

# Semantic Customer 360: A Context-Preserving Architecture for Unifying Heterogeneous Enterprise Customer Data

Swapna Putti

Salesforce Solution Manager, Pearson Education Inc, United States

| Article Info                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | ABSTRACT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Article history:</b></p> <p>Received, April, 2024<br/>Revised, April, 2024<br/>Accepted, April, 2024</p> <hr/> <p><b>Keywords:</b></p> <p>Context-Aware Computing;<br/>Customer 360;<br/>Customer Relationship Management;<br/>Data Lake Architecture;<br/>Data Protection Compliance;<br/>Entity Resolution;<br/>Knowledge Graphs;<br/>Linked Data;<br/>Master Data Management;<br/>Ontology-Based Data Access;<br/>Schema Matching;<br/>Semantic Data Integration</p> | <p>Customer information is handled in enterprise organisations through customer relationship management systems, transactional databases, support systems, and marketing systems, where the meaning is coded using incompatible local schemas. Without capturing the context of the source of those records, where and when they were created, and which regulatory contexts the records are relevant to, such unified views of customers are complete in content but lacking in interpretive fidelity. This paper aims to compile the literature in the fields of context awareness, linked data, entity resolution, ontology-based data access, master data management, knowledge graphs, data lake architecture and data protection regulation to propose an architecture for the Semantic Customer 360 that aims to preserve explicit contexts, and not merely data consolidation. The architecture places a governed data-lake ingestion zone, a hybrid schema-matching module, a probabilistic and graph-based entity-resolution pipeline, and a knowledge-graph-structured semantic layer under the hood of customer-facing analytics and service consumption. The synthesis shows that several candidate comparison techniques can lead to over 10 times fewer comparisons than naive pairwise comparison, that a choice of master data management strategy has measurable impact on governance overhead, that ontology-based mediation can also enable access for queries to sensitive data without moving or copying physical data, and that perceived regulatory compliance is measurable and correlated with customers' trust. The architecture proposed along with comparative analysis it provides can help enterprise data architects to gauge unstructured approaches to evaluating matching algorithms, integration strategies and governance mechanisms that can help them harmonize disparate customer data while keeping them easily traceable and audit-friendly in the present landscape of data protection requirements.</p> <p><i>This is an open access article under the <a href="#">CC BY-SA</a> license.</i></p> <div></div> |

|                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Corresponding Author:</b></p> <p>Name: Swapna Putti<br/>Institution: Pearson Education Inc, United States<br/>Email: <a href="mailto:pswapna.cse@gmail.com">pswapna.cse@gmail.com</a></p> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 1. INTRODUCTION

The ability to have a clear view of the customer's identity, preferences and behavior is a key element in the business of marketing,

service and customer retention for enterprise organizations. Customer information, on the other hand, is often spread out in transactional databases, customer relationship management systems, support

ticketing systems, and marketing automation systems, all with their own schemas, and local business rules. If heterogeneous records were combined without taking into account the situational and relational conditions under which they were created, a composite view of the customer would be lacking in interpretive fidelity. Context-aware computing is similar to the context-dependency problem that occurs in ubiquitous computing systems, where the information collected by sensors or applications needs to be understood in the context in which it was gathered instead of as a fact without context. This principle can be extended to enterprise customers to drive a context-preserving Customer 360 application architecture where the integration mechanisms carry the history, time validity, and meaning of the integrated values, as well.

Some of the key technical ingredients of such an architecture are the ability to represent customer attributes and relationships in a way that allows machines to interpret the meaning, not just tabular values. The idea of the linked data paradigm is to publish data with globally unique identifiers, standardized vocabularies, and typed relationships to allow the ability to interweave and traverse data as one navigable graph [1]. In an internal enterprise environment, the application of the principles of linked data enables that there are statements that express the attributes of a customer that come from separate operational systems, and the relationship between the attributes is explicit and not implicit in application code. This is the semantics layer that is embodied in the representation presented here is the basis for this paper's proposed architecture.

The other challenge is to determine which records in different data sources are for the same customer. Identifiers, naming conventions and data quality are different across systems, and naive key-based joins often fail to realize that multiple records refer to a single person or organization. The problem is addressed by entity resolution and record linkage techniques which compare pairs of candidate records on a number of similarity dimensions and determine whether they match, don't match, or are possibly similar that must be reviewed. This means that the unified customer profile generated by a customer 360 architecture is also dependent on the accuracy of its entity resolution component, and therefore a main design concern, not a peripheral data-cleaning task.

