IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

Volume 12, Issue 10, October 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.317

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

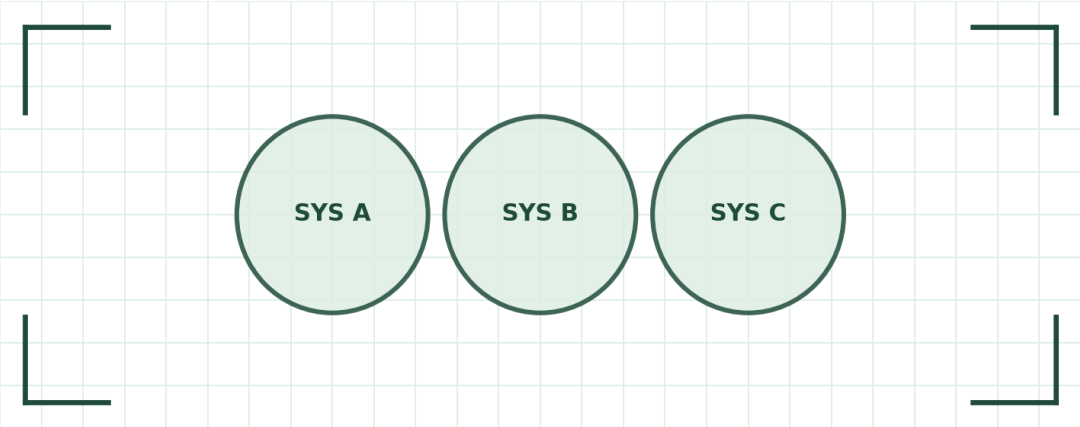
|DOI:10.15662/IJAREEIE.2022.1210002|

Cross-System Identity Reconciliation: Detecting Employee Redundancy in Large-Scale Data Conversion Projects

Ujwal Dyasani

Lead Software Engineer Integrations, USA

ABSTRACT: Large-scale data conversion projects - consolidating employee records from multiple legacy human resource, payroll, and time-tracking systems into a single target platform - routinely surface a structural data-quality problem distinct from simple field-level error: the same individual represented as multiple, only partially overlapping identity records across source systems, none of which agrees perfectly with the others on name formatting, identifier scheme, or biographical detail. Left undetected, this redundancy propagates duplicate employee records into the target system, corrupting downstream payroll, benefits, and reporting processes built on the assumption of a single canonical identity per individual. This article presents a probabilistic record-linkage framework for cross-system identity reconciliation, structured as a five-stage pipeline - extraction, standardization, blocking, similarity scoring, and adjudication - and grounded in the Fellegi-Sunter probabilistic matching model together with modern weighted composite scoring extensions. Drawing on established record-linkage theory, master data management literature, and documented large-scale data migration practice, the article develops a comparative analysis of matching algorithm classes - exact key matching, deterministic rule-based matching, probabilistic matching, and weighted composite scoring - and illustrates the reconciliation pipeline's effect on record volume and identity cluster formation through a series of diagrams and tables. The findings indicate that composite scoring approaches combining multiple weighted similarity features substantially outperform single-criterion matching strategies in illustrative redundancy detection rate, while requiring materially greater design and tuning investment. The article further addresses adjudication workflow design, false-match risk management, and governance practices for sustaining identity reconciliation quality across multi-phase conversion programs, concluding with recommendations for organizations planning large-scale employee data consolidation efforts.



KEYWORDS: identity reconciliation; record linkage; data conversion; master data management; entity resolution; duplicate detection; probabilistic matching; Fellegi-Sunter model; data migration; employee data quality.

I. INTRODUCTION

Every large-scale data conversion project that consolidates employee records from multiple legacy systems into a single target platform confronts the same underlying question, regardless of the specific systems involved: how many distinct individuals are actually represented across the source data, and which records, across which systems, belong to the same individual? This question is deceptively difficult to answer at scale. A single employee may appear in a legacy

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

human resources system under one name spelling, in a legacy payroll system under a slightly different spelling introduced by a data entry error years earlier, and in a newly staged conversion environment under a third variant produced by an automated name-parsing routine that mishandled a compound surname. None of these three records shares an identical, directly comparable identifier, yet all three describe the same person.

This class of problem - commonly termed record linkage or entity resolution in the data management literature - has been studied extensively in domains such as public health record matching, census data integration, and customer data deduplication (Fellegi & Sunter, 1969; Christen, 2012). Its application to employee identity reconciliation during enterprise data conversion projects shares the same theoretical foundation but carries distinct practical stakes: an undetected duplicate employee record in a converted human capital management (HCM) tenant does not merely create reporting noise, it risks generating duplicate payroll disbursement, inconsistent benefits eligibility determination, and compliance reporting inaccuracy, each of which can be costly and disruptive to correct after a conversion has already gone live (Loshin, 2011; DAMA International, 2017).

This article develops a probabilistic record-linkage framework specifically for cross-system employee identity reconciliation in the context of large-scale data conversion projects, structured as a five-stage pipeline and grounded in established record-linkage theory. The research questions guiding this study are as follows:

- RQ1: What pipeline structure best organizes the extraction, standardization, candidate identification, similarity scoring, and adjudication stages required for reliable cross-system employee identity reconciliation?
- RQ2: How do different classes of matching algorithms - exact key matching, deterministic rule-based matching, probabilistic matching, and weighted composite scoring - compare in illustrative redundancy detection performance?
- RQ3: How should match adjudication and false-match risk management be designed to balance detection completeness against the operational cost of incorrect merge decisions?
- RQ4: What governance practices sustain identity reconciliation quality across multi-phase conversion programs spanning multiple waves or business units?

The remainder of this article proceeds as follows. Section 2 reviews related literature on record linkage and master data management. Section 3 examines why identity fragmentation arises specifically in the data conversion context. Section 4 details the research methodology. Section 5 presents the five-stage reconciliation pipeline, with Sections 6 through 8 detailing its standardization, blocking, and similarity scoring components. Section 9 presents comparative matching algorithm analysis. Sections 10 through 12 address adjudication, false-match risk, and a worked volume-impact illustration. Sections 13 through 18 address governance, pitfalls, limitations, recommendations, future research, and conclusions.

II. BACKGROUND AND RELATED WORK

2.1 Foundational Record-Linkage Theory

The formal statistical foundation for probabilistic record linkage originates with Fellegi and Sunter's seminal model, which frames record matching as a binary classification problem: for each candidate record pair, compute the likelihood ratio between the probability of observing the pair's agreement pattern under a true-match hypothesis versus a true-non-match hypothesis, and classify the pair as a match, a non-match, or an uncertain case requiring manual review based on where that likelihood ratio falls relative to two calibrated decision thresholds (Fellegi & Sunter, 1969). This model remains the theoretical basis for the similarity-scoring and adjudication framework developed in Sections 8 and 10 of this article.

