INTERNATIONAL JOURNAL OF ACTA INFORMATICA
VOLUME (2024)

METADATA-DRIVEN PRIVACY CLASSIFICATION OF SQL SERVER STORED PROCEDURES IN ENTERPRISE WEB APPLICATIONS USING TRANSFORMER-BASED MODELS

Divya Sai Jaladi[1*]
[1]Application Developer, SCDMV, Charlotte, NC, UNITED STATES

ABSTRACT

Metadata facilitates semantic interoperability, data reutilization, and effective information retrieval across diverse systems. Nonetheless, current research exhibits disjointed methodologies and fails to provide a unified perspective on the tactics employed across many fields. This fragmentation results in a deficiency in comprehending the present state of the art and recognizing cross-domain trends and issues. This systematic study seeks to fill this vacuum by examining metadata integration strategies across many disciplines, including Health and Medicine, Smart Cities and IoT, Data Science, Geosciences, Cultural Heritage, and Library and Information Science. This evaluation adheres to the Kitchenham framework, which includes the preparation, execution, and reporting stages. A total of 81 peer-reviewed articles published from 2014 to 2023 were selected from five major databases, adhering to established inclusion and exclusion criteria and a systematic quality evaluation method. The findings indicate a majority of ontology utilization, succeeded by metadata-driven language, standards, processes, and standardized metadata schemas, alongside a burgeoning tendency towards automation via the implementation of AI-based methodologies. The recognized limitations encompass semantic heterogeneity, insufficient standardization, restricted automation, and usability concerns in existing tools and systems. We offer a thorough synthesis of current methodologies, emphasizing both domain-specific and cross-domain trends, while pinpointing research opportunities to enhance metadata integration via automation, semantic enrichment, and standardized procedures.

KEYWORDS: Metadata Integration, Interoperability, Metadata Mapping, Semantic Enhancement.

INTRODUCTION

The increasing volume and diversity of data across several areas has heightened the necessity for efficient data integration solutions, allowing information from disparate sources to be accessible, analyzed, and repurposed in a coherent and significant manner. In this context, metadata, often defined as data about data, provides descriptive, structural, and semantic information that enhances the comprehension, organization, and interoperability of data across diverse systems and platforms.

Consequently, metadata integration has become a crucial method for guaranteeing semantic compatibility and enhancing data utility in distributed architectures, including cloud-based settings, IoT, and healthcare information systems. Integration techniques facilitate data discovery, reuse, and governance by establishing linkages among various metadata standards and repositories.

Nonetheless, despite its significance, metadata integration continues to pose a formidable challenge owing to both technological and semantic obstacles. This encompasses schema heterogeneity, the absence of standards in information representation, and the intricacies associated with aligning various ontologies and controlled vocabularies. Ontologies and vocabularies are frequently developed independently to address specific domain requirements. This results in disparate modeling methodologies, differing degrees of specificity, and the use of several representation standards, including the Web Ontology Language (OWL), Resource Description Framework (RDF), and Simple Knowledge Organization System (SKOS). Thus, aligning these resources frequently necessitates advanced ontology matching algorithms, semantic reasoning abilities, and occasionally expert curation to guarantee meaningful integration [2].

A diverse array of strategies has been suggested to tackle these difficulties. This encompasses, but is not restricted to, ontology-driven frameworks [3], [4], and federated metadata.

registries, rule-based transformations, machine learning-based entity matching, semantic mappings, schema alignment approaches, utilization of restricted vocabularies, and metadata standardization protocols. This variability exemplifies the varied nature of metadata integration and its adjustment to different domain requirements and technology circumstances. Nevertheless, current contributions are dispersed across various disciplines, and a comprehensive knowledge of their methodologies, uses, and constraints remains absent.

This article provides a systematic literature review (SLR) on metadata integration, aiming to identify key strategies, methods, and techniques; analyze application domains and their specific requirements; and map significant contributions, challenges, gaps, and future directions documented in the literature. This study's primary contributions include a thorough classification of metadata integration strategies, methods, and techniques; the recognition of emerging trends and consistent patterns across diverse application domains; and the provision of a consolidated resource to assist researchers and specialists involved in data and metadata integration efforts.

Background Information Regarding Metadata Integration

This section delineates the essential definitions and vocabulary.

