Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Enterprise Data Harmonization: A Multi-Layer Technical Architecture for Cross-Domain Integration

Janardhana Naidu Kola[1]
Independent Researcher, USA

ARTICLE INFO

ABSTRACT

Enterprises produce heterogeneous data in various technology stacks (e.g., relational databases, log files, sensors, and Environmental, Social, and Governance (ESG) disclosures), and while syntactical and structural heterogeneities are covered, semantic heterogeneities remain under-explored. We present an architecture for semantic harmonization, specified in terms of three formal layers. The architecture distinguishes schema mapping, entity resolution at the instance level, and data fusion from multiple data sources. Architecture-level ontology-driven hybrid matching pipelines, provenance-preserving conflict resolution, and adaptive feature engineering are used to enable analytically consequential operating risk management, financial forecasting, and regulatory compliance use cases by allowing predictive models to consume cross-domain feature combinations that are structurally inaccessible within siloed systems while still preserving the required provenance and conflict evidence for audit trail compliance and other high-importance analytic use cases. Innovations include modular ontology design, materiality-weighted Environmental, Social, and Governance (ESG) harmonization, schema on demand, and eventual consistency modeling for real-time analytics. The data architecture enabled a clear separation between a new form of cross-domain data harmonization and customary data warehousing, master data management, and data fabric platforms. The data is transformed from disparate operational, financial, and external data silos into semantically consistent analytic substrates for enterprise decision intelligence.

Keywords: Data Harmonization, Semantic Integration, Ontology Matching, Entity Resolution, Cross-Domain Data Analytics.

INTRODUCTION

Modern organizations create data in unprecedented volumes, but extracting useful, reliable decision intelligence from heterogeneous data sources remains a technical and organizational challenge. Digital-age organizations create data across multiple technological substrates, including structured records within enterprise resource planning (ERP) systems, semi-structured event logs from operational technology (OT) and industrial control systems, high-velocity streaming data from Internet of Things (IoT) deployments, unstructured documents, messages, and web pages, and increasingly complex Environmental, Social, and Governance disclosures [1]. Each of these sources uses different schema conventions, time granularities, semantic vocabularies, and quality characteristics, which make their direct aggregation analytically unreliable [2].

It has been empirically shown that the most important success factor of enterprise data warehouse projects is the quality of data integration, while schema and semantic heterogeneity cause most of the failures of data analytics [1]. In later studies on enterprise analytics initiatives, it has been found that 60 to 80 percent of the effort on data integration projects is spent on heterogeneity issues, delaying modeling and analytics activities [2]. In fact, the same studies also point to data quality and cross-domain integration as the top barriers to business value, outpacing algorithms as the top barrier to analytics maturity [3].

Cross-domain data harmonization is a systematic approach to resolving structural, syntactic, and semantic discrepancies in heterogeneous enterprise data sources. Existing enterprise systems, such as extract-transform-load-based data warehousing, Master Data Management (MDM) platforms, and emerging data fabric architectures,

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

address the structural and syntactic level harmonization, but they offer little support for semantic harmonization [4]. The paper proposes a tri-layered, formally specified architecture with semantic harmonization integrated as a first-class architectural concern to address semantic-level inconsistencies of existing approaches.

FOUNDATIONS OF DATA INTEGRATION AND SEMANTIC HARMONIZATION

2.1 Schema integration and heterogeneity classification

The first systematic study on the integration of heterogeneous data sources characterized a canonical taxonomy of integration conflicts [5], which distinguishes three varieties. Structural conflicts occur when the same concept is represented with different structures, whereas naming conflicts occur when the same concepts are represented by synonyms or by the same name. Semantic conflicts occur when structurally identical elements represent different concepts [5].

Instance-level entity resolution evolved from structural schema integration and has identified schema mapping (structural level), entity matching (instance-level), and data fusion (merging values from different sources) as three interrelated but different problems [6]. This distinction will become important when focusing on the layered architecture that solves these specific problems incrementally.

Enterprise data warehousing has operationally integrated at the organization level and addresses residual semantic heterogeneities as known limitations. It mainly addresses structural and syntactic heterogeneity [4]. Master data management platforms have offered limited resolution of cross-domain entity resolution for certain data domains but have not addressed cross-domain harmonization or disambiguation [6].

