

GITG-RMF™

Grumpy Old IT Guy Record Matching Framework™

Version 1.0

**Copyright © 2026 Grumpy Old IT Guy Publishers /
Angelo R. Bobak**

Executive Summary

Modern organizations depend on accurate, trusted, and unified information to support operational systems, analytics, artificial intelligence, regulatory compliance, and customer engagement. Yet one of the most persistent challenges facing enterprises is duplicate data. Customer records, supplier profiles, products, assets, patients, students, citizens, and employees are frequently represented multiple times across applications using slightly different values, formats, abbreviations, or spelling variations.

Duplicate records reduce data quality, increase operational costs, complicate reporting, weaken customer experiences, and undermine confidence in enterprise information.

The **Grumpy Old IT Guy Record Matching Framework (GITG-RMF™)** provides a structured, repeatable methodology for identifying, evaluating, matching,

merging, and governing duplicate records throughout the enterprise. Rather than treating record matching as a one-time cleansing activity, GITG-RMF™ integrates matching into a continuous data quality lifecycle supported by governance, stewardship, monitoring, and measurable quality metrics.

The framework is technology-independent and can be implemented using Microsoft SQL Server, Data Quality Services (DQS), SQL Server Integration Services (SSIS), Microsoft Purview, Master Data Services (MDS), commercial master data platforms, cloud services, or custom matching engines.

Why Record Matching Matters

Organizations typically accumulate duplicate records through:

- Multiple operational systems
- Manual data entry
- Mergers and acquisitions
- Legacy migrations
- Different naming conventions
- Data entry errors
- Missing identifiers
- Inconsistent formatting
- International character variations
- Incomplete customer information

Examples include:

- Robert Smith
- Bob Smith
- Rob Smith

- R. Smith
- Robert J. Smith

Although these records may represent the same individual, traditional exact matching techniques often fail to recognize them.

Without effective record matching:

- Customers receive duplicate mailings.
- Financial reporting becomes inaccurate.
- Healthcare records become fragmented.
- Marketing costs increase.
- Fraud becomes more difficult to detect.
- AI models learn from inconsistent information.
- Master data initiatives struggle to succeed.

What is GITG-RMF™?

The **Grumpy Old IT Guy Record Matching Framework (GITG-RMF™)** is a structured enterprise methodology that standardizes how organizations discover duplicate records, evaluate similarity, determine business survivorship, merge data, and continuously monitor data quality.

Unlike simplistic duplicate detection, GITG-RMF™ recognizes that record matching combines technical algorithms with business rules and governance processes.

The framework supports:

- Exact matching
- Fuzzy matching
- Probabilistic matching

- Deterministic matching
- Hybrid matching strategies
- AI-assisted similarity analysis
- Manual stewardship
- Continuous monitoring

Objectives

The framework seeks to:

- Reduce duplicate records
- Improve enterprise data quality
- Increase confidence in reporting
- Improve customer experience
- Enable trusted analytics
- Support master data management
- Reduce operational costs
- Improve regulatory compliance
- Support AI and machine learning initiatives

The GITG-RMF™ Lifecycle

The framework consists of seven integrated phases.

Phase 1 — Discovery

The organization first identifies where duplicate records may exist.

Activities include:

- Source system inventory
- Data profiling
- Candidate key analysis

- Duplicate frequency analysis
- Attribute completeness review
- Business process assessment

Deliverables:

- Source inventory
- Duplicate assessment report
- Data profiling results
- Risk assessment

Phase 2 — Standardization

Matching quality depends upon standardized data.

This phase normalizes:

- Names
- Addresses
- Phone numbers
- Email addresses
- Company names
- Dates
- Country values
- State abbreviations
- Product descriptions

Common techniques include:

- Parsing
- Data cleansing
- Reference data validation
- Address standardization
- Name normalization
- Abbreviation expansion

Deliverables:

- Standardization rules
- Cleansed datasets
- Reference lookup tables

Phase 3 — Candidate Identification

Rather than comparing every record to every other record, the framework first identifies likely candidates.

Techniques include:

- Blocking keys
- Soundex grouping
- Phonetic encoding
- Geographic clustering
- Hash indexing
- Business key grouping

Candidate generation dramatically improves matching performance while maintaining high accuracy.

Deliverables:

- Candidate record groups
- Blocking strategy
- Candidate statistics

Phase 4 — Similarity Evaluation

Each candidate pair is evaluated using one or more matching algorithms.

Common techniques include:

Exact Matching

Used for:

- Social Security Number
- Customer ID
- Tax ID
- Employee Number

Fuzzy Matching

Algorithms include:

- Levenshtein Distance
- Damerau-Levenshtein
- Jaro
- Jaro-Winkler
- N-Gram Similarity
- Cosine Similarity
- Jaccard Similarity

Phonetic Matching

Examples include:

- Soundex
- Double Metaphone
- NYSIIS

AI-Assisted Matching

Modern AI models evaluate semantic similarity across multiple attributes simultaneously.

Deliverables:

- Similarity scores
- Match confidence
- Candidate rankings

Phase 5 — Match Decision

Similarity scores alone do not determine duplicates.