Summary of the primary categories of metadata integration. We additionally provide a synopsis of pertinent research on this subject.

Introduction to Metadata

The terminology for metadata integration in the literature is inconsistent, with phrases like harmonization and standardization sometimes used interchangeably. In [20], metadata integration denotes the amalgamation and standardization of metadata from several sources to generate a cohesive data set. In [3], standardization was characterized as a method for enhancement.

Interoperability of metadata. In [14], harmonization is defined as the establishment of a cohesive metadata model, whereas [15] describes it as the process of guaranteeing both structural and syntactic interoperability among resources, exemplified by the consistent application of shared metadata properties (e.g., the language property from Dublin Core), as well as semantic alignment across datasets. While these terms have been employed, we assert that metadata integration necessitates uniformity.

Classifications of Metadata Integration

The understanding of metadata integration differs among the research. Certain authors have characterized the alignment of data types, encoding formats, and domain constraints as syntactic integration across heterogeneous systems [21], [22], [23]. From this viewpoint, integration entails the conversion or standardization of metadata representations, such as the transformation of XML Schema or JSON Schema specifications into a unified format (e.g., RDF), enabling uniform interpretation of attributes across repositories. Figure 1 depicts a standard metadata representation derived from [22]. Simultaneously, other studies delineate a structural perspective on information integration, emphasizing the organization and interrelation of metadata pieces inside their schemas. Integration encompasses not only the standardization of syntactic representations but also the alignment of models and conceptual frameworks [5], [24]. This structural viewpoint frequently entails recognizing the parallels between entities and attributes, addressing schema variability, and aligning analogous constructs.

Despite the standardization of field names and data formats, inconsistencies in interpretation may still occur. Attributes that are structurally distinct may denote the same concept across many systems. For instance, data file A use the field full_name, while data file B employs the term name, both denoting an individual's complete name. Despite the syntactic alignment of these variables post-standardization, semantic integration is essential for a uniform interpretation. Semantic integration utilizes ontologies that formally delineate concepts and their interrelations to solve such

scenarios. Utilizing OWL, a mediator can assert equivalence between attributes and individuals, thereby facilitating interoperability.

capability across diverse sources. Take into account the subsequent information:

full_name ≡ name full_name(JoaoRecordA, 'João Silva') name(JoaoRecordB, 'João Silva') JoaoRecordA ≡ JoaoRecordB

In this instance, the attributes:full_name and:name is aligned using owl:equivalentProperty, and the individuals :JoaoRecordA and :JoaoRecordB are connected using owl:sameAs. Such semantic mappings enable the unification of theoretically equivalent data representations, hence enhancing integration efforts beyond mere structural compatibility.

Consequently, metadata integration may commence with the syntactic uniformity of formats and types, but frequently progresses to structural integration and, eventually, semantic integration.

Connection to Previous Systematic Reviews

This study integrated two prior systematic reviews, [16] and [26], both concentrating on metadata practices within the medical field. The review in [16] performed a scoping analysis of metadata-driven methodologies for data harmonization in healthcare, emphasizing technological infrastructure, interoperability frameworks, and policy-related factors. Conversely, [26] examined the conceptual underpinnings of metadata through the analysis of definitions, classifications, and application domains in biomedical informatics.

This review diverges in significant respects. Initially, it expands the reach beyond healthcare to include many fields such as cultural heritage, geosciences, smart cities, and digital libraries. Secondly, it classifies the chosen research by application domain, facilitating cross-domain comparisons. The study is organized around well defined research issues, facilitating the discussion of integration methodologies, obstacles, and the application of semantic enrichment in various contexts.

This review enhances the analytical rigor and scope of prior studies, contributing to a more thorough comprehension of metadata integration advancement in various disciplines. It consolidates repeating methodological trends into a framework for metadata integration. This roadmap delineates the standard procedures of integration, matching, mapping, transformation, fusion, and enrichment across several domains, providing a systematic and domain-neutral approach.

Research Methodology

This systematic review adhered to Kitchenham's recommendations [27] and was organized into three phases: planning, conducting, and reporting. During the planning phase, we established the study topic, search string, information bases, and criteria for

inclusion and exclusion. In the conducting phase, we screened the articles, eliminated duplicates, and implemented the criteria.