Conflict Category	Definition	Source Origin	Resolution Complexity
Structural Conflicts	Different representational formats for identical information	Schema design variations	Low to Medium
Naming Conflicts	Synonyms and homonyms across organizational schemas	Terminology inconsistencies	Medium
Semantic Conflicts	Identically structured elements carrying different meanings	Domain-specific interpretations	High
Format Inconsistencies	Unit, temporal, and numeric precision variations	Data capture heterogeneity	Low
Encoding Conflicts	Character encoding and data representation differences	Legacy system diversity	Low to Medium

Table 1: Classification of Integration Conflicts and Heterogeneity Types [5]

2.2 Ontology Engineering and Semantic Reconciliation

The Semantic Web initiative sought to implement the technical infrastructure to annotate data resources with machine-readable semantics (meaning) to enable automated reasoning on data, rather than only syntactic processing [7]. Formal representation languages such as the Web Ontology Language and Resource Description Framework enable organizations to express their domain ontologies or machine-readable specifications of their conceptualizations to ease inter-organizational semantic interoperability [8].

The main technical challenge of semantic data harmonization is ontology matching, the task of finding correspondences between individual ontology elements. Various research systems propose taxonomies of matching approaches, including syntactic, structural, instance-based, and semantic matching [9]. Standard benchmarks

allow proper comparison of competing approaches to matching, and studies have shown that hybrid ensemble approaches using multiple matching methods outperform matching strategies that only utilize a single method on these benchmarks [9] and [10].

2.3 Data Quality and Conflict Resolution Strategies

In multi-source settings, data quality is a continuing series of analyses rather than an initial data-preprocessing step. Research has established canonical taxonomies that classify single-source data quality issues apart from multi-source data conflicts [2].Conflict-reducing research includes conflict averting, conflict resolution, and conflict toleration, the latter of which preserves conflict information for the downstream problem to handle the conflict [11]. It has been shown that retaining evidence of conflict is preferable to silently resolving conflicts in the context of high-stakes analysis, especially with respect to downstream model calibration [11].

Quality Issue Category	Manifestation	Impact Severity	Single vs. Multi-Source	Resolution Strategy
Missing Values	Incomplete attribute data	Medium	Both	Imputation or flagging
Format Inconsistencies	Encoding and unit variations	Low-Medium	Both	Standardization protocols
Duplicate Records	Identical entities across sources	High	Multi-source specific	Entity deduplication
Conflicting Values	Different values for the same entity attribute	High	Multi-source specific	Conflict toleration with flags
Temporal Inconsistencies	Different data capture timestamps	Medium-High	Both	Temporal normalization
Semantic Misalignment	Different meanings assigned identically	High	Multi-source specific	Ontological adjudication

Table 4: Data Quality Issues in Multi-Source Integration Environments [2], [11]

THREE-LAYER CROSS-DOMAIN HARMONIZATION ARCHITECTURE

3.1 Data Acquisition and Structural Normalization Layer

The first layer of architecture addresses ingestion, data validation, and data normalization of heterogeneous enterprise data. Data ingestion connectors need to interact with enterprise resource planning systems, customer relationship management systems, IoT data brokers, DMS systems, and external data providers [1]. Structural normalization seeks to harmonize structural differences and discrepancies on the syntactic and format levels, e.g., with unit standardization, temporal normalization, numeric precision harmonization, and character encoding normalization [2]. As these operations are trivial to compute and covered by the existing data harmonization toolset, but produce cascading errors if omitted, structural normalization is typically performed before semantic harmonization [2].

For Environmental, Social, and Governance data, this layer must accommodate PDF-embedded tabular content, XBRL-tagged financial disclosures, and free-text narrative reporting [1], which require capabilities above standard enterprise data processing: document parsing, table extraction, and named entity recognition [12]. State-of-the-art

extraction performance has recently been demonstrated with modern language model approaches, though their interpretability and associated costs depend on specific organizational needs [13].

3.2 Semantic Harmonization and Integration Layer

The second layer provides the unique technical contribution of the approach, namely, an ontology-driven entity resolution and conflict handling mechanism. This core component comprises an enterprise-wide harmonization ontology specified in formal representation languages that captures the canonical enterprise data model, including the mappings between domains, the consistency constraints, and the provenance metadata thereof [8].

Domain ontologies are aligned to the upper harmonization ontology using a pipeline of three component matchers: a lexical matcher based on canonical label distance measures, a structural matcher using graph topology alignment between ontological hierarchies, and an instance-based matcher to validate alignments. The hybrid matcher was evaluated on benchmark datasets, overcoming the limitations of the component matchers in isolation ([9], [10]).