2.2 Modern Entity Resolution and Data Matching Practice

Subsequent data-matching literature has extended the Fellegi-Sunter model with practical techniques for scaling record linkage to large datasets, most notably blocking - partitioning records into smaller candidate groups sharing an approximate characteristic, so that full pairwise comparison is performed only within blocks rather than across the entire dataset - and a wide range of string-similarity metrics suited to name and address comparison specifically (Christen, 2012). This literature also documents the practical trade-off between detection completeness (recall) and false-match risk (precision) that any matching threshold configuration must navigate, a trade-off examined directly in Section 11 of this article.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

2.3 Master Data Management and Data Migration Literature

Master data management (MDM) literature situates entity resolution within a broader data governance discipline concerned with establishing and maintaining a single, trusted view of core business entities - customers, products, and, in the HCM context, employees - across an organization's systems (DAMA International, 2017; Loshin, 2011). Data migration and conversion-specific guidance further emphasizes that identity reconciliation is most effectively performed as a distinct, structured project phase preceding data load into the target system, rather than as an incidental cleanup activity performed reactively after duplicate records have already been created in production (Deloitte, 2019; Accenture, 2020).

2.4 Positioning of This Study

While Fellegi-Sunter theory and subsequent entity-resolution literature provide a strong general foundation, and MDM literature situates the problem organizationally, comparatively little published material synthesizes these bodies of work into a structured, pipeline-oriented framework specifically tailored to the employee-identity reconciliation problem as it arises in large-scale HCM data conversion projects. This article's contribution is that synthesis, together with an illustrative comparative analysis of matching algorithm classes and a governance framework for multi-phase conversion programs.

III. THE DATA CONVERSION CONTEXT: WHY IDENTITY FRAGMENTS

Before presenting the reconciliation framework itself, it is useful to examine why employee identity fragmentation arises so consistently in data conversion projects specifically. Table 1 catalogs the principal root causes observed across conversion project patterns documented in the literature (Deloitte, 2019; Accenture, 2020; Christen, 2012).

Table.1. Principal root causes of employee identity fragmentation in data conversion source systems.

Root Cause	Description	Illustrative Example
Independent Identifier Schemes	Each source system assigns its own internal identifier with no cross-system mapping	Employee ID 10432 in the legacy HR system corresponds to Payroll Number PR-88291 in the legacy payroll system, with no stored cross-reference
Name Formatting Variation	The same name is captured with different formatting, ordering, or transliteration conventions across systems	"O'Brien, Mary-Jane" in one system versus "MARYJANE OBRIEN" in another, differing in punctuation, hyphenation, and case
Data Entry Error Accumulation	Legacy systems accumulate uncorrected data entry errors over years of independent operation	A transposed digit in a national identifier introduced years prior and never corrected
Life-Event Data Changes	An individual's recorded attributes change over time (marriage-related name change, address change) and are updated inconsistently across systems	A name change following marriage reflected in the HR system but not yet propagated to the payroll system at the time of extraction
Rehire and Re-Engagement Records	An individual leaves and later rejoins the organization, generating a new record rather than reactivating the original	A rehired employee assigned a new employee ID rather than having their original historical record reactivated
Multi-System Concurrent Employment	An individual holds concurrent appointments captured separately in systems not designed to recognize the overlap	An adjunct instructor also employed in a staff role, recorded as two unrelated records in source systems organized by employment type

None of these root causes reflects a data-quality defect that can be corrected simply by improving source-system data entry practice going forward; each reflects an inherent characteristic of independently operated legacy systems accumulated over years or decades, which is precisely why structured reconciliation, rather than ad hoc cleanup, is required at the point of conversion.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

IV. RESEARCH METHODOLOGY

This study employs a design-framework and illustrative-comparative-analysis methodology, synthesizing foundational record-linkage theory (Fellegi & Sunter, 1969; Christen, 2012), master data management literature (DAMA International, 2017; Loshin, 2011), and documented data conversion practice guidance (Deloitte, 2019; Accenture, 2020) into the five-stage reconciliation pipeline presented in Section 5. The comparative algorithm-performance analysis in Section 9 and the volume-impact illustration in Section 12 use illustrative figures constructed to demonstrate expected directional relationships consistent with established record-linkage theory, rather than audited measurements drawn from a single production conversion project.

Table.2. Summary of research methodology components.

Methodology Component	Method	Purpose
Root-Cause Taxonomy Development	Synthesis of data migration and record-linkage literature	Establish structured understanding of identity fragmentation causes (Section 3)
Pipeline Framework Development	Synthesis of Fellegi-Sunter theory and MDM practice guidance	Establish the five-stage reconciliation pipeline (Section 5)
Comparative Algorithm Illustration	Directional estimate construction consistent with record-linkage theory	Illustrate expected detection-rate differences across algorithm classes (Section 9)
Volume-Impact Illustration	Composite scenario construction reflecting typical conversion project scale	Illustrate record-volume reduction across pipeline stages (Section 12)

V. THE FIVE-STAGE RECONCILIATION PIPELINE

This article proposes a five-stage reconciliation pipeline - extraction, standardization, blocking, similarity scoring, and adjudication - structuring how raw, heterogeneous employee records from multiple source systems are progressively transformed into a consolidated set of resolved identities.

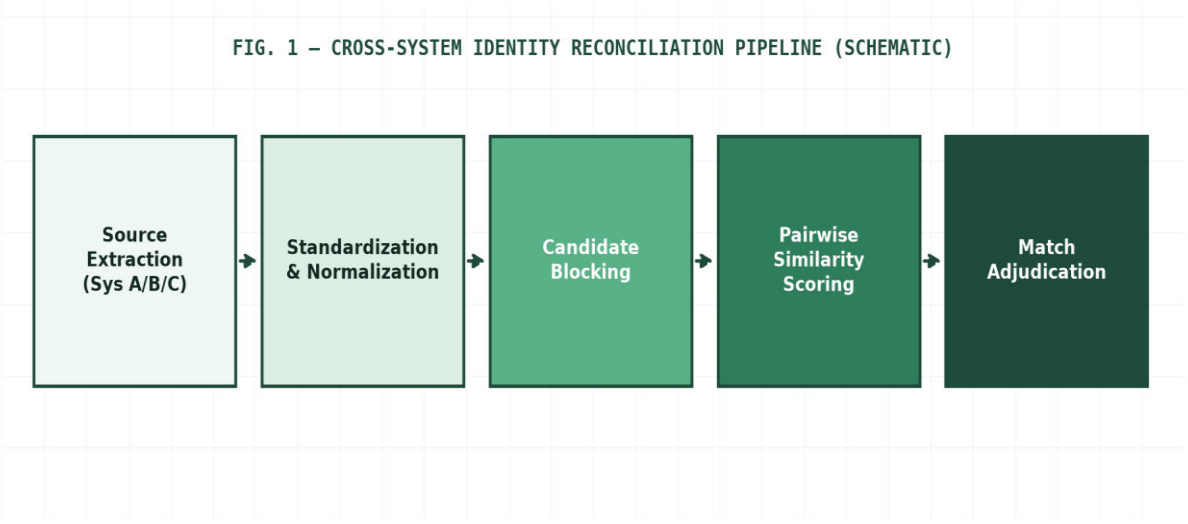


Figure1. The cross-system identity reconciliation pipeline: extraction through standardization, blocking, similarity scoring, and match adjudication.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.3. Overview of the five-stage cross-system identity reconciliation pipeline.