All supplemental materials, encompassing filtered datasets, classification spreadsheets, quality assessment scores, and intermediate synthesis files, are accessible in a designated repository to guarantee repeatability [28].

Research Inquiries

The subsequent research questions were established:

1) RQ1: What strategies, methodologies, and approaches are employed for metadata integration?

2) RQ2: What are the contributions and problems associated with metadata integration across various domains?

The search technique was devised to utilize five databases pertinent to both computing and health sectors. The databases utilized in this review were as follows:

• Association for Computing Machinery

• Institute of Electrical and Electronics Engineers

• PubMed

• Scopus

• Web of Science

Table 1 presents the search string that adheres to a logical framework of (title OR abstract) AND author keywords. The title must incorporate the phrase metadata. The term harmonization, a synonym for integration, was incorporated and may or may not be present in the title. Supplementary conditions like Framework, platform, and architecture were integrated to expand the search parameters.

Selection of Studies

A total of 2,375 articles were obtained, of which 885 were duplicates. In the screening procedure, utilizing title and abstract examination alongside the inclusion and exclusion criteria depicted in Figure 2, 1,249 publications were discarded, while 241 were considered suitable for full-text evaluation.

Subsequent to an extra screening phase, which involved a comprehensive examination of titles, abstracts, conclusions, selected text excerpts, and figures, alongside the application of the quality assessment criteria specified in Table 2, a final compilation of 81 articles was chosen for full-text analysis.

Conversely, the inclusion criteria (IC) emphasized the selection of peer-reviewed papers (IC1) that specifically addressed metadata integration and offered practical or applied examples (IC2). The inclusion criterion (IC3) limited the review to papers

published from 2014 to 2023, preserving the designated ten-year span but integrating recent research to reflect current advancements and trends in the area.

This study seeks to consolidate insights from a body of literature that is both thematically relevant to metadata integration and methodologically robust, reflecting the current state of the art. Figure 3 illustrates the classification of 81 selected publications by their publication year. The incorporation of works published till 2023 guarantees that the evaluation encompasses the latest advancements in metadata integration, mirroring current research trends and novel methodologies. This temporal filtering was crucial for preserving the study's relevance and offering a contemporary review of methodological and technological advancements in the field.

Data Integration

The reporting phase depicted in Figure 2 integrates the outcomes of the comprehensive review process, linking study selection to analysis, debate, and final findings. Each paper was analyzed concerning the research questions, facilitating a systematic synthesis of the methodologies, problems, and contributions to metadata integration. During the analytical phase, the categorization of publications was particularly hard, since numerous research covered multiple techniques or overlapped across other fields. The ambiguities were rigorously analyzed and addressed in the Results and Discussion sections to guarantee conceptual clarity and analytical coherence. This thorough report addresses its guiding questions and offers a clear description of the interpretative decisions taken during the process.

The contributions of each study were categorized based on their application domains, including Health and Medicine (HM), Smart Cities and IoT (SCIoT), Data Science (DS), Culture and Digital Heritage (CDH), Information Science and Libraries (ISL), and Geosciences and Environment (GE). This facilitated a cross-domain investigation of the integration methodologies and contextual difficulties pertinent to each domain.

The issues identified in the investigations were consolidated to reveal similar obstacles, including metadata heterogeneity, absence of standards, and contextual mediators. Research possibilities were similarly classified to emphasize new trends, such as the application of AI for automated enrichment, the adoption of findable, accessible, interoperable, reusable (FAIR) principles, and federated metadata management.

Findings and Analysis

Ontologies

The analysis of the examined papers indicated that ontologies represent the primary method, utilized in 42 out of 81 articles (51.85%). Ontologies provide a formal and explicit delineation of shared concepts and relationships, acting as a semantic basis that facilitates consistent metadata alignment across varied schemas and diverse domains

that utilize varying terminologies and data formats. Ontology-based integration functions at a more profound semantic level, rather than depending exclusively on syntactic or structural similarities like schema matching or string comparison [35], [79]. Mediation eliminates ambiguity and inconsistency by offering a unified conceptual framework for mapping various information items.