Ontology-based data access techniques link a single global ontology to multiple underlying data sources by defining declarative mapping rules that enable queries to be made against the ontology, and automatically translate those queries into the underlying source-specific queries, without actually moving all the data into one repository [2], [3]. The benefit of this mediation strategy is that it is appealing in enterprise environments where for operational, contractual or technical reasons, it is not possible to consolidate the source systems. For this, this paper consolidates the literature covering the areas of context-aware computing, semantic web technologies, entity resolution, master data management, and knowledge graph construction, and discusses the data protection regulation, as well as the integration of these techniques into a proposed Semantic Customer 360 architecture that keeps the context information intact throughout the integration lifecycle.

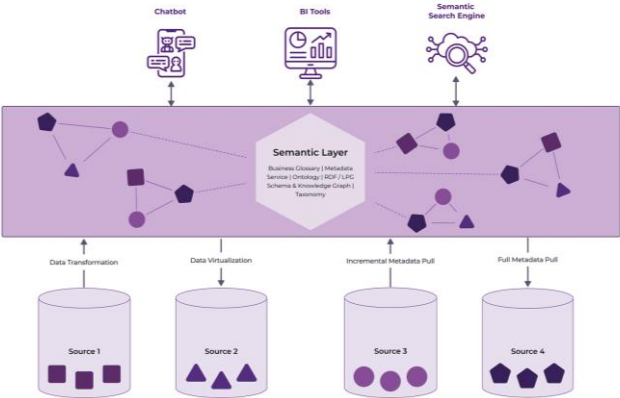


Figure 1. Semantic Customer 360 Layered Architecture (Enterprise Knowledge, 2019)

2. LITERATURE REVIEW

2.1 Artificial Intelligence and Customer Relationship Management

With the introduction of artificial intelligence, customer relationship management platforms have continued to embrace features such as predictive scoring, personalization and segmentation automation. A preliminary analysis of the examples found in the literature on the use of AI systems in customer relationship management reveals positive empirical results regarding the performance of customer engagement functions and the efficiency of the processes in use, but also organizational issues linked to data readiness and process redesign [4]. The results indicate that AI-enhanced CRM can only be truly effective when the data layer is integrated and well-managed, not just in terms of integration, but the quality of the data as well. The findings underscore the need for a robust data-unification architecture as the foundation

for downstream AI-driven recommendations.

2.2 Organizational Adoption of AI-Integrated Systems

In addition to the technical performance, the organizational adoption of AI-integrated CRM systems has been studied using longer versions of the technology acceptance models that also include behavioral, trust-related and situational factors that influence intention to use. A unified theory of acceptance and use of technology (UTAUT) approach was used to systematically analyze the findings, which suggest that performance expectancy, effort expectancy and perceived data trustworthiness collectively influence users' willingness to rely on systems that integrate AI [5]. The literature suggests that while a Customer 360 architecture can accurately connect data, it also must deliver integrated views in a way that enables users to trust and interpret the information, or despite its technical capability, the architecture will not be adopted.

| Table 1. Comparative Overview of Context-Aware and Semantic Integration Paradigms |                                                     |                                                                           |      |
|-----------------------------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------|------|
| Paradigm                                                                          | Core Mechanism                                      | Primary Contribution                                                      | Ref. |
| Context-Aware Computing                                                           | Adapts data interpretation to situational metadata  | Frames provenance and situational validity as first-class data properties | [6]  |
| Linked Data / Semantic Web                                                        | Globally unique identifiers and typed relationships | Enables cross-source navigability without physical consolidation          | [1]  |
| Ontology-Based Data Access                                                        | Global ontology mediates queries via mapping rules  | Supports querying without migrating source data                           | [2]  |

| Paradigm               | Core Mechanism                                            | Primary Contribution                                      | Ref. |
|------------------------|-----------------------------------------------------------|-----------------------------------------------------------|------|
| Knowledge Graphs       | Typed nodes and edges representing entities and relations | Provides flexible, extensible schema for evolving domains | [7]  |
| Data Lake Architecture | Staged ingestion, storage, processing, governance zones   | Accommodates heterogeneous formats and update frequencies | [8]  |

Source: Table entries synthesized from the reviewed literature; bracketed numbers reference the corresponding source.