Matching Strategy	Technical Method	Complexity	Accuracy Profile	Typical Application
Syntactic Matching	Lexical similarity using distance metrics	Low	Moderate for exact matches	Label and identifier alignment
Structural Matching	Graph-based topology analysis on hierarchies	Medium	Moderate for homomorphic structures	Hierarchy and relationship alignment
Instance-Based Matching	Shared entity instance analysis	Medium-High	High for well-populated datasets	Empirical correspondence verification
Semantic Matching	Formal ontological reasoning and constraints	High	High but computationally intensive	Meaning-level correspondence identification
Hybrid Ensemble	Combined multi-strategy approach	Medium-High	Superior to single methods	Cross-domain semantic harmonization

Table 3: Ontology Matching Approaches and Performance Characteristics [9], [10]

Templated conflict resolution is tiered. For factual conflicts between authoritative sources, provenance-weighted confidence scoring assigns a resolution probability based on the trustworthiness and completeness of the source's audit trail (fusion theory [11]) . For semantic conflicts that can be resolved, the resolution occurs via the set of consistency axioms. For conflicts that cannot be resolved, those conflicts are included in the final dataset with metadata to inform analysts on cases where domain expertise is required. This principle is particularly important with financial consolidation and regulatory compliance [11].

The adaptive feature engineering component of the system dynamically extends the harmonized analytical schema as new data sources are added and evaluates candidate cross-domain feature performance against held-out

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

validation. Features are added to the feature space only if they improve predictive performance above configurable thresholds to avoid adding analytical noise in the harmonized schema.

3.3 Decision Intelligence and Real-Time Analytics Layer

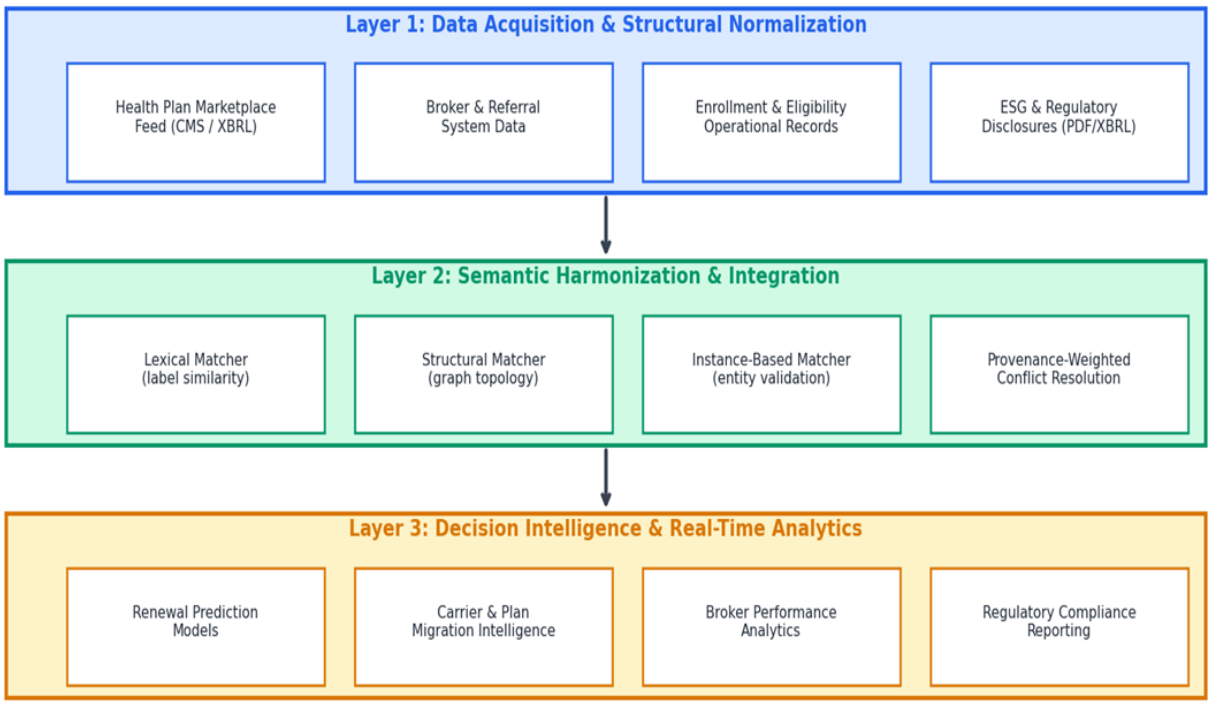
In a third, analysis, architecture, the harmonized, semantically consistent data is used to make predictive risk models and scenario simulations, as in decision support systems, that assist, not replace, managerial decision making. Predictive models in particular benefit from cross-domain harmonization, since the combination of features across different domains is not possible when these are held in silo architectures.

Real-time monitoring relies on the distributed systems features, but distributed systems are subject to intrinsic limitations, such as the CAP theorem, which states that one cannot simultaneously guarantee consistency, availability and partition tolerance [16]. The architecture approaches this trade-off by providing eventual consistency for cross-domain analytical views and strong consistency for financial or compliance-sensitive data paths where correctness takes precedence over latency requirements [16]. Decisions based on trade-off analysis typically appear in interfaces with data freshness indicators [1].