Ontologies facilitate reasoning and inference procedures, enabling the discovery of latent linkages within metadata and the generation of new knowledge from integrated sources [35]. Furthermore, ontologies can be augmented and enhanced by associating information pieces with external resources, such as DBpedia or Wikidata, hence improving discoverability and providing contextual depth [14], [49].

Various methodologies may be employed for ontology-driven integration. Ontology mapping and alignment approaches establish correspondences between metadata schemas and ontology components, frequently augmented by artificial intelligence methods, including Natural Language Processing (NLP) for concept identification and classification, and machine learning for disambiguation tasks [56]. Mediation solutions employ a global ontology as a reference framework to standardize queries and maintain consistency across diverse datasets [62]. In situations marked by domain convergence, integration is typically supported by a reference ontology that provides a common conceptual framework and serves as a central unifying knowledge repository [35]. Linked Data practices facilitate the publication of metadata in RDF and its interconnection via ontologies, enhancing interoperability within the Semantic Web [49].

The development or modification of ontologies necessitates significant work and knowledge, whereas semantic heterogeneity, characterized by variations in scope, granularity, or conceptualization within ontologies, frequently complicates alignment [35].

Metadata-Driven Language, Standards, and Processes (MDLSP)

The second most prevalent strategy entails employing a metadata-driven language, standards, and processes (30 articles, 37.04%), which seeks to position metadata at the nucleus of system design, utilizing it not solely for descriptive functions but also to facilitate data integration, transformation, and retrieval processes [41], [42], [69]. MDLSP establishes standardized structures, semantics, and operational protocols.

In addition to technological harmonization, the MDLSP offers advantages in governance and transparency [44]. The metadata-centric approach, which documents all integration procedures, including provenance details and transformation specifications, enhances the reproducibility, auditability, and long-term preservation of integrated resources [36], [62]. It additionally facilitates sophisticated features including semantic querying, reasoning, and the development of knowledge graphs

[36], [45].

The Resource Description Framework (RDF) is a standard for encoding metadata in a machine-readable and interoperable format. RDF triples, or assertions, constitute connected graphs where subjects and objects serve as nodes and predicates function as edges [36], [53]. RDF organizes metadata on the Web, and its formal semantics and compatibility with ontologies establish a foundation for metadata-driven techniques, rendering integration both descriptive and interoperable as well as extendable [49], [53].

Standardized Metadata Schema

A prevalent metadata schema, found in 29 articles (35.08%), functioned as a cohesive framework that integrated local descriptions into a collective representation. In [22], the common metadata schema is characterized as a mediator, functioning as an intermediary layer that standardizes disparate sources by integrating them into a cohesive framework [10]. The provision of a unified representation paradigm facilitates the integration, retrieval, and reuse of metadata from diverse applications.

Numerous initiatives have exemplified this philosophy. For instance, in [4], the Dublin Core Metadata Initiative (DCMI) was employed to delineate bibliographic metadata standards, encompassing aspects such as title, creator, subject, description, and publisher. ISO/IEC 11179 delineates principles for metadata registries and promotes syntactic and semantic interoperability; for instance, the caDSR (Cancer Data Standards Registry and Repository) employs encoded public identifiers to ensure traceability via source information and to uphold version control with accessible historical versions [5]. In the healthcare sector, Health Level Seven – Fast Healthcare Interoperability Resources (HL7 FHIR) serves as a standardized framework to enable the sharing of clinical and biomedical data.

Nevertheless, the adoption of a unified metadata standard poses obstacles. The formulation and approval of a communal schema necessitate consensus among stakeholders, which is frequently challenging to attain. Moreover, integrating varied and domain-specific attributes into a unified schema may lead to a loss of granularity, as certain aspects might not conform to the standardized model. Continuous evolution is essential, as metadata schemas must adjust to emerging data kinds, technologies, and user demands.

Rule-Based Metadata Integration

Rule-based methodologies for metadata integration, documented in 27 papers (33.33%), depended on the explicit articulation of rules to align, convert, and integrate disparate metadata across sources. This approach depends on established integration rules that dictate the matching and reconciliation of traits, relationships, or data values.

For instance, in [68], extraction rules were incorporated into data model fields. These rules may represent equivalence, is-a relationships, or transformation requirements, so facilitating a regulated and comprehensible integration process.