2.3 Entity Resolution at Enterprise Scale

In general, end-to-end entity resolution pipelines for large data sets include blocking, pairwise comparison, classification and clustering steps, each with its own set of compromises between computational expense and accuracy of the matches. A detailed description of these pipelines highlights the evolution of the switch from rule-based pipelines to learning-based and hybrid pipelines that can operate on heterogeneous, high volume data sources [9]. This is a topic that is directly connected to Customer 360 efforts, which are likely to unravel the identities of potentially millions of customers from multiple channels, all while adhering to the requirement for acceptable processing latency for near real-time customer service applications.

2.4 Data Lake Architectures as Ingestion Substrates

Enterprise customer data comes in many different formats and frequency of updates, so many enterprise organizations use data lake architectures to ingest raw source data before they transform it. A data lake architecture framework that distinguishes data ingestion, storage, processing and governance zones aims to keep the lake

from becoming an uncontrolled data repository [8], [10]. Having those zoning principles in a customer 360 pipeline creates a staging environment to capture context information, such as the source system, extraction time, data quality metrics etc., that precede semantic transformation and entity resolution.

3. METHODS

3.1 Search Strategy

Adopting a systematic literature synthesis approach, this review involved the identification of academic sources on context-aware computing, semantic data integration, entity resolution, master data management, knowledge graphs, and data protection compliance, and the thematic organization. Similar systematic methods used for information systems studies of the adoption of master data management have shown that structured identification of literature can yield organizational determinants that may be otherwise missed in narrative literature reviews [11]. The sources were taken from peer-reviewed journals and conference proceedings, with a focus on articles that studied the architectural, algorithmic, and governance aspects of enterprise customer data unification.

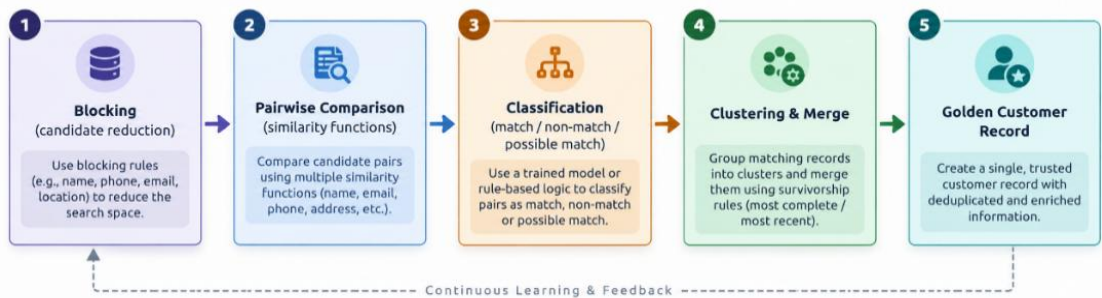


Figure 2. Entity Resolution Pipeline for Customer Identity Consolidation

Table 2. Thematic Classification of Reviewed Literature (n = 24)

| Thematic Category                           | Representative Focus                                            | Sources |
|---------------------------------------------|-----------------------------------------------------------------|---------|
| Entity Resolution & Schema Matching         | Matching, blocking, probabilistic and machine-learning methods  | 6       |
| Semantic Web, Ontologies & Knowledge Graphs | Linked data, ontology mediation, knowledge graph representation | 6       |
| Data Lake & Data Integration Architecture   | Ingestion zoning, metadata management, theoretical integration  | 4       |
| Master Data Management                      | Adoption determinants, implementation strategy, business value  | 3       |
| AI-Enabled CRM & Technology Adoption        | AI-embedded CRM, organizational acceptance                      | 2       |
| Data Protection & Customer Trust            | GDPR compliance, customer trust in data handling                | 2       |
| Context-Aware Computing                     | Situational interpretation of distributed data                  | 1       |

Source: Categories are mutually exclusive and sum to the twenty-four references underlying this synthesis.