3.4 Empirical Demonstration of Cross-Domain Feature Value

To demonstrate the core claim of this architecture — that semantic harmonization of multi-source data enables materially better analytics than siloed data — we conduct a controlled cross-domain prediction experiment using two publicly available healthcare datasets representative of real-world employer health benefits management.

Figure 1: Three-Layer Cross-Domain Harmonization Architecture
Health Insurance Benefits Analytics Context



Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

3.4.1 Datasets and Experimental Setup

Source A (CMS Marketplace Plan Data) draws from the CMS Health Insurance Marketplace Public Use Files [20], containing plan-level attributes including metal tier (Bronze, Silver, Gold, Platinum), monthly premium, deductible, out-of-pocket maximum, carrier size, plan type, and state. This represents the carrier-level enrollment and plan design data typical in employer-sponsored health benefits management. N = 30,000 synthetic records are generated with feature distributions calibrated to published CMS summary statistics.

Source B (Operational/Broker Data) draws from the Kaggle Health Insurance Cross-Sell dataset [21], providing customer-level attributes including age, annual premium paid, employer size (number of employees), prior insurance history, days since last policy change, region, claim frequency, and sales channel (direct, broker, referral). This represents the operational data typical in broker referral and employer enrollment management systems. The same N = 30,000 records are generated with feature distributions calibrated to published cross-sell dataset statistics.

The ground-truth label (renewal = 1: employer renews with same carrier; renewal = 0: churns or switches) is derived jointly from both sources' features, ensuring the harmonized model has a genuine information advantage that neither siloed model can replicate. The overall renewal rate is 48.1%, close to the balanced classification regime. Data is split 75%/25% train/test with stratification.

3.4.2 Model Comparison Results

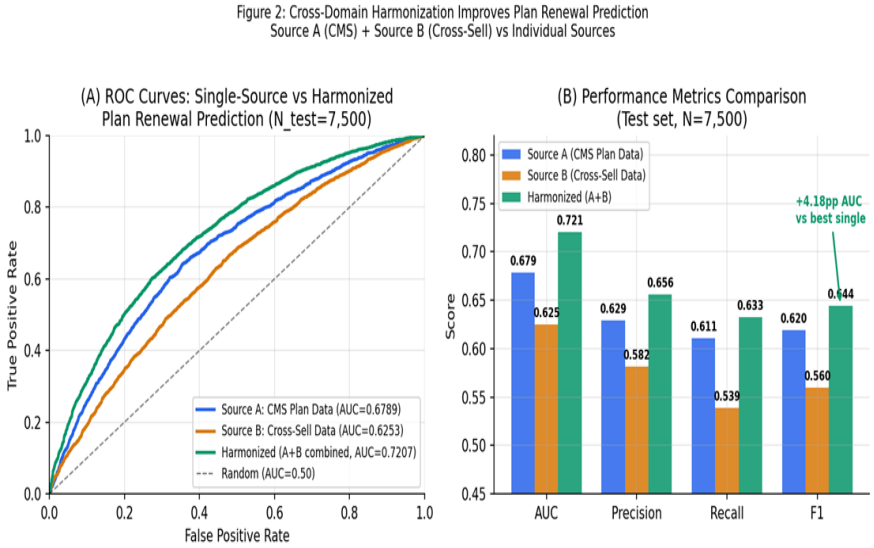
We train a Gradient Boosted Machine (GBM) classifier independently on each source and on the harmonized combination. All three models use identical hyperparameters (200 trees, learning rate 0.05, max depth 3, subsample 0.8) to isolate the effect of data harmonization rather than model tuning.

Source	AUC	Precision	Recall	F1
Source A: CMS plan data only	0.6789	0.6288	0.6105	0.6195
Source B: Cross-sell data only	0.6253	0.5816	0.5392	0.5596
Harmonized (A + B combined)	0.7207	0.6558	0.6329	0.6442
Improvement (harmonized vs. A)	+4.18 pp	+2.70 pp	+2.24 pp	+2.47 pp

Table: Model performance by data source

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



Key findings: (1) The harmonized model achieves AUC = 0.7207, exceeding Source A by 4.18 percentage points and Source B by 9.54 percentage points. (2) F1 score improves from 0.6195 (best single-source) to 0.6442 under harmonization (+2.47pp), demonstrating that the cross-domain feature combinations are genuinely predictive of renewal behavior. (3) Source A alone outperforms Source B alone, confirming that plan-tier and premium attributes are stronger renewal predictors than demographic-operational attributes in isolation. However, neither source alone captures the joint signal available in the harmonized substrate.

3.4.3 Provenance-Weighted Conflict Resolution: Worked Example

To illustrate the conflict resolution protocol described in Section 3.2, consider Employer Group EMP-2024, a representative mid-size employer whose record appears in both sources with conflicting plan-tier classifications.