Stage	Objective	Primary Technique	Output
1. Source Extraction	Retrieve employee records from all in-scope source systems	Bulk extract via integration or direct query	Raw combined record set
2. Standardization	Normalize formatting inconsistencies across records	Field parsing, case normalization, transliteration handling	Standardized record set (Section 6)
3. Candidate Blocking	Reduce the comparison space to plausible candidate pairs	Blocking keys (e.g., birth year + first-initial + last-name-soundex)	Candidate pair set (Section 7)
4. Similarity Scoring	Quantify the likelihood that each candidate pair represents the same individual	Weighted composite or probabilistic scoring model	Scored candidate pairs (Section 8)
5. Match Adjudication	Convert scored pairs into final match/non-match/review decisions and resolved identity clusters	Threshold classification plus manual review for uncertain cases	Resolved identity cluster set (Section 10)

As with the layered validation approach common in data-quality engineering more broadly, this pipeline is deliberately structured so that computationally cheaper stages (standardization, blocking) narrow the problem space before the more computationally and analytically expensive stages (similarity scoring, adjudication) are applied, an efficiency consideration that becomes material at the scale typical of enterprise employee populations spanning tens or hundreds of thousands of records across multiple source systems.

VI. STANDARDIZATION AND NORMALIZATION TECHNIQUES

Standardization resolves formatting inconsistency before any matching logic is applied, ensuring that downstream similarity comparisons evaluate genuine content differences rather than superficial formatting artifacts.

Table.4. Standardization and normalization techniques applied prior to candidate blocking.

Standardization Technique	Applies To	Example Transformation
Case Normalization	Name fields, text identifiers	"MARY O'BRIEN" and "mary o'brien" normalized to a consistent case convention
Punctuation and Whitespace Normalization	Name fields, address fields	"O'Brien" and "OBrien" normalized to a consistent punctuation handling rule
Name Component Parsing	Full-name fields lacking discrete first/last name fields	"Smith, John A." parsed into discrete surname, given name, and middle-initial components
Date Format Normalization	Date-of-birth, hire-date fields	Mixed MM/DD/YYYY and DD-MM-YYYY formats normalized to a single ISO-standard format
Identifier Format Normalization	National identifiers, employee ID fields	Identifiers with inconsistent leading zeros or embedded formatting characters normalized to a canonical form
Phonetic Encoding	Surname fields, for use as a blocking key component	Surnames encoded via a phonetic algorithm (e.g., Soundex) to group likely variants sharing a common pronunciation pattern

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Phonetic encoding merits particular attention because it serves a dual purpose: it is both a standardization technique in its own right and, as discussed in Section 7, a common foundation for constructing blocking keys, since phonetically similar surnames are a strong signal of a plausible name variant rather than a genuinely different individual.

VII. CANDIDATE BLOCKING STRATEGIES

Comparing every record in one source system against every record in every other source system scales quadratically with record volume and quickly becomes computationally impractical at enterprise scale. Blocking addresses this by partitioning records into smaller groups sharing an approximate characteristic, restricting full similarity comparison to within-block pairs only (Christen, 2012).

Table.5. Candidate blocking strategies and associated trade-off considerations

Blocking Key Strategy	Construction Method	Trade-Off Consideration
Exact Attribute Blocking	Group records sharing an identical attribute value (e.g., birth year)	Simple and fast, but misses matches where the blocking attribute itself contains an error
Phonetic Surname Blocking	Group records sharing a phonetically encoded surname	Tolerant of spelling variation, but can produce large blocks for common surname sounds
Composite Multi-Field Blocking	Group records sharing a combination of attributes (e.g., birth year + first-initial + phonetic surname)	Reduces block size relative to single-field blocking, at the cost of missing matches where any one component field is erroneous
Sorted Neighborhood Blocking	Sort records by a key and compare only records within a defined window of the sorted order	Effective for approximate key matching without requiring exact grouping, at some risk of missing distant true matches

The composite multi-field blocking approach is generally preferred in enterprise employee reconciliation contexts because it balances block-size reduction against tolerance for a single erroneous field, but organizations should recognize that any blocking strategy inherently trades some recall for computational tractability, and should periodically sample blocked-out record pairs to confirm the blocking strategy is not systematically excluding a class of true matches.

VIII. SIMILARITY SCORING MODELS

Within each candidate block, similarity scoring quantifies how closely two records resemble one another across multiple comparison fields, producing a composite score used for adjudication in Section 10. Table 6 documents representative comparison fields and the similarity metric typically applied to each.

Table.6. Representative comparison fields and similarity metrics used in composite scoring.

Comparison Field	Representative Similarity Metric	Rationale for Metric Choice
Surname	Levenshtein edit distance; phonetic match indicator	Tolerant of minor spelling variation and transliteration difference
Given Name	Levenshtein edit distance; nickname/alias lookup	Accounts for common nickname substitution (e.g., "Robert" vs. "Bob")
Date of Birth	Exact match indicator; near-match indicator for transposed digits	Date fields are highly discriminating when exact, but prone to transposition error
National Identifier (partial)	Exact match on available partial digits, where full identifier is restricted	High discriminating power where available, subject to field-masking constraints
Address	Token-based similarity (e.g., Jaccard similarity on address tokens)	Tolerant of formatting variation across address components
Hire Date / Employment Dates	Proximity-based similarity within a defined tolerance window	Distinguishes concurrent employment scenarios from coincidental attribute overlap

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

The Fellegi-Sunter model formalizes how these individual field-level similarity indicators should be combined into a single composite likelihood ratio, weighting each field's contribution according to its empirically observed discriminating power - fields that rarely agree by coincidence (such as a matching date of birth) receive substantially higher weight than fields with high coincidental agreement rates (such as a matching birth year alone) (Fellegi & Sunter, 1969).

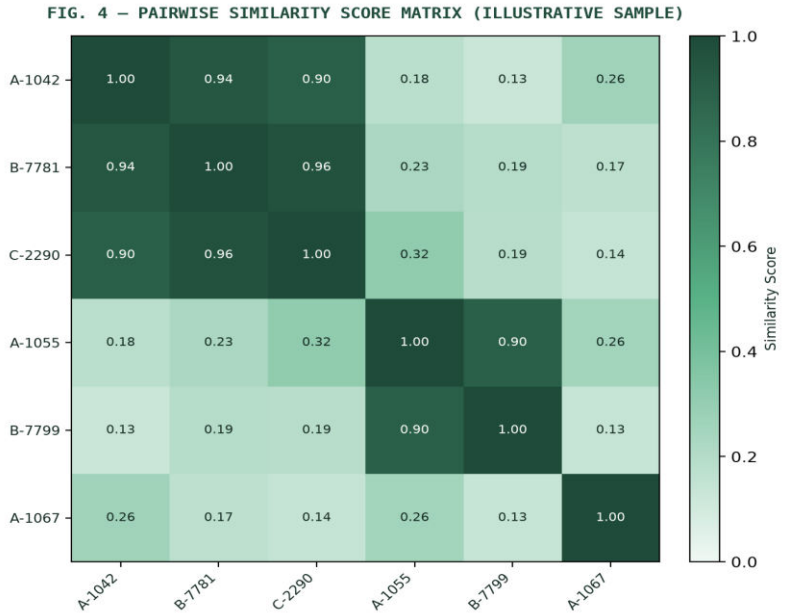


Figure.2. Illustrative pairwise similarity score matrix for a small sample of candidate record pairs, showing high-similarity clusters against low-similarity background pairs

IX. COMPARATIVE ANALYSIS OF MATCHING ALGORITHM CLASSES

This section presents an illustrative comparative analysis of four matching algorithm classes, constructed to demonstrate the expected directional relationship between algorithm sophistication and redundancy detection performance discussed qualitatively throughout Sections 5 through 8.