### 3.2 Inclusion Criteria and Thematic Classification

It included publications that touched on one or more of the following areas: Semantic or Ontology based data integration, Methodologies for entity resolution, Master data management strategy, Knowledge graph construction, Data Lake architecture, or regulatory governance of customer data. To demonstrate the practical relevance of integration strategy, for example between registry-based, consolidation-based and coexistence-based, in relation to organizational structure and data ownership of an international hearing healthcare company, a case-study analysis is provided [12]. The classification scheme used for the categorization of the literature reviewed is based on the case-based evidence and provided in Table 2.

### 3.3 Architecture Synthesis Method

To synthesize the proposed Semantic Customer 360 architecture, the rules of constructing a knowledge graph [7] that describes how heterogeneous facts can be organized into typed nodes and relationships for human understanding and automated reasoning were used, by mapping each identified thematic category onto an architectural layer. The mapping exercise was done in an iterative manner; for each architectural

component, its capabilities and limitations were documented in the literature reviewed and the component was improved until some substantiating body of prior work could be used to support the architectural component.

## 4. RESULTS AND DISCUSSION

### 4.1 Semantic Layer Design

The semantic layer of the proposed architecture is the integrated customer data, represented as a knowledge graph, with nodes representing things such as individual customers, organizational accounts, products, and interactions, and edges representing typed relations between them. The survey results corroborate the fact that knowledge graph representation, acquisition and usage is supported by survey evidence, showing that the graph representation provides flexibility in schema evolution, and multi-hop reasoning was only possible with significant redesign effort for relational schemas [13]. In a customer 360 use case, this flexibility enables the use of new customer touchpoints or data sources to be added over time rather than re-architecting existing tables, a significant feature for one that can always add new digital engagement points.

#### 4.2 Context Preservation Mechanisms

The problem of data integration theory is that it focuses on integrating heterogeneous data sources by reconciling a global schema with local schemas from independent sources by making mapping assertions that define the relationships between the global concepts and the data in the local sources [14]. The proposed architecture also differs by preserving context information, which includes provenance and validity attributes like the source system identifier, extraction time, confidence score, and applicable jurisdictional scope of the statements as they are ingested into the semantic layer, instead of discarding them as is normally done in an extract-transform-load pipeline.

#### 4.3 Probabilistic Entity Resolution

In addition to deterministic matching rules, probabilistic entity resolution methods approximate the probability that two given records are associated with the same entity in the real world by modelling the partition structure over the space of possible assignments of entities to records. When match decisions have material consequences, downstream decision making, e.g., consent management or credit assessment, for the individual involved, then principled uncertainty quantification is important, particularly when using Bayesian graphical approaches with exchangeable random partition priors. A simplified log-likelihood ratio formulation is often used to obtain a composite match score by combining the probabilities of agreement at each field (assuming they are independent) as in Equation 1, wherein the fields are weighted by their discriminating power between a true match and chance agreement.

$$\log L(r) = \sum_i w_i \cdot \left[ m_i \log \left( \frac{m_i}{m_i + u_i} \right) + (1 - m_i) \log \left( \frac{1 - m_i}{1 - m_i + u_i} \right) \right] \quad (1)$$

In Equation 1,  $m_i$  denotes the probability that the two candidates agree on field  $i$  when the two are a true match,  $u_i$  denotes the probability that the two candidates agree on field  $i$  when they are not a true match, and  $w_i$  weights the contribution of field  $i$  to the composite score, with candidate pairs whose composite score is above a calibrated threshold being classified as matches.

#### 4.4 Ontology-Based Integration in Regulated Domains

Ontology-based integration has shown to be useful in regulated areas, such as enterprise customer management, where there are structural similarities, like the need to reconcile clinical and research data under different consent and access restrictions. A shared conceptual model has been shown to be able to provide access to sensitive information in clinical and research information systems without physical merging [15]. This precedent serves as an example of how a similar approach can be used to mediate customer data stored in systems of different regulatory sensitivity, for example, a marketing system and a payment or health related service system.

#### 4.5 Data Lake Ingestion and Management Challenges

While data lakes offer flexibility, they also have known challenges, such as schema drift, data quality issues and loss of metadata if not properly managed. Data lake management challenges analysis reveals that there are a number of common problems that exist between the “promised land” and the “real lake” of data lakes, including metadata cataloging and data-quality monitoring [16]. The proposed architecture mitigates this risk by having the data lake be a staging area that is promoted to the semantic layer

only after the data has been validated for the schema and enriched for metadata – not the source of the system of record for customer queries.

