shvaiko and J. Euzenat, “A Survey of Schema-Based Matching Approaches,” Journal on Data Semantics, vol. 4, pp. 146–171, 2005.

[23] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, “Ontology-Based Integration of Information—A Survey of Existing Approaches,” Proceedings of the IJCAI Workshop on Ontologies and Information Sharing, pp. 108–117, 2001.

.