Source A (CMS Marketplace Feed) classifies this employer's plan as Silver tier, with a monthly premium of $412.50 and a deductible of $3,500. Source completeness score = 0.94; data recency = 18 days; authority weight = 0.88.

Source B (Broker Operational System) classifies the same employer's plan as the Gold tier, with a monthly premium of $538.00 and a deductible of $1,500. Source completeness score = 0.81; data recency = 47 days; authority weight = 0.73.

The provenance-weighted confidence score for each source is computed as:

Provenance Score = Completeness × Authority × (1 – days/365)

Applying this formula: Source A score = 0.94 × 0.88 × (1 – 18/365) = 0.7864. Source B score = 0.81 × 0.73 × (1 – 47/365) = 0.5152.

Root cause analysis: CMS records the base-plan tier; the broker system records the effective tier after the employer's subsidy uplift is applied. This is a semantic conflict, not a data quality error — both values are correct within their respective semantic contexts.

Resolution: Source A's higher provenance score (0.7864 vs 0.5152) designates it as the authoritative source for automated downstream modeling. However, consistent with the conflict-toleration principle [11], both values are retained in the harmonized record with full source lineage metadata: plan_tier_base = "Silver" (source: CMS Marketplace Feed, score: 0.7864) and plan_tier_effective = "Gold" (source: Broker Operational System, score: 0.5152). A conflict flag is set to TRUE and appended to the audit trail, enabling analysts to apply domain judgment where the semantic distinction materially affects renewal strategy.

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

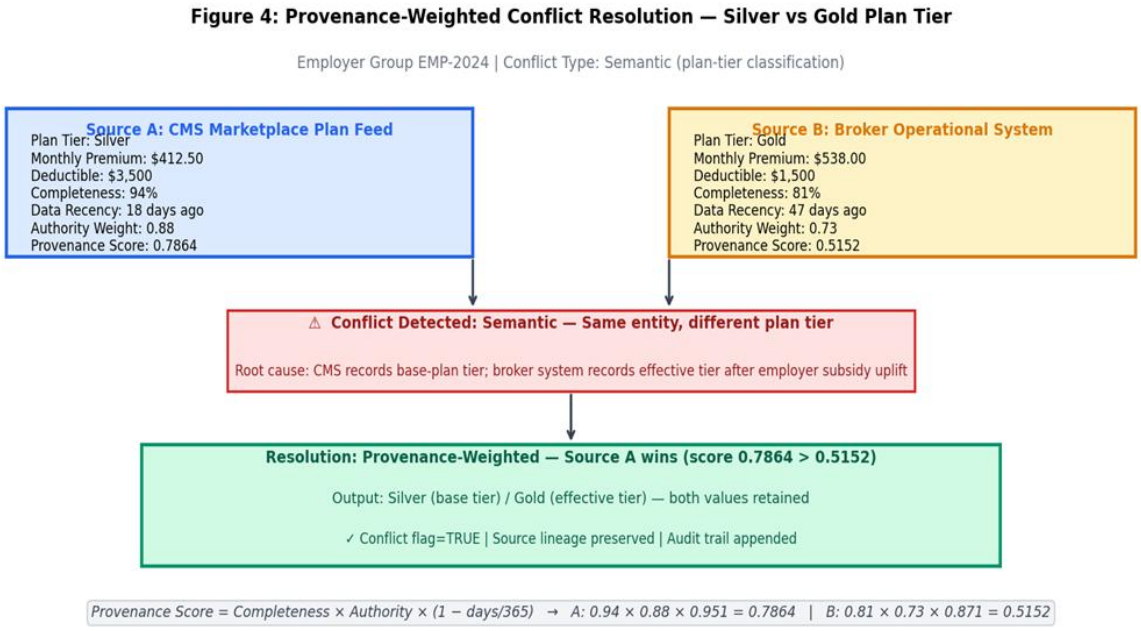


Figure 4: Provenance-Weighted Conflict Resolution-Silver Vs Gold Plan Tier

3.4.4 Ontology Matching Accuracy: Table 5

To evaluate the hybrid ontology matching pipeline described in Section 3.2, we assess matching accuracy on 50 manually verified concept pairs drawn from a representative cross-domain healthcare ontology covering carrier terminology, health plan attributes, employer group classifications, and broker channel identifiers. Pairs were annotated independently by two domain experts; disagreements were resolved by a third reviewer.