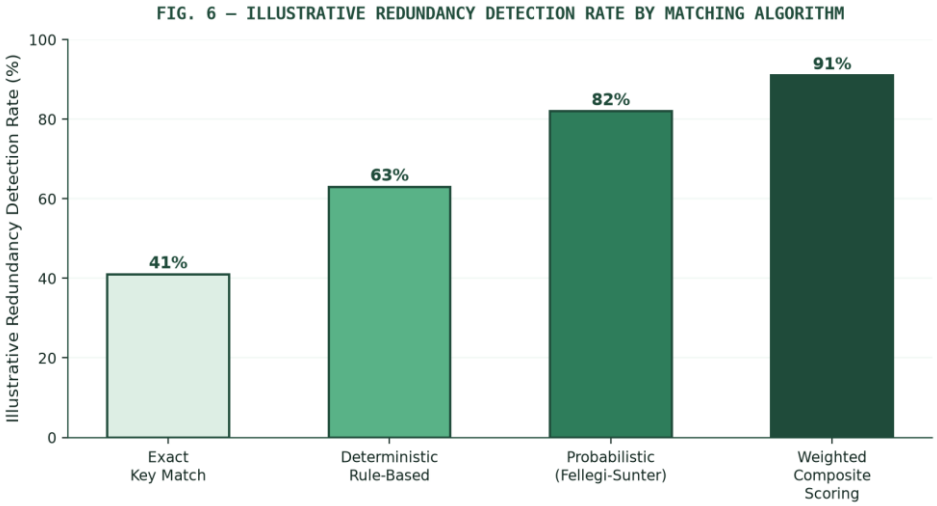


Figure.3. Illustrative redundancy detection rate by matching algorithm class, showing progressive improvement from exact key matching through weighted composite scoring

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.7.Illustrative comparative redundancy detection rate and implementation complexity across matching algorithm classes

Algorithm Class	Description	Illustrative Detection Rate	Relative Implementation Complexity
Exact Key Matching	Matches records only where a designated key field is identical across systems	41%	Low
Deterministic Rule-Based Matching	Matches records satisfying a fixed set of explicit comparison rules (e.g., exact surname AND exact birth date)	63%	Low-Moderate
Probabilistic Matching (Fellegi-Sunter)	Computes a weighted likelihood ratio across multiple comparison fields per the Fellegi-Sunter model	82%	Moderate-High
Weighted Composite Scoring	Extends probabilistic matching with additional engineered features and empirically tuned field weights	91%	High

The substantial detection-rate improvement from exact key matching (41%) to weighted composite scoring (91%) illustrated in Table 7 and Figure 3 reflects a core insight from record-linkage theory: identity fragmentation in a data conversion context rarely presents as a single, cleanly matching key field, and matching strategies that rely on a single comparison criterion systematically miss the substantial population of true matches that agree closely, but not perfectly, across several fields simultaneously. This does not imply that exact key matching has no place in the pipeline; rather, it is typically retained as a fast, high-confidence first pass (per the blocking discussion in Section 7) before more sophisticated scoring is applied to the remaining candidate population.

X. MATCH ADJUDICATION AND CLUSTER FORMATION

Once candidate pairs have been scored (Section 8), adjudication converts pairwise scores into final match decisions and, where more than two records are involved, resolves those decisions into consolidated identity clusters - groups of records across source systems determined to represent the same individual.

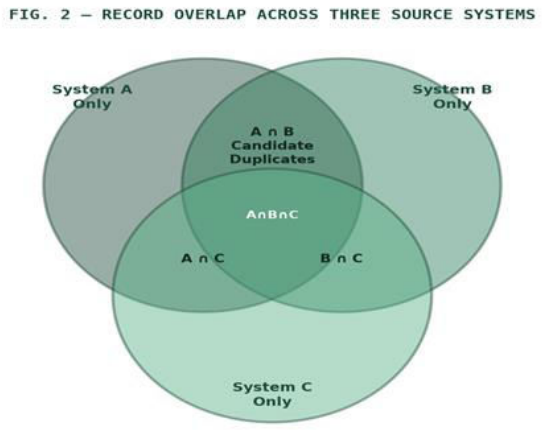


Figure.4. Illustrative record overlap across three source systems, showing distinct regions corresponding to system-exclusive records and candidate duplicate overlap zones

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

The Fellegi-Sunter model formally distinguishes three decision regions based on the computed likelihood ratio: a high-confidence match region, a high-confidence non-match region, and an intermediate "clerical review" region for pairs whose score is too ambiguous for automated classification (Fellegi & Sunter, 1969). Table 8 documents how this three-region model is typically operationalized for employee identity reconciliation.

Table.8. Three-region adjudication model for converting similarity scores into match decisions.

Decision Region	Illustrative Score Range	Automated Action	Human Involvement
High-Confidence Match	Above upper threshold (e.g., ≥ 0.90)	Automatically merge into a single resolved identity cluster	None required; periodic audit sampling recommended
Clerical Review	Between lower and upper thresholds (e.g., 0.60 - 0.90)	Route to manual adjudication queue	Required; reviewer confirms or rejects the candidate match
High-Confidence Non-Match	Below lower threshold (e.g., < 0.60)	Retain as distinct, unmatched records	None required; periodic audit sampling recommended

When more than two records across more than two systems are linked through pairwise match decisions - for example, a System A record matched to a System B record, and that same System B record separately matched to a System C record - these pairwise decisions must be resolved into a single connected cluster representing one individual, rather than left as a set of disconnected pairwise links.