A prevalent application of rule-based integration involves the utilization of constraint languages and reasoning mechanisms, such as the Semantic Web Rule Language (SWRL) or the Shapes Constraint Language (SHACL), which facilitate the formulation of rules that validate, transform, or infer new metadata statements. For instance, regulations may stipulate that two attributes from different schemas be comparable or that a specific value must be transformed to conform to the units or formats anticipated in the target schema [72]. The RDF Mapping Language (RML) illustrates a rule-based methodology for metadata integration by enabling the declarative definition of the transformation of diverse data sources into RDF [43]. In reality, it serves as a mapping language that encodes regulations for translating any format.

Converting structured or semi-structured data into a Resource Description Framework (RDF) [8].

Rule-based metadata integration can utilize first-order logic to formally articulate conditions and constraints, thereby enhancing the rigor and transparency of the integration process [59]. Moreover, heuristic rules may be utilized, generally structured as IF conditions, THEN decision, wherein domain expertise or experience delineates the conditions that direct the allocation of metadata categories or classes [46]. This strategy poses obstacles due to its demand for subject expertise and its time-consuming nature.

Schema Alignment

Schema matching methodologies, recognized in 20 studies (24.69%), pertain to the identification of correspondences among elements of disparate metadata schemas that delineate semantically analogous concepts, as defined by [85] and referenced in [26]. This approach facilitates the alignment of properties, relationships, and constraints among various data sources, thus promoting data interoperability and knowledge integration [7], [8].

Schema matching involves various complementary strategies that identify correspondences among schema items. String similarity algorithms assess the degree of resemblance between two textual labels, facilitating the identification of possible mappings between metadata attributes [49]. Common metrics encompass the Levenshtein distance, which determines the smallest number of insertions, deletions, or substitutions necessary to convert one string into another [17], [33], and the Jaccard coefficient, which evaluates the intersection between characters or tokens.

Sets derived from two strings [64]. While proficient in identifying syntactic correspondences, lexical similarity does not ensure semantic equivalence; homonyms may produce erroneous matches (e.g., Leonardo da Vinci signifying either the artist or the airport in Fiumicino) [56], whereas synonyms can obscure legitimate correspondences when different terms convey the same concept [53]. Thus, string similarity is frequently integrated with semantic methods to enhance alignment accuracy [46].

In addition to lexical similarity, schema matching employs structural and value-based strategies to identify more profound links among schema elements. Structural methodologies examine hierarchical and relational dependencies, including parent–child relationships, cardinality constraints, and domain–range compatibility [7], [9], [23], [36], [54], [83]. Path-based similarity takes into account the hierarchical context of an element to guarantee that correspondences align with schema organization [46], [52], [54]. Type alignment ensures datatype compatibility among schemas, while normalization addresses format discrepancies, including variations in date formats [46], [52], [58].

Value-oriented methodologies enhance structural techniques by analyzing instance-level data. Unit normalization transforms measurement units into standardized representations [12], [21], [61], while data value comparison employs IDs, controlled vocabularies, or date values to reveal equivalences despite variations in labels and structures [58]. Collectively, these techniques diminish ambiguity and improve the dependability of schema alignment, facilitating the mapping of local properties to a unified metadata schema.

Notwithstanding considerable advancements, schema matching continues to encounter obstacles. Terminological ambiguities, wherein identical phrases may possess disparate meanings across domains [32] or distinct names may represent the same concept [10], persist as a challenge. Moreover, schema matching frequently relies on the accessibility and caliber of schema descriptions, which are not always thoroughly or consistently articulated [6], [11], [26].

Linked Data Resources

Linked data found in 20 papers (24.69%) offered structured, interoperable, and semantically rich material suitable for reuse across several disciplines [11], [53]. They utilize uniform resource identities (URIs) to distinctly identify entities, relationships, and APIs, facilitating programmatic access and interoperability across platforms [11]. Commonly utilized resources encompass DBpedia, which derives structured information from Wikipedia; WordNet, which delineates terminological and semantic relationships among terms; Wikidata, a collaboratively assembled knowledge graph that consolidates cross-domain facts; and GeoNames, which offers comprehensive

geospatial metadata. Metadata integration systems can enhance descriptions, clarify ambiguities, and harmonize diverse vocabularies by associating local properties and identifiers with external references [31], [53], [56].