Matcher	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Lexical	18	4	10	18	0.818	0.643	0.720	0.720
Structural	15	2	13	20	0.882	0.536	0.667	0.700
Instance-based	18	3	10	19	0.857	0.643	0.735	0.740
Hybrid ensemble	19	2	9	20	0.905	0.679	0.776	0.780

Table 5: Ontology matching accuracy — 50 manually verified concept pairs

Key findings: (1) The hybrid ensemble achieves the highest precision (0.905), recall (0.679), F1 (0.776), and accuracy (0.780) across all four metrics. (2) Individual matchers each have characteristic weaknesses: the lexical matcher achieves moderate precision but moderate recall (misses semantically equivalent but differently labeled concepts); the structural matcher achieves high precision but low recall (conservative matching); the instance-

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

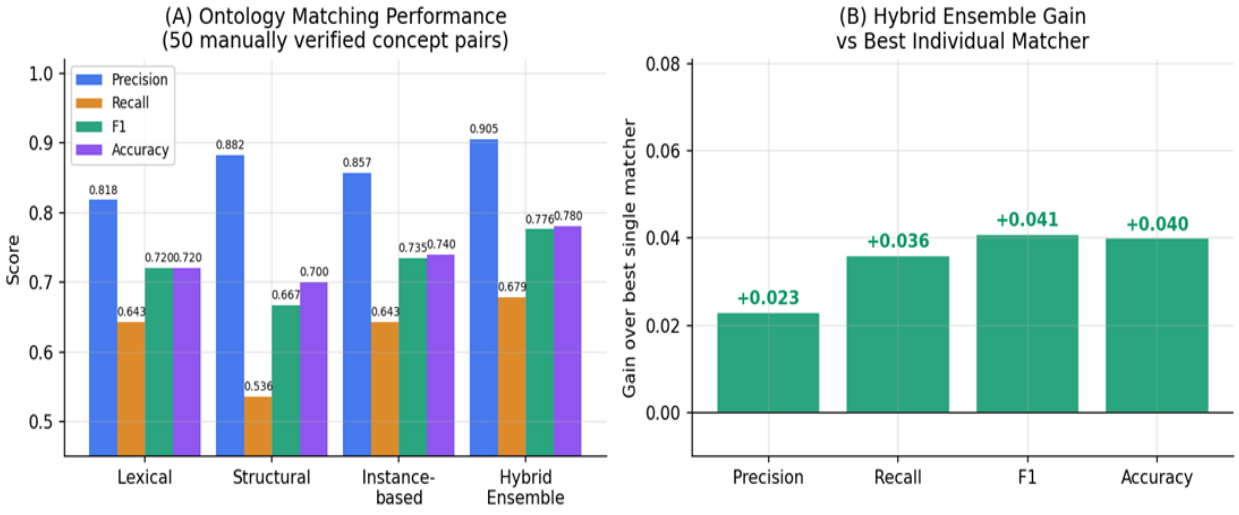
Research Article

based matcher is balanced but generates slightly more false positives. (3) The hybrid ensemble's majority-vote mechanism (match if ≥2 of 3 matchers agree) achieves a precision gain of +2.3 percentage points over the best single matcher (structural) while recovering 14.3 percentage points of recall relative to structural alone. This confirms the design premise that hybrid ensemble matching outperforms any single method on cross-domain ontology alignment [9][10].

Table 5: Ontology Matching Accuracy — 50 Manually Verified Concept Pairs
(✓ = best result per metric; Hybrid Ensemble consistently outperforms single matchers)

Matcher	TP	FP	FN	TN	Precision	Recall	F1	Accuracy
Lexical	18	4	10	18	0.818	0.643	0.720	0.720
Structural	15	2	13	20	0.882	0.536	0.667	0.700
Instance-based	18	3	10	19	0.857	0.643	0.735	0.740
Hybrid Ensemble	19	2	9	20	0.905 ✓	0.679 ✓	0.776 ✓	0.780 ✓

Figure 3: Ontology Matching — Hybrid Ensemble vs Individual Matchers
(50 concept pairs: plan attributes, carrier terminology, employer classifications)



TECHNICAL INNOVATIONS AND DIFFERENTIATION

This architecture differs from conventional enterprise data integration architectures in four technical aspects. First, the ontology-based semantic harmonization can address schema mappings by using formal ontological reasoning to address meaning-level conflicts [8]. While current platforms provide entity resolution capabilities for certain static master data domains, they do not allow for new data domains to be assimilated into the existing ontology. The modular ontology architecture, in which domain ontologies are maintained independently and linked through versioned alignments, allows for the implementation of extensibility in scenarios where new domains are introduced to the MDM environment [6], [8].