Business rules establish confidence thresholds.

Typical categories include:

High Confidence

Automatically matched.

Medium Confidence

Requires Data Steward review.

Low Confidence

Rejected.

Organizations often configure thresholds such as:

- 95–100% Automatic Match
- 80–94% Manual Review
- Below 80% Reject

Deliverables:

- Match decisions
- Steward review queue
- Approved duplicate list

Phase 6 — Merge and Survivorship

After identifying duplicates, organizations determine which values survive.

Survivorship rules may prioritize:

- Most recent value
- Highest quality source
- Most complete record
- Trusted application
- Steward-approved value
- Longest customer history

Merged records produce a single trusted Golden Record.

Deliverables:

- Golden Record
- Survivorship rules
- Merge history
- Audit trail

Phase 7 — Governance and Monitoring

Record matching never ends.

Continuous monitoring measures:

- Duplicate rates
- Matching accuracy
- False positives
- False negatives
- Steward workload
- Merge success
- Data quality trends

Governance ensures:

- Business ownership
- Policy enforcement
- Steward accountability
- Continuous improvement

Deliverables:

- Quality dashboards
- KPI reports
- Governance metrics
- Continuous improvement plans

Matching Techniques Supported

The framework supports multiple matching methodologies.

Technique	Best Used For
Exact Match	Unique identifiers
Deterministic	Business rules
Probabilistic	Large customer datasets
Fuzzy Matching	Names and addresses
Phonetic Matching	Spoken names
Rule-Based Matching	Industry-specific logic
AI-Assisted Matching	Complex enterprise matching
Hybrid Matching	Highest accuracy

Deliverables

A complete implementation typically produces:

- Record Matching Strategy
- Business Matching Rules
- Standardization Rules
- Candidate Identification Rules
- Matching Algorithms
- Similarity Score Matrix
- Steward Review Process
- Survivorship Rules
- Golden Record Definition
- Governance Policies
- Quality Dashboards
- Audit Reports
- Continuous Monitoring Metrics

Required Roles

Successful implementations involve multiple business and technical roles.

- Executive Sponsor - Provides strategic direction and funding.
- Data Governance Council - Approves enterprise matching policies.
- Business Data Owners - Define business rules.
- Data Stewards - Review uncertain matches.
- Data Quality Analysts - Analyze duplicate patterns.
- Data Engineers - Implement matching pipelines.
- Database Administrators - Optimize performance.
- Application Owners - Integrate operational systems.
- Compliance Officers - Ensure regulatory adherence.

Technology Components

The framework is platform-independent and supports technologies including:

Microsoft SQL Server

- Data Quality Services (DQS)
- Integration Services (SSIS)
- SQL Server Fuzzy Matching functions
- SQL Server Machine Learning Services

If using SQL Server 2025 or above:

Microsoft Purview

- Metadata management
- Data governance
- Lineage
- Policy enforcement

Master Data Platforms

- Master Data Services (legacy)
- Commercial MDM platforms
- Cloud-native master data services

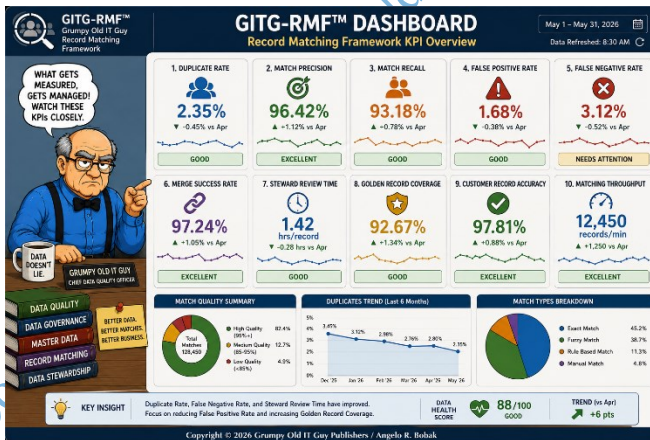
AI Technologies

- Machine learning similarity models
- Large Language Models
- Entity resolution services
- Intelligent data classification

Key Performance Indicators

Organizations implementing GITG-RMF™ should monitor:

- Duplicate Rate
- Match Precision
- Match Recall
- False Positive Rate
- False Negative Rate
- Merge Success Rate
- Steward Review Time
- Golden Record Coverage
- Customer Record Accuracy
- Matching Throughput



Sample RMF Dashboard

Benefits

Organizations implementing GITG-RMF™ typically experience:

- Higher customer data quality
- Reduced duplicate records
- Improved operational efficiency
- Better analytics
- More trusted AI models
- Lower mailing costs
- Faster customer service
- Improved regulatory compliance
- Stronger master data governance
- Increased executive confidence in enterprise data

Relationship to Other GITG™ Frameworks

GITG-RMF™ is designed to operate as part of the broader Grumpy Old IT Guy data management methodology.

