## **Input Files**

1. **Missing Owners Report** (missing\_owners.xlsx)
   * Contains owners from the old system not found in current production
   * ~2,484 owners needing ID migration
   * Fields: Owner ID, Name, Address, City, State, ZIP
2. **Excluded Owners ID Comparison** (excluded\_owners.xlsx)
   * Contains the mapping reference data with new IDs
   * ~4,767 owner records
   * Fields: OLD\_OWNER\_ID, NEW\_OWNER\_ID, Name, Address, City, State, ZIP

## **Data Preprocessing**

### **Name Cleaning**

* Converts to lowercase
* Removes punctuation and special characters
* Strips extra whitespace
* Extracts attention lines (e.g., "Attn: John Smith")
* Standardizes common abbreviations (LLC, LTD, CO, etc.)

### **Address Normalization**

* Uses usaddress library to parse address components
* Standardizes street types (ST, STREET, AVE, AVENUE)
* Normalizes directionals (N, NORTH, E, EAST)
* Handles suite/unit variations (STE, SUITE, #)

## **Multi-Step Matching Strategy**

### **Step 1: Exact Name Match**

* **Confidence:** 100%
* **Method:** Direct string comparison on cleaned names
* **Validation:** Ensures state matches
* **Use Case:** Owners with identical names and no variations

### **Step 2: Fuzzy Name Match with Address Validation**

* **Confidence:** 75-95%
* **Method:**
  + Uses RapidFuzz token\_sort\_ratio for name matching
  + Requires name similarity ≥ 75%
  + Validates with address similarity ≥ 70%
* **Protection:** Rejects matches with weak address scores to prevent false positives
* **Use Case:** Names with minor variations (e.g., "Smith Energy LLC" vs "Smith Energy")

### **Step 3: Address-First Match (Attention Line Handler)**

* **Confidence:** 85-95%
* **Method:**
  + Prioritizes address matching (≥ 80% similarity)
  + Checks if attention line matches owner name in excluded database
  + Validates name similarity ≥ 55%
* **Special Case:** Handles "Attn: Company X" scenarios where attention line is the real owner
* **Example:** "SBI West Texas; Attn: Extex Operating" → Matches to "Extex Operating"

### **Step 4: Global Search (Cross-State Fallback)**

* **Confidence:** Capped at 80%
* **Method:**
  + Removes state restriction for difficult matches
  + Uses high thresholds (name ≥ 85%, address ≥ 75%)
  + Flags as CROSS\_STATE\_REVIEW
* **Use Case:** Companies that relocated or have addresses in different states

### **Step 5: Partial Name Review**

* **Confidence:** 70-85%
* **Method:**
  + Uses partial\_ratio for substring matching
  + Requires ≥ 70% partial match
  + Validates with address
* **Use Case:** Name expansions (e.g., "Steko" vs "Steko Investments of Texas")

### **Step 6: Confirmed Fuzzy Match**

* **Confidence:** 80-90%
* **Method:**
  + Multiple fuzzy algorithms (token\_sort, token\_set, partial)
  + All must score ≥ 80%
  + Address validation required
* **Use Case:** High-confidence fuzzy matches with address confirmation

## **Confidence Score Calculation**

**Formula:**

****Confidence = (name\_score × 0.70) + (address\_score × 0.30)

**Adjustments:**

* Cross-state matches: Capped at 80%
* Exact matches: Always 100%
* Address-first matches: Weighted toward address similarity

## **Output Structure**

### **mapped\_owners.csv**

Contains all 2,484 missing owners with match results:

**Columns:**

* **Previous Owner ID** - Original ID from missing owners
* **Owner Name** - Original name
* **Cleaned Name** - Normalized version
* **Attention Name** - Extracted attention line (if any)
* **Last Known Address** - Full address string
* **State** - State code
* **mapped\_new\_id** - New owner ID (or NULL if unmatched)
* **match\_step** - Which strategy found the match
* **matched\_on\_attention** - Yes/No - Was attention line used?
* **confidence\_score** - Overall confidence (0-100)
* **name\_score** - Name similarity percentage
* **address\_score** - Address similarity percentage
* **status** - Human-readable status
* **review\_priority** - HIGH/MEDIUM/LOW for unmatched records
* **suggested\_action** - Recommended next step

**Sorting:**

* Rows 1-N: Matched owners (sorted by confidence 100 → 75)
* Rows N+1-end: Unmatched owners (sorted by review priority)

## **Match Results Summary**

### **Current Performance:**

* **Total Owners:** 2,484
* **Matched:** 147 (5.9%)
* **Unmatched:** 2,337 (94.1%)

### **Match Type Breakdown:**

* **ADDRESS\_NAME\_MATCH:** 54 matches - Address + name fuzzy match
* **EXACT\_NAME:** 41 matches - Perfect name match
* **ADDRESS\_ATTN\_MATCH:** 20 matches - Attention line matched to owner
* **PARTIAL\_NAME\_REVIEW:** 12 matches - Substring/partial matching
* **FUZZY\_NAME:** 11 matches - Standard fuzzy name match
* **CONFIRMED\_FUZZY:** 5 matches - Multi-algorithm confirmation
* **CROSS\_STATE\_REVIEW:** 4 matches - Cross-state fallback

### **Why Low Match Rate?**

* Many owners are legitimately inactive (sold rights, deceased, dissolved)
* Client expects ~64% to be missing from current production
* Excluded owners database may have data quality issues
* 147 matches recovered owners that would have been lost

## **Error Prevention & Data Quality**

### **False Positive Protection:**

1. **Address validation required** - Name-only matches rejected if address differs significantly
2. **State verification** - Cross-state matches flagged for review
3. **Attention line safeguards** - Prevents matching to wrong company when attention line present
4. **Confidence thresholds** - Low confidence matches flagged for manual review

### **Duplicate Handling:**

* Same owner with multiple addresses: All rows kept during matching
* System picks best address match for each old owner ID
* No premature deduplication that could lose matching data

## **Dependencies**

* **pandas** - Data manipulation and CSV/Excel processing
* **rapidfuzz** - High-performance fuzzy string matching
* **usaddress** - Address parsing and normalization
* **openpyxl** - Excel file reading

## **Recommendations for Review**

### **Priority 1: High Confidence (≥90%)**

* 41 exact matches + high fuzzy matches
* Safe to apply automatically

### **Priority 2: Medium Confidence (75-89%)**

* Spot check 10-20% of these
* Look for obvious false positives

### **Priority 3: Low Confidence (<75%)**

* Review carefully before applying
* May require manual verification

### **Unmatched Records:**

* HIGH priority unmatched (65-74% similarity) - Worth manual investigation
* MEDIUM/LOW priority - Likely legitimately inactive