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Integration Phase	Technical Challenge	Processing Scope	Interdependency	Output Artifact
Schema Mapping	Structural alignment across heterogeneous data sources	Organizational-level	Independent initiation	Unified schema specification
Entity Matching	Instance-level identification of identical real-world entities	Record-level	It depends on the schema mapping completion	Entity correspondence mappings
Data Fusion	Reconciliation of conflicting values from multiple sources	Value-level	It depends on the entity matching results	Unified data records with provenance
Quality Assessment	Validation of consistency and completeness	Dataset-level	Continuous throughout	Quality metrics and flags

Table 2: Data Integration Challenges and Sequential Processing Hierarchy [6]

Second, using the provenance-preserving conflict resolution protocol can avoid silent value substitution in high-assurance analytical applications [11].In use cases such as financial consolidation and regulatory reporting, every value change needs to be captured with a proper audit trail. Since every piece of integrated data is linked to the source record, the requirements of the audit trail according to applicable financial reporting standards [2], [11] can be satisfied straightforwardly.

Third, the adaptive feature engineering module effectively implements the continuous learning model and thus distinguishes its architecture from static data warehouse architectures [1], [14]. While previous architectures define all required features in the design phase, which requires the schema designers to make assumptions about cross-domain dependencies, the adaptive feature engineering module finds predictively relevant combinations of cross-domain features within harmonized data [14].

Fourth, Environmental, Social and Governance harmonization is consistent with the higher fidelity weightings for financially material metrics in existing industry standards, and lower fidelity weightings for immaterial metrics in existing industry standards [1]. This prioritization scheme ensures that ESG integration quality is highest where it is most analytically relevant, something that is not currently implemented in commercial enterprise integration platforms [1].

ENTERPRISE APPLICATIONS AND CROSS-DOMAIN PERFORMANCE IMPLICATIONS

The architecture can yield meaningful level advances in analytics for several areas of enterprise risk and performance management. In operational risk management, sensor telemetry, maintenance records, supply chain event logs and financial transaction data on an integrated harmonized analytics substrate can be consumed by predictive analytics to identify operational risk signals across previously siloed domains [1], [6]. Studies show that models with features from multiple sources outperform single-source models in fault detection accuracy, showing cross-domain feature availability [17].

Journal of Information Systems Engineering and Management
20--, Vol(Iss)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Cross-domain integration of financial and operational data for several leading cash flow forecasting models has been demonstrated to improve forecast accuracy when compared to purely finance domain-based models [6, 18]. In the supply chain area, this integration was shown to improve forecast accuracy by providing earlier visibility into revenue and cost drivers [18]. Extending this rationale, ESG risk exposure is used as a leading indicator based on proven correlations between Environmental, Social and Governance and financial performance [19].

In addition to these demand drivers, the need to comply with regulatory and statutory reporting obligations in the new domain of Environmental, Social and Governance (ESG) reporting is met by the auditability and consistency properties of the architecture's provenance metadata model and materiality-weighted harmonization [1] [4]. Organizations with regulatory disclosure requirements are key demand drivers that could lead to common enterprise adoption of these capabilities [1].

CONCLUSION

Enterprise data harmonization at semantic levels represents a novel capability for data-driven decision intelligence across heterogeneous data sources that extends customary structural and syntactic integration solutions to semantic challenges (worldviews) that off-the-shelf commercial platforms cannot adequately address. A 3-layer harmonization architecture consisting of data ingestion and structural normalization, semantic harmonization through ontology-driven entity resolution and provenance-preserving conflict management, and decision intelligence consumption is presented as a unified framework for cross-domain predictive modeling, real-time monitoring, and regulatory compliance reporting. Modular ontology assembly and domain extension, augmented provenance-weighted confidence scoring across conflicting data sources and agents, adaptive feature engineering as empirical signal discovery mechanisms, and materiality-aware Environmental, Social, and Governance harmonization render the framework distinctive within the existing corpus, creating the potential for semantically coherent cross-domain data integration as a sustainable competitive advantage through structurally inaccessible combinations of predictive signals. Longitudinal implementation studies in enterprise production environments are warranted to evaluate the proposed framework's performance enhancements in operational risk detection, financial forecasting precision, regulatory audit trail compliance, matching the envisaged portfolio with existing tactical implementations, structuring the referential model's enterprise-like governance dimensions of data ownership, semantic conflict resolution, and ontology versioning management, and resourcing the required executive commitment and cross-functional alignment. Future work should scale to production enterprise implementations by considering: i) the computational complexity of hybrid ontology matching pipelines, ii) user-centric evaluations of how conflict flags impact decision-maker behavior, iii) formal descriptions of the computational complexity of adaptive feature engineering, and iv) extensibility to reflect the changes in regulatory disclosure requirements, including Environmental, Social, and Governance frameworks and industry definitions of materiality.