#### 4.6 Matching Algorithm Selection

There are several factors to consider when choosing a suitable one-to-one entity resolution matching function to use, such as the precision, recall and computational efficiency, as well as the nature of the data, such as the completeness of the data fields, the variation in naming conventions, and the

rate of data duplication. A comparative evaluation of one-to-one matching techniques shows that there is no single technique that performs well under all data conditions; and the performance order may vary under different data densities and noise types [17]. Table 3 summarizes the qualitative trade-off between the different families of matching methods cited in the reviewed literature and gives a decision guide for architects to choose a matching method for a specific customer data landscape.

Table 3. Comparative Trade-offs Among Entity Resolution Matching Algorithm Families

| Algorithm Family               | Matching Basis                                   | Precision Profile     | Recall Profile  | Cost          | Ref. |
|--------------------------------|--------------------------------------------------|-----------------------|-----------------|---------------|------|
| Rule-Based / Deterministic     | Exact or threshold field agreement               | High on clean data    | Low under noise | Low           | [18] |
| Probabilistic (Fellegi–Sunter) | Likelihood ratio over agreement patterns         | Moderate–High         | Moderate–High   | Moderate      | [19] |
| Machine-Learning-Based         | Learned similarity function over labeled pairs   | High (data-dependent) | High            | Moderate–High | [17] |
| Graph-Based / Collective       | Joint resolution exploiting relational structure | High                  | High            | High          | [9]  |

Source: Profiles are qualitative syntheses of reported trends rather than results of a single controlled benchmark.

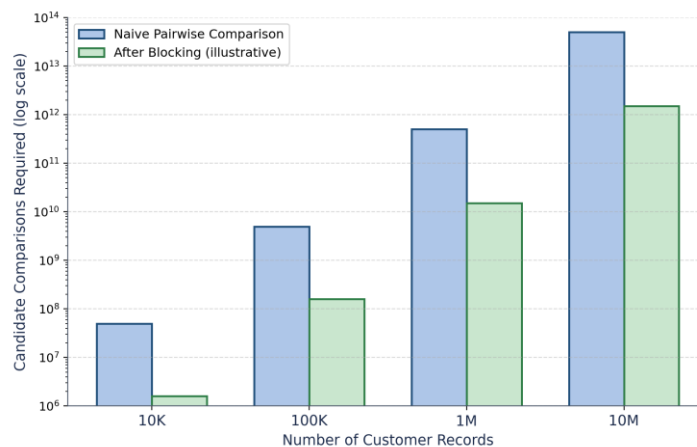


Figure 3. Effect of Blocking on Entity Resolution Comparison Volume (Illustrative)

#### 4.7 Blocking for Scalability

When comparing all candidate records with all records in the database, the number of comparisons needed to make before classifies increases quadratically with the number of records, so blocking and filtering techniques partition the candidate set into smaller groups that are more likely to contain a

match, thus reducing the number of comparisons that need to be made before a classification decision is reached. Based on the evidence from the surveys, well-designed blocking keys can achieve the vast majority of the elimination of unwanted comparisons, while preserving most true matches, thereby making entity resolution at the enterprise scale

computationally tractable [20]. In the figure 3, the curves are showing that blocking method was more effective at reducing comparison volume than naive pairwise comparison method with increasing record volume.

#### 4.8 Schema Matching Automation

Before entity resolution, the schema for each source needs to be aligned in order to identify the fields that are semantically equivalent. For instance, an account holder name field in one system is recognized as equivalent to a customer full name field in another, despite its different name. Automatic schema matching is an area of foundational work that categorizes matching techniques based on either schema-level structure or instance-level data or using auxiliary linguistic resources, and shows that more than one technique is needed to achieve a better performance than using only one technique, in general [21]. The planned

architecture is a hybrid schema-matching module upstream from the semantic and to reduce the amount of manual work needed in the case of new source systems.

#### 4.9 Regulatory Compliance Considerations

A Customer 360 architecture brings together personal data from various scenarios, so it needs to meet requirements for data protection, including those relating to lawful basis, data purpose limitation, and data subject rights. The importance of privacy protection measures, such as access control, pseudonymization, and audit logging, in big data systems is stressed, noting that they need to be integrated into the system from the outset, rather than added as an afterthought [22]. Table 5 shows the provisions of the regulation that are especially relevant to unified customer data platforms, all of which demand the ability to trace the architecture at the entity level of all integrated sources.