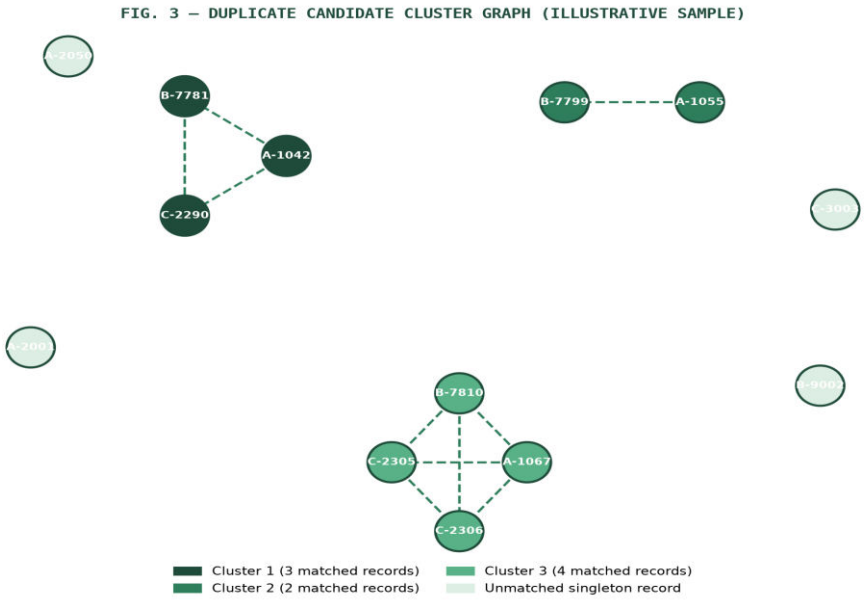


Figure.5. Illustrative duplicate candidate cluster graph, showing connected components of matched records forming resolved identity clusters, alongside unmatched singleton records.

Cluster formation techniques - commonly implemented via connected-component analysis over the graph of pairwise match decisions, as illustrated in Figure 5 - must additionally guard against a specific failure mode known as transitive over-merging, in which a chain of individually plausible pairwise matches (A matches B, B matches C) produces a merged cluster (A, B, C) even though A and C, compared directly, would not themselves score as a plausible match. Table 9 documents recommended safeguards against this failure mode.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.9. Recommended safeguards against transitive over-merging during cluster formation.

Safeguard	Description	Trade-Off
Direct Pairwise Confirmation	Require every pair within a proposed cluster, not just chain links, to independently exceed a minimum similarity threshold	Increases adjudication rigor at the cost of additional pairwise comparisons within larger clusters
Cluster Size Cap with Manual Review	Route any proposed cluster exceeding a defined size (e.g., more than 3 records) to mandatory manual review	Adds review burden but limits automated risk from unusually large clusters
Minimum Cluster Cohesion Score	Require an aggregate cohesion metric across all pairs in a cluster to exceed a threshold before automated merge	Requires additional scoring logic beyond simple pairwise threshold checks

XI. FALSE-MATCH RISK AND THRESHOLD CALIBRATION

Threshold calibration - setting the score boundaries separating the three decision regions in Section 10 - directly determines the balance between detection completeness (recall) and false-match risk (precision), a core trade-off in record-linkage theory (Christen, 2012; Fellegi & Sunter, 1969).

FIG. 5 — DISTRIBUTION OF MATCH CONFIDENCE SCORES ACROSS CANDIDATE PAIRS

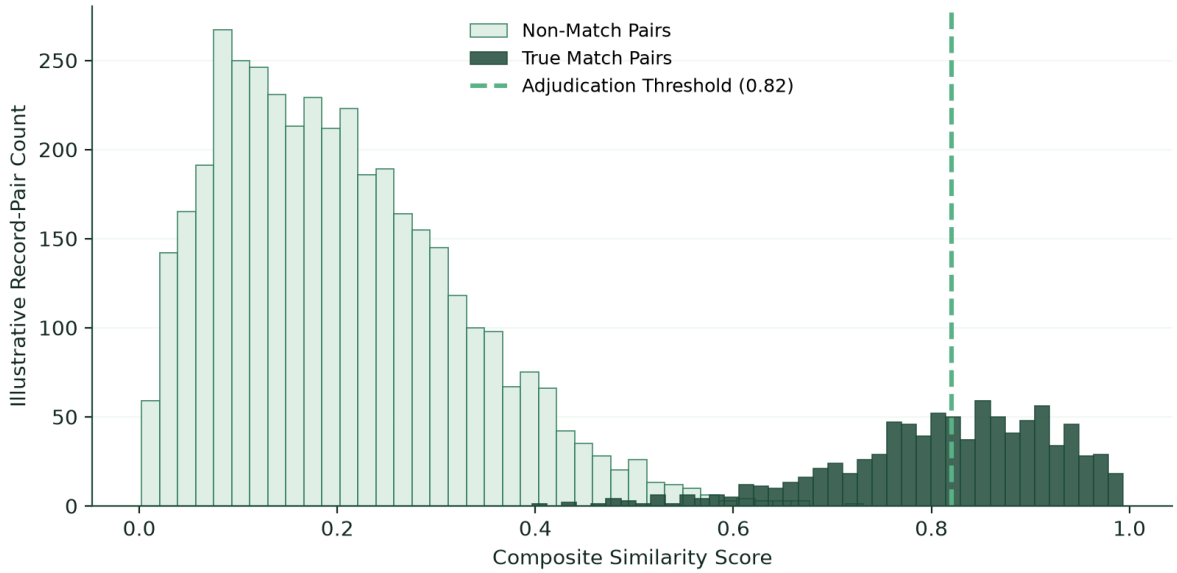


Figure.6. Illustrative distribution of composite similarity scores across true-match and non-match candidate pairs, with an adjudication threshold positioned in the region of maximal separation.

As Figure 6 illustrates, true-match and non-match score distributions typically overlap to some degree even under a well-designed scoring model, meaning that no single threshold eliminates both false positives (incorrectly merging distinct individuals) and false negatives (failing to merge records that represent the same individual) simultaneously. Because the operational cost of these two error types differs substantially in an employee identity reconciliation context, threshold calibration should be informed by that asymmetry rather than set purely to maximize an undifferentiated accuracy metric.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.10.Comparative severity of false-positive versus false-negative errors in employee identity reconciliation

Error Type	Description	Illustrative Downstream Consequence	Relative Severity
False Positive (Incorrect Merge)	Two records representing distinct individuals are incorrectly merged into a single identity	Potential compensation, benefits, or employment-history data corruption affecting two real individuals	High - difficult and costly to unwind post-conversion
False Negative (Missed Match)	Two records representing the same individual are not merged and remain as separate identities	A duplicate record persists in the target system, requiring downstream reconciliation	Moderate - generally correctable via post-conversion cleanup without affecting a second individual's data

Because Table 10 indicates that false positives generally carry higher downstream severity than false negatives in this context, most documented conversion project practice favors a conservative upper threshold - accepting a somewhat higher clerical review workload in the intermediate region (Section 10) in exchange for lower risk of an automated incorrect merge - rather than optimizing purely for minimizing total review volume (Deloitte, 2019; Accenture, 2020).

XII. PIPELINE VOLUME IMPACT: A WORKED ILLUSTRATION

This section presents a worked, illustrative volume-impact scenario tracing a representative record population through each stage of the five-stage pipeline (Section 5), to demonstrate the cumulative effect of standardization, blocking, scoring, and adjudication on the raw combined record count.