The presence of SPARQL endpoints significantly improves the integration process by enabling real-time queries of external knowledge bases and the retrieval of contextual information to augment local metadata [12]. This functionality facilitates activities including entity linkage, semantic enrichment, and ontology alignment, allowing local terms to be correlated with external ideas using recognized semantics [14], [36], [49]. Linking metadata to Linked Data resources enhances integration by improving interoperability, discoverability, and semantic accuracy [36]. Simultaneously, these resources present obstacles, such as the necessity to manage disparate levels of data quality, incompleteness, and dynamic schemas [14]. Nonetheless, the utilization of Linked Data has been implemented in semantic integration methodologies, allowing systems to augment local representations with global ones collective intelligence.

Regulated Lexicon

The controlled vocabularies identified in 19 articles (23.46%) were standardized and curated sets of terms designed to ensure precision in the description and organization of information [12]. By limiting language to a uniform lexicon, they reduced ambiguity, eliminated differences in phrasing, and improved interoperability among systems [31]. In contrast to free-text descriptions that might lead to discrepancies, controlled vocabularies offer a standardized reference that enhances the precision of data indexing, searching, and retrieval [42].

They may manifest in various forms, including basic taxonomies and thesauri, as well as more intricate ontologies [10], [18], [47]. Every structure offers hierarchical classification and semantic connections that facilitate advanced navigation and reasoning [9], [21], [72].

The Simple Knowledge Organization System (SKOS) is a prevalent method for publishing and managing vocabularies on the semantic web, encoding thesauri, taxonomies, and categorization schemes as RDF graphs [18], [47], [51], [53]. A SKOS-based vocabulary enables the unique identification of concepts through URIs and enhances them with semantic associations, including skos:broader, skos:narrower, and skos:related [5], [53]. This facilitates interoperability and enables sophisticated activities such as semantic search, query expansion, and hierarchical reasoning [5], [48]. In the realm of metadata integration, SKOS enables the alignment of local terminologies with standardized vocabularies, thus diminishing semantic heterogeneity and enhancing cross-domain discovery. Synonym detection tackles the issue of employing various names that denote the same or closely associated topics. By identifying and normalizing synonyms, systems enhance search precision and recall,

diminish redundancy, and streamline the mapping of characteristics and entities [5], [23]. SKOS facilitates this process by features like skos:altLabel and skos:hiddenLabel, which explicitly represent synonyms within controlled vocabularies, and skos:exactMatch, which creates equivalence across concepts across many vocabularies or ontologies [5], [18], [23], [48], [51].

Health and Medicine

A significant area of research in the HM domain has involved the creation of ontologies and collaborative platforms that facilitate semantic annotation, hence enhancing metadata integration, interoperability, and ultimately fostering a unified language and knowledge base [20], [69], [74]. A Bipolar Disorder Ontology was created to consolidate various knowledge, data, and metadata regarding illness susceptibility from different heterogeneous sources [79]. GenEpiO and FoodOn were created to offer contextual knowledge by associating data fields with their semantic significance in laboratory, clinical, and food safety environments, while also standardizing data elements across diverse sources and incorporating established food-related ontologies and databases [78].

Certain studies have highlighted the significance of metadata-driven methodologies via metadata repositories (MDRs) that consolidate attributes, transformation rules, and schema mappings, thereby diminishing manual labor [16], [19], [72], [73]. When integrated with rule-based systems and user-friendly interfaces, these repositories enhance accessibility and metadata integration in healthcare environments. Semantic annotation is facilitated in these repositories through the standardization of data element (DE) mapping to terminology systems, assuring internal semantic consistency and enabling integration across diverse datasets. As per [74], collaborative schema management across Distributed sites are supported, along with semantic labeling via controlled vocabularies.

Initiatives targeting standardization deficiencies in clinical imaging and digital pathology have employed FHIR healthcare data exchange resources and specialized pipelines to convert, pseudonymize, and transmit Digital Imaging and Communications in Medicine (DICOM) metadata, aligning clinical metadata with controlled vocabularies such as SNOMED-CT to improve semantic interoperability. Related contributions examine the translation of clinical study data from the CDISC Operational Data Model, a standard for representing clinical trial data, to FHIR, creating FHIR-based models that encapsulate metadata, thereby facilitating enhanced semantic enrichment and interoperability across observational studies [71].