Table 4. Comparison of Master Data Management Implementation Strategies

| Strategy                       | Data Ownership Model                            | Organizational Fit                             | Governance Overhead | Ref. |
|--------------------------------|-------------------------------------------------|------------------------------------------------|---------------------|------|
| Registry (Reference-Only)      | Sources retain data; hub holds cross-references | Loosely coupled business units                 | Low                 | [11] |
| Consolidation                  | Data physically copied into central hub         | Reporting and analytics-focused units          | Moderate            | [23] |
| Coexistence                    | Hub synchronized bidirectionally with sources   | Needs both central and local control           | High                | [12] |
| Centralized (System of Record) | Hub is the sole authoritative source            | Highly regulated, transaction-critical domains | High                | [23] |

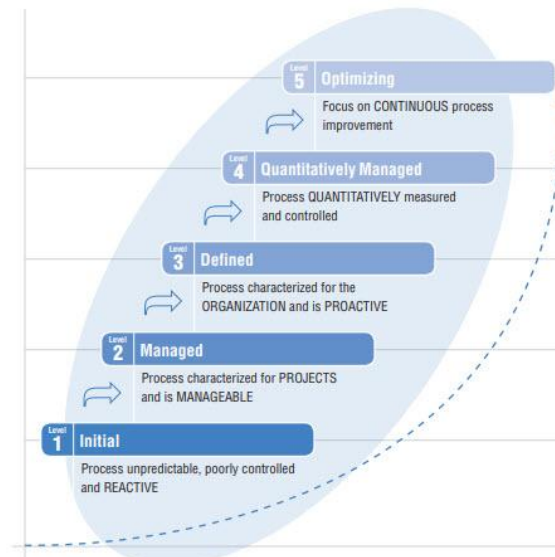


Figure 4. Master Data Management Governance Maturity Progression (LightsOnData, 2019)

## 5. CONCLUSION

This review has synthesized literature spanning context-aware computing, semantic web technologies, entity resolution, ontology-based data integration, master data management, knowledge graphs, data lake architecture, and data protection regulation to propose a Semantic Customer 360 architecture oriented around context preservation rather than mere data consolidation. The architecture layers a governed data-lake ingestion zone, a hybrid schema-matching module, a probabilistic and graph-based entity resolution pipeline, and a knowledge-graph-structured semantic layer that retains provenance, temporal validity, and regulatory scope alongside integrated customer attributes. Findings drawn from the synthesized literature indicate that no single technical component is sufficient in isolation:

matching algorithm selection must be paired with adequate blocking strategies to remain tractable at scale, ontology mediation must be paired with versioning discipline to remain maintainable as source systems evolve, and technical integration accuracy must be paired with transparent governance to sustain customer trust [24]. Organizations pursuing enterprise customer data unification are therefore advised to treat context preservation, encompassing provenance, consent scope, and data lineage, as a first-class architectural requirement rather than an auxiliary compliance concern. Future work building on the reviewed literature might further examine the operational performance of hybrid probabilistic-graph entity resolution pipelines under production-scale customer data volumes and the longitudinal effect of transparent context preservation on measured customer trust.

## REFERENCES

- [1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," in *Linking the world's information: Essays on tim berners-lee's invention of the world wide web*, 2023, pp. 115–143. doi: 10.4018/jswis.2009081901.
- [2] D. Calvanese *et al.*, "The MASTRO system for ontology-based data access," *Semant. Web*, vol. 2, no. 1, pp. 43–53, 2011, doi: 10.3233/SW-2011-0029.
- [3] H. Wache *et al.*, "Ontology-Based Integration of Information: A Survey of Existing Approaches," in *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, 2001, pp. 108–117. [Online]. Available: <https://dblp.org/rec/conf/ijcai/WacheVVSSNH01>
- [4] S. Chatterjee, R. Chaudhuri, and D. Vrontis, "AI and digitalization in relationship management: Impact of adopting AI-embedded CRM system," *J. Bus. Res.*, vol. 150, pp. 437–450, 2022, doi: 10.1016/j.jbusres.2022.06.033.
- [5] S. Chatterjee, N. P. Rana, S. Khorana, P. Mikalef, and A. Sharma, "Assessing organizational users' intentions and behavior to AI integrated CRM systems: A meta-UTAUT approach," *Inf. Syst. Front.*, vol. 25, no. 4, pp. 1299–1313, 2023, doi: 10.1007/s10796-021-10181-1.