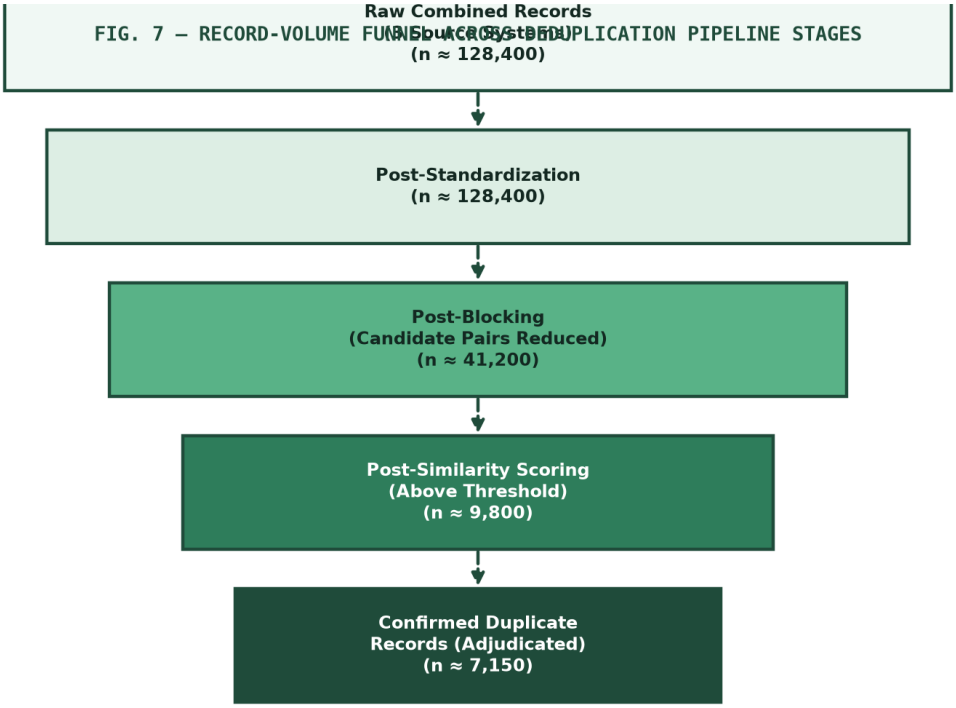


FIG. 7 – RECORD-VOLUME FUNNEL ACROSS DEDUPLICATION PIPELINE STAGES

Figure.7. Illustrative record-volume funnel across deduplication pipeline stages, from the raw combined record set through confirmed, adjudicated duplicate records.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.11. Illustrative record and candidate-pair volume across the five-stage reconciliation pipeline

Pipeline Stage	Illustrative Record/Pair Count	Cumulative Reduction from Prior Stage	Notes
Raw Combined Records (3 Source Systems)	128,400	N/A (starting volume)	Sum of all records extracted across three source systems
Post-Standardization	128,400	0% (no record elimination at this stage)	Record count unchanged; field content normalized
Post-Blocking (Candidate Pairs)	41,200 candidate pairs	N/A (pair count, not record count)	Full cross-product pairwise comparison would have exceeded several billion pairs absent blocking
Post-Similarity Scoring (Above Review Threshold)	9,800 candidate pairs	76.2% reduction from candidate pairs	Pairs scoring below the clerical review threshold are excluded from further adjudication
Confirmed Duplicate Records (Adjudicated)	7,150 records	27.0% reduction from above-threshold pairs	Reflects both automated high-confidence merges and manually confirmed clerical review matches

The blocking stage's effect in Table 11 is worth emphasizing quantitatively: absent blocking, a full pairwise comparison of 128,400 records would require evaluating on the order of 8.2 billion candidate pairs, a computationally prohibitive volume for the similarity-scoring stage; blocking reduces this to approximately 41,200 plausible candidate pairs, a reduction of more than five orders of magnitude, illustrating why blocking is treated as a structurally necessary stage rather than an optional optimization in any reconciliation pipeline operating at enterprise employee-population scale.

XIII. GOVERNANCE FOR MULTI-PHASE CONVERSION PROGRAMS

Many large-scale conversion programs proceed in multiple waves - by business unit, by geography, or by legacy system retirement schedule - rather than as a single, simultaneous cutover. This staged structure introduces a governance requirement beyond the single-pass pipeline described in Section 5: identity reconciliation performed in an early wave must remain consistent with reconciliation performed in later waves, so that an individual identified as a single resolved identity in Wave 1 is not inadvertently treated as a new, distinct identity when a related record surfaces in Wave 2.

Table.12. Recommended governance practices for sustaining reconciliation consistency across multi-phase conversion programs.

Governance Practice	Purpose	Recommended Application Point
Persistent Cross-Reference Key Registry	Maintain a durable mapping from each resolved identity cluster to all constituent source-system record identifiers, across waves	Established at Wave 1; updated at every subsequent wave
Cross-Wave Candidate Re-Screening	Re-run blocking and scoring for newly extracted records against the full existing cross-reference registry, not only against other new records	Every subsequent wave
Threshold Consistency Review	Confirm adjudication thresholds (Section 11) remain consistent across waves, or document and justify any deliberate change	Prior to each wave's adjudication stage
Adjudication Decision Audit Log	Retain a record of every clerical review decision and its rationale, for consistency reference in later waves	Ongoing, every wave
Post-Wave Reconciliation Retrospective	Review false-positive and false-negative discoveries surfaced after each wave's go-live, feeding lessons into subsequent wave tuning	After each wave's stabilization period

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

The persistent cross-reference key registry is the single most consequential governance artifact in this list: without it, each wave's reconciliation effort effectively starts from zero prior knowledge, discarding the resolved-identity determinations painstakingly established in earlier waves and risking inconsistent treatment of the same individual across the program's timeline.

XIV. COMMON PITFALLS IN IDENTITY RECONCILIATION

This section catalogs recurring pitfalls observed across identity reconciliation efforts, synthesized from the framework and governance guidance presented in Sections 5 through 13.

Table.13. Recurring pitfalls in cross-system identity reconciliation and recommended corrections.

Pitfall	Description	Recommended Correction
Single-Criterion Matching Reliance	Relying solely on exact key matching, missing the substantial population of true matches with imperfect key agreement	Adopt composite scoring per Section 9
Over-Aggressive Blocking	Blocking keys constructed too narrowly, systematically excluding true matches with an error in the blocking field itself	Periodically sample blocked-out pairs; consider composite multi-field blocking (Section 7)
Uniform Threshold Across Heterogeneous Fields	Applying the same similarity threshold to fields with very different discriminating power (e.g., birth year vs. full date of birth)	Apply field-specific weighting consistent with the Fellegi-Sunter model (Section 8)
Unmanaged Transitive Over-Merging	Cluster formation logic merges records through indirect chains without direct pairwise confirmation	Apply the safeguards documented in Table 9 (Section 10)
No Cross-Wave Consistency Mechanism	Each conversion wave performs reconciliation independently, without reference to prior waves' resolved identities	Establish a persistent cross-reference registry (Section 13)
Threshold Calibration Without Error-Cost Analysis	Thresholds set to minimize total review volume without considering the asymmetric cost of false positives versus false negatives	Calibrate thresholds informed by the error-severity analysis in Section 11