Methods like federated semantic metadata registries enhance ISO/IEC 11179 by incorporating linked open data principles to facilitate interoperability in patient registries, while semantic data dictionaries and pragmatic metadata repositories provide

machine-readable, community-driven solutions that encourage the reuse, standardization, and traceability of clinical data elements. Automated methodologies, encompassing deep learning–driven mapping maintenance systems [6] and platforms like GEM, have enhanced scalability by employing text mining and machine learning classifiers to accurately discover metadata mappings [7].

RDF-based alignment algorithms facilitate context-aware mappings by creating correspondences between entities through structural similarity and the integration of semantic context, thereby addressing both structural and semantic heterogeneity in clinical and genomic data [20], [82]. Semantic Web technologies enhance metadata workflows by facilitating the collaborative application of shared vocabularies and ontology terms [20], while web-based tools like ODMedit underscore the necessity for standardized semantic annotations backed by resources such as the Unified Medical Language System (UMLS) [80]. Metadata-driven and ELT processes enhance the integration of real-world data in multisite clinical investigations, while ontology and rule-based systems permit automated conversions between heterogeneous standards [16].

Cross-Domain Observations

Metadata integration across several domains demonstrates recurring patterns that amalgamate ontologies (e.g., OBO, BFO, SNOMED-CT, SSN, CIDOC-CRM, GenEpiO, FoodOn).

Includes Semantic Web standards including RDF, RDFS, OWL, SPARQL, JSON-LD, SKOS, DCAT, SHACL, and PROV.

and supplementary domain-specific standards (Dublin Core, Schema.org, ISO/IEC 11179, ISO 19115, HL7 FHIR,

LOINC, UMLS, DICOM, OMOP CDM. This diversity is underpinned by an extensive array of tools for modeling, reasoning, and storing (Protégé, Apache Jena, Neo4J, ArangoDB, RDF4J, Pellet, Hermit, FaCT++), in addition to external resources (DBpedia, WordNet, Wikidata, GeoNames). Thesauri and knowledge base application programming interfaces (APIs). The methodology encompasses strategies such as the implementation of controlled vocabularies and standardized metadata schemas, ontology-driven and metadata-driven languages, established standards and processes, rule-based and schema matching techniques, utilization of linked data resources, AI-driven methodologies, semantic annotations, and graph-oriented models. These factors collectively establish a basis for handling heterogeneity, hence enhancing interoperability in interdisciplinary contexts, as seen in Figure 6.

Conclusion and Future Research Directions

This systematic study offers a thorough examination of the strategies, methods, and

approaches employed for metadata integration across several domains. The study, through the examination of 81 peer-reviewed articles from 2014 to 2023, revealed prevalent approaches like ontology-based frameworks, schema and semantic mapping, centralized metadata models, and the increasing utilization of AI-assisted enrichment.

The assessment emphasizes ongoing issues, including information heterogeneity, insufficient automation, lack of standardization, usability obstacles, and the absence of semantic reasoning capabilities in numerous integration technologies. Tools and platforms exhibit significant variability in maturity and adoption, with the majority necessitating manual curation or domain-specific modification.

Additional research is recommended to automate metadata management, focusing on the implementation of machine learning algorithms and natural language processing techniques, while also enhancing support for semantic inference systems. The creation and implementation of reusable, domain-specific ontologies may enhance reasoning and interoperability. Moreover, federated and decentralized metadata repositories adhering to FAIR principles can facilitate scalable integration among systems and institutions.

Future studies should investigate hybrid methodologies that combine ontologies with developing techniques like LLMs to facilitate more intelligent and flexible metadata systems, as indicated by the conclusions of this review and prevailing research trends. Expanding the scope of metadata integration to include inference generation and cross-domain harmonization is a viable subject for further exploration.

REFERENCES

[1] Kota, T. K. (2022, May). Automated Data Classification and Metadata Management Using Machine Learning in Python. In *International Conference on Intelligence-Based Transformations of Technology and Business Trends* (pp. 433-448). Cham: Springer Nature Switzerland.

[2] Chennareddy, R. K., & Sethuraman, P. (2024c). Decision-Centric Architectures for Intelligent and Networked Wireless Computing Environments Operating at Scale and Uncertainty. *International Journal of Emerging Trends in Computer Science and Information Technology*, *5*(3), 150-160.