- [6] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 2, no. 4, pp. 263–277, 2007, doi: 10.1504/IJAHUC.2007.014070.
- [7] A. Hogan *et al.*, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, 2021, doi: 10.1145/3447772.
- [8] C. Giebler, "The data lake architecture framework," 2021, doi: 10.18420/btw2021-19.
- [9] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, "An Overview of End-to-End Entity Resolution for Big Data," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–42, 2020, doi: 10.1145/3418896.
- [10] P. N. Sawadogo and J. Darmont, "On data lake architectures and metadata management," *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 97–120, 2021, doi: 10.1007/s10844-020-00608-7.
- [11] F. Haneem, N. Kama, N. Taskin, D. Pauleen, and N. A. A. Bakar, "Determinants of master data management adoption by local government organizations: An empirical study," *Int. J. Inf. Manage.*, vol. 45, pp. 25–43, 2019, doi: 10.1016/j.ijinfomgt.2018.10.007.
- [12] A. Haug, A. M. Staskiewicz, and L. Hvam, "Strategies for master data management: a case study of an international hearing healthcare company," *Inf. Syst. Front.*, vol. 25, no. 5, pp. 1903–1923, 2023, doi: 10.1007/s10796-022-10323-z.
- [13] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022, doi: 10.1109/TNNLS.2021.3070843.
- [14] M. Lenzerini, "Data integration: A theoretical perspective," in *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2002, pp. 233–246. doi: 10.1145/543613.543644.
- [15] S. Mate *et al.*, "Ontology-based data integration between clinical and research systems," *PLoS One*, vol. 10, no. 1, p. e0116656, 2015, doi: 10.1371/journal.pone.0116656.
- [16] F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, "Data lake management: challenges and opportunities," *Proc. VLDB Endow.*, vol. 12, no. 12, pp. 1986–1989, 2019, doi: 10.14778/3352063.3352116.
- [17] G. Papadakis, V. Efthymiou, E. Thanos, O. Hassanzadeh, and P. Christen, "An analysis of one-to-one matching algorithms for entity resolution," *VLDB J.*, vol. 32, no. 6, pp. 1369–1400, 2023, doi: 10.1007/s00778-023-00791-3.
- [18] D. G. Brizan and A. U. Tansel, "A Survey of Entity Resolution and Record Linkage Methodologies," *Commun. IIMA*, vol. 6, no. 3, p. 5, 2006, doi: 10.58729/1941-6687.1324.
- [19] N. G. Marchant, B. I. P. Rubinstein, and R. C. Steorts, "Bayesian graphical entity resolution using exchangeable random partition priors," *J. Surv. Stat. Methodol.*, vol. 11, no. 3, pp. 569–596, 2023, doi: 10.1093/jssam/smac030.
- [20] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and Filtering Techniques for Entity Resolution: A Survey," *ACM Comput. Surv.*, vol. 53, no. 2, pp. 1–42, 2020, doi: 10.1145/3377455.
- [21] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching," *VLDB J.*, vol. 10, no. 4, pp. 334–350, 2001, doi: 10.1007/s007780100057.
- [22] M. Rhahla, S. Allegue, and T. Abdellatif, "Guidelines for GDPR compliance in Big Data systems," *J. Inf. Secur. Appl.*, vol. 61, p. 102896, 2021, doi: 10.1016/j.jisa.2021.102896.
- [23] S. Singh and J. Singh, "A survey on master data management techniques for business perspective," in *Cyber Intelligence and Information Retrieval: Proceedings of CIIR 2021*, Springer, 2021, pp. 609–617. doi: 10.1007/978-981-16-4284-5\_54.
- [24] J. Zhang, F. Hassandoust, and J. E. Williams, "Online Customer Trust in the Context of the General Data Protection Regulation (GDPR)," *Pacific Asia J. Assoc. Inf. Syst.*, vol. 12, no. 1, pp. 86–122, 2020, doi: 10.17705/1pais.12104.