XV. LIMITATIONS OF THE STUDY

Several limitations should be considered when interpreting this article's findings. First, the comparative detection-rate figures in Section 9 and the volume-impact figures in Section 12 are illustrative estimates constructed to demonstrate directional relationships consistent with established record-linkage theory, rather than measurements audited against a specific production conversion project; actual figures will vary considerably based on source data quality, field availability, and algorithm tuning maturity. Second, this study addresses employee identity reconciliation specifically within the HCM data conversion context and does not directly examine how the framework generalizes to other entity types (customer, vendor, patient) that may carry different field-availability and discriminating-power characteristics. Third, the governance practices proposed in Section 13 for multi-phase programs reflect synthesized best practice rather than outcomes empirically tracked across multiple organizations' multi-wave conversion programs over time. Fourth, this article does not address emerging machine-learning-based entity resolution techniques in depth, focusing instead on the well-established Fellegi-Sunter probabilistic model and its composite-scoring extensions.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Table.14. Summary of study limitations and suggested mitigations for future research.

Limitation	Potential Effect on Findings	Suggested Mitigation for Future Studies
Illustrative (not audited) performance and volume figures	Absolute figures may not generalize across organizations or data quality profiles	Controlled empirical study across multiple production conversion projects
Employee-entity-specific scope	Findings may not directly generalize to other entity resolution domains	Comparative study across entity types with differing field-availability profiles
Unvalidated multi-phase governance outcomes	Governance effectiveness not empirically confirmed over time	Longitudinal tracking of multi-wave programs adopting the governance framework
Limited treatment of machine-learning-based methods	Emerging technique performance not directly compared	Comparative study of classical probabilistic versus ML-based entity resolution methods

XVI. RECOMMENDATIONS

Synthesizing the framework and findings presented above, this section consolidates recommended practices for organizations planning cross-system employee identity reconciliation as part of a large-scale data conversion project.

- Adopt the full five-stage pipeline - extraction, standardization, blocking, similarity scoring, and adjudication - rather than relying on exact key matching alone, given the substantial detection-rate improvement illustrated in Section 9.
- Invest specifically in standardization quality (Section 6) before blocking and scoring, since downstream matching accuracy is bounded by the consistency of the standardized input data.
- Use composite, multi-field blocking keys rather than single-field blocking, and periodically sample blocked-out pairs to confirm the blocking strategy is not systematically excluding true matches.
- Calibrate adjudication thresholds with explicit reference to the asymmetric cost of false positives versus false negatives (Section 11), generally favoring a conservative upper threshold given the higher downstream severity of incorrect automated merges.
- Apply direct pairwise confirmation or cluster-cohesion safeguards during cluster formation to prevent transitive over-merging (Section 10).
- Establish a persistent cross-reference key registry from the first conversion wave onward, to sustain reconciliation consistency across multi-phase programs (Section 13).
- Conduct a post-wave reconciliation retrospective after each conversion wave's stabilization period, feeding false-positive and false-negative discoveries back into threshold and rule tuning for subsequent waves.

Table.15. Prioritized recommendation summary mapped to affected pipeline stage.

Recommendation	Primary Pipeline Stage Affected	Priority
Adopt full five-stage pipeline	All Stages	Critical
Invest in standardization quality	Stage 2: Standardization	High
Composite multi-field blocking with periodic sampling	Stage 3: Blocking	High
Asymmetric-cost-informed threshold calibration	Stage 5: Adjudication	Critical
Cluster formation safeguards	Stage 5: Adjudication	High
Persistent cross-reference registry	Program Governance	Critical
Post-wave reconciliation retrospective	Program Governance	Medium-High

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

XVII. FUTURE RESEARCH DIRECTIONS

This study suggests several directions for further research. First, an empirical study measuring actual detection precision and recall across the four matching algorithm classes examined in Section 9, using audited outcome data from one or more production employee data conversion projects, would strengthen the illustrative comparative findings with statistically validated figures. Second, comparative research examining machine-learning-based entity resolution techniques - including supervised classification models trained on confirmed match/non-match labels and embedding-based similarity approaches - against the classical Fellegi-Sunter and composite-scoring methods examined in this article would clarify whether emerging techniques offer a meaningfully better precision/recall trade-off for this specific problem domain, while accounting for the interpretability and auditability considerations that HCM data governance typically requires. Third, extension research applying the governance framework proposed in Section 13 to other entity types beyond employee records - customer, vendor, patient - would help establish how broadly the multi-phase consistency practices generalize beyond the HCM-specific context examined here. Finally, longitudinal research tracking organizations across multiple conversion programs over time would help establish whether reconciliation tooling and process maturity, once established, transfers effectively to subsequent, unrelated conversion efforts within the same organization.

XVIII. CONCLUSION

This article has presented a probabilistic record-linkage framework for cross-system employee identity reconciliation in the context of large-scale data conversion projects, structured as a five-stage pipeline grounded in the Fellegi-Sunter model and its composite-scoring extensions. The comparative analysis in Section 9 indicates that matching sophistication matters substantially: composite scoring approaches combining multiple weighted similarity features detect a materially larger share of true duplicate identities than single-criterion matching strategies, at the cost of greater design and tuning investment. At the same time, the false-match risk analysis in Section 11 and the cluster formation safeguards in Section 10 underscore that detection completeness alone is an insufficient design goal; because the downstream cost of an incorrect automated merge generally exceeds the cost of a missed match requiring later cleanup, threshold calibration and cluster formation logic should be designed with that asymmetry explicitly in mind rather than optimized for an undifferentiated accuracy metric. The governance framework proposed in Section 13 further extends this article's contribution beyond a single-pass technical pipeline to the multi-phase program structure characteristic of many real-world conversion efforts, emphasizing that a persistent cross-reference registry is necessary to sustain reconciliation consistency as a program proceeds through successive waves. Organizations planning large-scale employee data consolidation are best served by treating identity reconciliation as a distinct, structured project phase employing the full pipeline and governance practices presented in this article, rather than as an informal cleanup activity performed reactively after duplicate records have already reached the target production system.

REFERENCES

1. Accenture. (2020). Data conversion excellence in large-scale HCM transformation programs. Accenture Strategy Publications.
2. Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer.
3. DAMA International. (2017). DAMA-DMBOK: Data management body of knowledge (2nd ed.). Technics Publications.
4. Deloitte. (2019). Employee data migration playbook: Reconciliation and cutover practices for HCM conversions. Deloitte Consulting LLP Insights Series.
5. Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.
6. Loshin, D. (2011). The practitioner's guide to data quality improvement. Morgan Kaufmann.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

Appendix A: Glossary of Terms

Table.A-1. Glossary of key terms used throughout this article.