[3] Sethuraman, P. (2022). Latency-Aware Scheduling and Resource Control Algorithms for Emergency and Public Safety Wireless Networks. *International Journal of Emerging Research in Engineering and Technology*, *3*(4), 133-140.

[4] Chennareddy, R. K. (2020). Engineering Intelligence Systems Using Big Data and Cloud Architectures for Modern Data Intensive Applications. *International Journal of AI, BigData, Computational and Management Studies*, *1*(2), 41-50.

[5] Li, T., Liang, F., Quan, J., Chuang, H., Wang, T., Huang, R., ... & Hu, X. (2022, March). Taste: Towards Practical Deep Learning-based Approaches for Semantic Type Detection in the Cloud. In *EDBT* (pp. 324-336).

[6] Sethuraman, P. (2023). Implicit Channel Inference Techniques for Pilotless OFDM Reception in Next-Generation Wireless Systems. *International Journal of Emerging Research in Engineering and Technology*, *4*(1), 143-152.

[7] Chennareddy, R. K. (2021). Designing Data and Analytics Ecosystems for High Volume Transaction Processing Applications. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2),

95-106.

[8] Ray, P. P. (2022). A review on vibe coding: Fundamentals, state-of-the-art, challenges and future directions. *Authorea Preprints*.

[9] Sethuraman, P., & Chennareddy, R. K. (2022a). Intelligent Vehicular Traffic Flow Prediction Using Learning-Based Spatio-Temporal Models for Data-Driven Wireless Transportation and Urban Analytics Systems. *International Journal of Emerging Trends in Computer Science and Information Technology*, *3*(2), 111-121.

[10] Chennareddy, R. K., & Sethuraman, P. (2024a). AI-Enabled Data-Driven Decision Frameworks for Enterprise Platforms and Tactical Defense Wireless Networks. *American International Journal of Computer Science and Technology*, *6*(4), 39-49.

[11] Sethuraman, P., & Chennareddy, R. K. (2022b). Machine Learning Assisted Design of Wireless Access Systems for Reliable and Low-Latency Financial and Smart Commerce Services. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(4), 133-142.

[12] Arul, K. (2022). Data Engineering Challenges in Multi-cloud Environments: Strategies for Efficient Big Data Integration and Analytics. *International Journal of Scientific Research and Management (IJSRM)*, *10*(06).

[13] Chennareddy, R. K. (2023). Enterprise-Scale AI and Analytics Strategy for End-to-End Business Transformation across Global Organizations. *International Journal of AI, BigData, Computational and Management Studies*, *4*(3), 134-145.

[14] Sethuraman, P., & Chennareddy, R. K. (2023a). AI-Based Fraud Detection and Prevention at the Radio Access Network: Architectures and Mechanisms for Financial Wireless Service. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *4*(4), 132-141.

[15] Chennareddy, R. K., & Sethuraman, P. (2023). Enterprise and RAN-Aware Data and Analytics Platforms for Mission-Critical and Low-Latency Digital Services. *International Journal of Emerging Trends in Computer Science and Information Technology*, *4*(4), 184-192.

[16] Sethuraman, P., & Chennareddy, R. K. (2023b). System-Level Design and Orchestration of Large-Scale Cellular Access Networks for Regulatory-Compliant Financial Services. *International Journal of Emerging Research in Engineering and Technology*, *4*(3), 140-150.

[17] Martins, M., Jardim, B., Neto, M. D. C., & Barriguinha, A. (2021). Talking to Data: A Systematic Review of the Rise of Conversational Agents for Visual Analytics. *IEEE Access*, *13*, 208902-208931.

[18] Chennareddy, R. K., & Sethuraman, P. (2024b). Data and Analytics Workflows for Decision Systems Enabled by Learning-Based RAN Intelligence across Distributed Computing Environments. *International Journal of Emerging Trends in Computer Science and Information Technology*, *5*(2), 149-158.

[19] Sethuraman, P., & Chennareddy, R. K. (2024). RAN-AI Architectures Supporting Personalized Customer Interaction and Virtual Assistance in Banking Services. *American International Journal of Computer Science and Technology*, *6*(6), 57-66.