Typical implementation sequence:

1. **GITG-DPF™** — Data Profiling Framework identifies data quality issues and duplicate patterns.
2. **GITG-DCF™** — Data Cleansing Framework standardizes and corrects data to improve match quality.
3. **GITG-RMF™** — Record Matching Framework identifies duplicate entities and establishes trusted Golden Records.
4. **GITG-CDSF™** — Customer Data Standardization Framework applies standardized customer

structures and business rules across enterprise systems.

5. **GITG-DQGF™** — Data Quality Governance Framework ensures that matching policies, stewardship responsibilities, and quality metrics are continuously monitored and improved.

Together, these frameworks form an integrated enterprise data quality lifecycle that transforms fragmented, inconsistent data into trusted, governed information assets.

Microsoft SQL Server Fuzzy Matching Functions

MICROSOFT SQL SERVER FUZZY MATCHING FUNCTIONS					
Built-in functions for measuring string similarity and supporting fuzzy matching in SQL Server.					
PURPOSE		HOW TO USE			
These functions help detect similar values that are not identical. They are essential for data quality, deduplication, and master data management.		Use these functions in T-SQL queries, views, stored procedures, and matching workflows to score and compare text values.			
FUNCTION	DESCRIPTION	RETURNS	RANGE	BEST USED FOR	EXAMPLE
 EDIT_DISTANCE	Calculates the Levenshtein distance between two expressions. It counts the minimum number of single-character edits (insertions, deletions, substitutions) required to change one string into the other.	int	0 to n (0 = exact match)	Measuring how many changes sequence two strings.	<code>SELECT BEST_DISTANCE('toston','titting') AS Distance;</code>
 JARO_WINKLER_SIMILARITY	Returns a similarity score based on the Jaro-Winkler algorithm, which gives higher scores to strings that match from the start (prefix weighting).	float	0.0 to 1.0 (1.0 = exact match)	Comparing names and short strings with common prefixes (e.g., "Jonathan" vs "Jonahen").	<code>SELECT JARO_SIMILARITY(SIMILARITY('Jonathan','Jonathan') AS Similarity);</code>
 SOUNDEX	Converts a string to its Soundex code, which represents how the string sounds in English. Strings with the same code sound similar.	varchar	4 characters	Finding phonetic matches for names (e.g., "Toson" vs "Tosyn").	<code>SELECT SOUNDEX('Robert'), SOUNDEX('Robert') AS Code;</code>
 DIFFERENCE	Returns an integer rating that indicates the degree of difference between two strings based on a phonetic algorithm.	int	0 to 4 (0 = similar)	Quick phonetic comparisons of names and words.	<code>SELECT DIFFERENCE('toston','tosing') AS DiffScore;</code>
 NGRAM_DISTANCE	Calculates the distance between two strings based on the number of matching bigrams. Lower scores mean more similar.	float	0.0 to 1.0 (0.0 = exact match)	Comparing longer strings such as addresses or descriptions.	<code>SELECT NGRAM_DISTANCE('123 Main Street','123 Main St.') AS NgramDist;</code>
 METAPHONE	Returns a metaphone code for a string, exposing phonetic matching for non-English names and more complex words.	varchar	variable	Phonetic matching across diverse languages and surnames.	<code>SELECT METAPHONE('Schneider') AS MetaCode;</code>
 PHONETIC_SIMILARITY	Compares two strings using phonetic algorithms (Smetaphone) and returns a similarity score.	float	0.0 to 1.0 (1.0 = exact match)	Scoring phonetic similarity between names.	<code>SELECT PHONETIC_SIMILARITY('Catherine','Katherine') AS PhoneticSim;</code>
 ASCII	Returns the ASCII value of the left-most character in the expression.	int	0 to 255	Supporting custom normalization and character comparisons.	<code>SELECT ASCII('A') AS AsciiValue;</code>
 TIPS - Combine multiple functions for higher accuracy. - Use thresholds to decide match vs. non-match. - Example: <code>JARO_WINKLER_SIMILARITY() > 0.90 AND SOUNDEX() match.</code> - Consider context (e.g., country, data type) along with fuzzy scores.  Better Data. Better Decisions. Better Business.					

Conclusion

Duplicate records are not merely a technical inconvenience—they represent a strategic business risk that affects customer relationships, operational efficiency, regulatory compliance, analytics, and artificial intelligence. Effective record matching requires more than sophisticated algorithms; it demands standardized data, well-defined business rules, governance, stewardship, and continuous measurement.

The **Grumpy Old IT Guy Record Matching Framework (GITG-RMF™)** provides organizations with a comprehensive methodology for discovering duplicate records, evaluating similarity, making consistent match decisions, creating trusted Golden Records, and sustaining high-quality data over time. By integrating modern matching techniques with enterprise governance practices, GITG-RMF™ enables organizations to establish a reliable foundation for master data management, advanced analytics, and AI-driven decision-making while ensuring that record matching becomes a repeatable and continuously improving business capability rather than a one-time project.

© 2026 Angelo R. Bobak
All rights reserved.

First Edition: February 2026

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the prior written permission of the publisher.

Grumpy Old IT Guy Publishers

Westchester County, NY, USA

www.grumpyolditguy.com

Copyright 2026 - Grumpy old IT Guy Publishers