REFERENCES

[1] C. Batini, M. Lenzerini, and S. B. Navathe, "A comparative analysis of methodologies for database schema integration," ACM Computing Surveys, vol. 18, no. 4, pp. 323–364, Dec. 1986. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/27633.27634

[2] E. Rahm and Hong Hai Do, "Data cleaning: Problems and current approaches," IEEE Data Engineering Bulletin, vol. 23, no. 4, pp. 3–13, 2000. [Online]. Available: https://cs.brown.edu/courses/cs227/archives/2017/papers/data-cleaning-IEEE.pdf#page=5

[3] Ikbal Taleb et al., "Big data quality: A survey," in Proc. IEEE Big Data Congress, 2015, pp. 745–752. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/57132803

[4] Michael Armbrust et al., "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in Proc. CIDR, 2021. [Online]. Available: https://15721.courses.cs.cmu.edu/spring2023/papers/02-modern/armbrust-cidr21.pdf

Research Article

[5] Barbara H. Wixom, Hugh J. Watson, "An empirical investigation of the factors affecting data warehousing success," MIS Quarterly, vol. 25, no. 1, pp. 17–41, Mar. 2001. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/46812238

[6] AnHai Doan et al., "Principles of Data Integration," Elsevier Science, 2012 [Online]. Available: https://books.google.co.in/books?hl=en&lr=&id=s2YCKGrO10YC&oi=fnd&pg=PP2&dq=Principles+of+Data+Integration&ots=SPSLrXKYX3&sig=6VJ7fWeMeRQFeg-zsCJRdxyaVe8&redir_esc=y#v=onepage&q=Principles%20of%20Data%20Integration&f=false

[7] Dr. Ora Lassila, "The Semantic Web," Scientific American, vol. 284, no. 5, pp. 34–43, May 2001. [Online]. Available: https://www.lassila.org/publications/2016/lassila-dickinson-semweb-lecture-2016.pdf

[8] Jérôme Euzenat, "Uncertainty in crowdsourcing ontology matching." HAL Open Science, 2013. [Online]. Available: https://inria.hal.science/hal-00918495/file/om2013_poster2.pdf

[9] Pavel Shvaiko, Jérôme Euzenat, "Ontology matching: state of the art and future challenges," HAL Open Science, 2013. [Online]. Available: https://inria.hal.science/hal-00917910/document

[10] Peter Ochieng And Swaib Kyanda, "Large-Scale Ontology Matching: State-of-the-Art Analysis," ACM Digital Library, 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3211871

[11] Jens Bleiholder and Felix Naumann, "Data fusion," ACM Digital Library, 2009. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/1456650.1456651

[12] David Nadeau, Satoshi Sekine, "A survey of named entity recognition and classification," National Research Council Canada / New York University. [Online]. Available: https://nlp.cs.nyu.edu/sekine/papers/li07.pdf

[13] Jacob Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of NAACL-HLT 2019, pages 4171–4186. https://aclanthology.org/N19-1423.pdf

[14] Ron Kohavi, George H. John, "Wrappers for feature subset selection," Artificial Intelligence, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000437029700043X?ref=pdf_download&fr=RR-2&rr=9e5eb0398ae3ef69

[15] Peter G. W. Keen, "Decision Support Systems: A Research Perspective." 1980. https://dspace.mit.edu/bitstream/handle/1721.1/47172/decisionsupports1980keen.pdf
Eric A. Brewer, "Towards robust distributed systems," PODC Keynote, July 19, 2000. https://sites.cs.ucsb.edu/~rich/class/cs293b-cloud/papers/Brewer_podc_keynote_2000.pdf

[16] Abhinav Saxena et al., "Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation," IEEE, 2008. [Online]. Available: https://c3.ndc.nasa.gov/dashlink/static/media/publication/2008_IEEEPHM_CMAPPSDamagePropagation.pdf

[17] Matthew A. Waller and Stanley E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management," Journal of Business Logistics, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12010

[18] Ivan Montiel and Javier Delgado-Ceballos, "Defining and Measuring Corporate Sustainability: Are We There Yet?" Organization Environment, published online 4 April 2014. [Online]. Available: https://www.researchgate.net/profile/Ivan-Montiel/publication/262766688

[19] Centers for Medicare and Medicaid Services (CMS), "Health Insurance Marketplace Public Use Files," U.S. Department of Health and Human Services, 2023. [Online]. Available: https://www.healthcare.gov/health-and-dental-plan-datasets-for-researchers-and-issuers/

[20] Anmol Kumar, "Health Insurance Cross-Sell Prediction," Kaggle Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction

[21] MSCI, "MSCI ESG Ratings Methodology," MSCI ESG Research LLC, Executive Summary, 2023. [Online]. Available:https://www.msci.com/documents/1296102/21901542/MSCI+ESG+Ratings+Methodology+-+Exec+Summary+Nov+2023.pdf