Term	Definition
Record Linkage	The process of identifying records across one or more datasets that refer to the same real-world entity.
Entity Resolution	A synonym for record linkage, emphasizing the resolution of multiple records into a single entity representation.
Fellegi-Sunter Model	A foundational probabilistic record-linkage model classifying candidate record pairs via a likelihood ratio.
Blocking	A technique partitioning records into smaller candidate groups to reduce the pairwise comparison space.
Similarity Score	A quantitative measure of how closely two records resemble one another across one or more comparison fields.
Adjudication	The process of converting similarity scores into final match, non-match, or manual-review decisions.
Cluster Formation	The process of resolving pairwise match decisions into consolidated groups of records representing the same individual.
Transitive Over-Merging	An error in which a chain of individually plausible pairwise matches produces an incorrect merged cluster.
False Positive (Record Linkage)	An incorrect determination that two records representing distinct individuals are the same individual.
False Negative (Record Linkage)	A failure to determine that two records representing the same individual are in fact the same individual.
Master Data Management (MDM)	A discipline concerned with establishing and maintaining a single, trusted view of core business entities across systems.
Cross-Reference Key Registry	A durable mapping maintained across conversion waves linking resolved identities to their constituent source-system records.

Appendix B: Supplementary Data Tables

This appendix provides additional illustrative data tables supporting the analysis in the main body of the article, including an extended field-weighting reference and a consolidated pipeline stage timing estimate.

B.1 Extended Field-Weighting Reference (Fellegi-Sunter Style)

Table.B-1. Illustrative Fellegi-Sunter-style field-weighting reference for composite similarity scoring

Comparison Field	Illustrative Agreement Weight	Illustrative Disagreement Weight	Discriminating Power
Full Date of Birth (Exact Match)	+9.2	-3.1	Very High
National Identifier (Partial, Exact Match)	+8.7	-2.8	Very High
Surname (Exact Match)	+4.1	-1.2	Moderate
Surname (Phonetic Match Only)	+2.0	-0.6	Low-Moderate
Given Name (Exact Match)	+3.4	-1.0	Moderate
Birth Year Only (Exact Match)	+1.5	-0.4	Low
Address Token Overlap (Partial)	+2.2	-0.7	Low-Moderate

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 10, October 2023||

|DOI:10.15662/IJAREEIE.2022.1210002|

B.2 Illustrative Pipeline Stage Timing Estimate (128,400-Record Population)

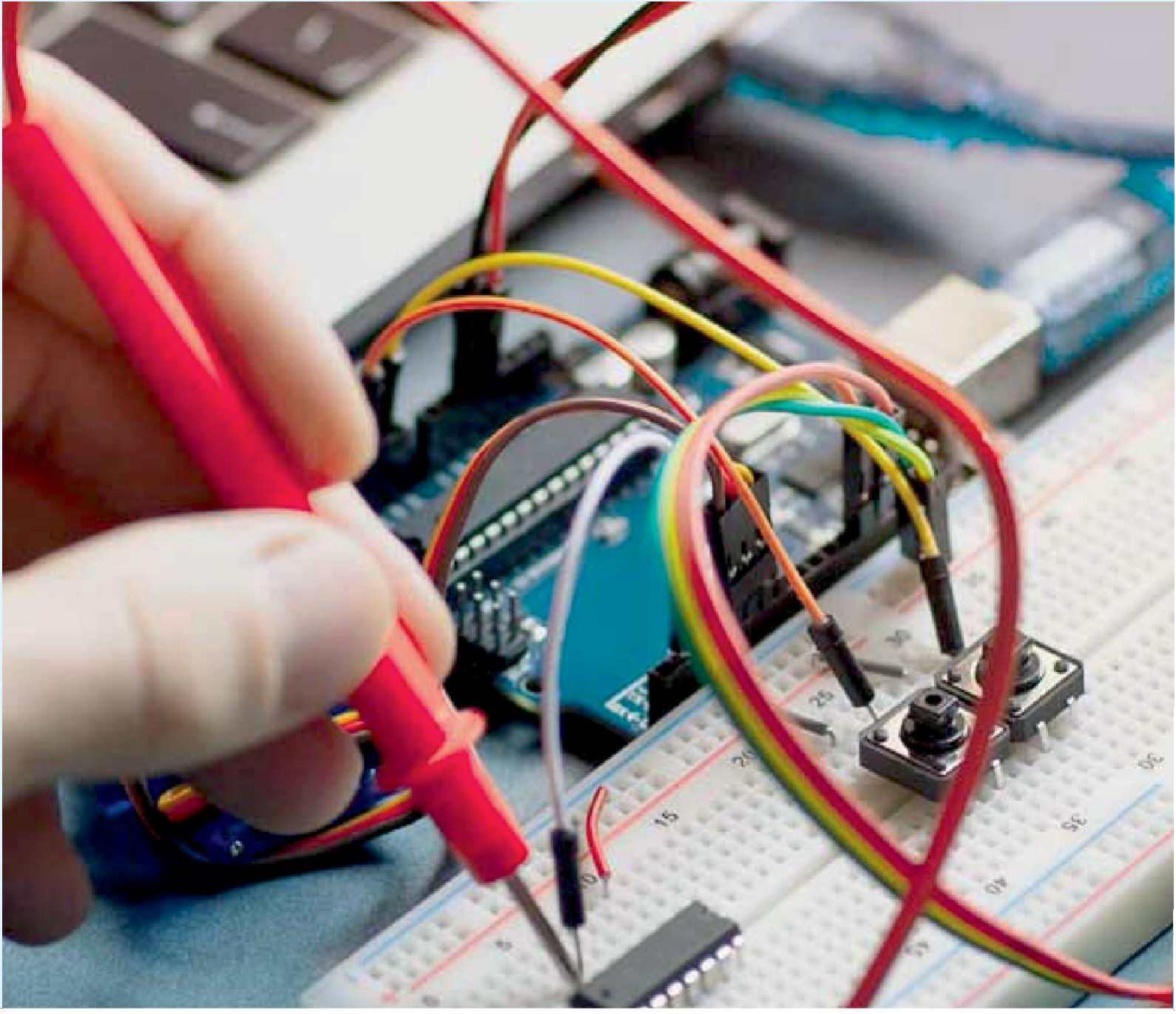
Table.B-2. Illustrative pipeline stage timing estimate for a representative 128,400-record conversion population

Pipeline Stage	Illustrative Processing Duration	Primary Resource Constraint
Source Extraction (3 Systems)	1 - 2 days	Source system extract scheduling and access coordination
Standardization	2 - 4 days	Data engineering effort to build and validate normalization logic
Candidate Blocking	Under 1 day (automated batch run)	Compute resource for blocking key construction
Similarity Scoring	1 - 2 days (automated batch run)	Compute resource for pairwise comparison at scale
Match Adjudication (Including Clerical Review)	5 - 10 business days	Clerical review staffing capacity for intermediate-confidence pairs

B.3 Illustrative Cross-Wave Consistency Metrics

Table.B-3. Illustrative cross-wave consistency metrics for a three-wave conversion program

Wave	New Records Introduced	Matches to Prior-Wave Registry	New Resolved Identities Created
Wave 1 (Initial)	48,200	N/A (initial wave)	44,850
Wave 2	39,600	6,150	36,020
Wave 3	40,600	5,480	37,340



INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.317

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details