

Push Instructions: EPC Address Matching Debug & Fix

Branch Information

- **Branch Name:** fix/epc-matching-debug
- **Commit Hash:** 8822e69
- **Status:** Successfully pushed to GitHub
- **PR URL:** <https://github.com/CliveCaseley/soldcomp-analyser2/pull/new/fix/epc-matching-debug>

Summary

Enhanced EPC certificate matching with comprehensive logging, text normalization, and defensive checks to debug and fix incorrect certificate selection.

What Was Fixed

Problem

The Apify actor was reportedly selecting wrong EPC certificates:

- Input: "32, Summerfields Drive, Blaxton, Doncaster DN9 3BG"
- Expected: Certificate 0587-3024-5207-2897-6200 (for "32 Summerfields Drive")
- Got: Certificate 2510-0044-8002-0798-5706 (for "2 Summerfields Drive")

Investigation Results

After thorough testing, the matching algorithm was found to be **working correctly** when tested in isolation. The issue likely occurs due to:

- Missing or malformed address data in production
- Fallback mechanism triggering when address is empty/null
- Punctuation inconsistencies in production data

Solution Implemented

1. **Text Normalization** (`normalizeTextForMatching()`)
 - Removes all punctuation (commas, periods, semicolons)
 - Collapses multiple spaces
 - Converts to lowercase
 - Ensures consistent comparison

2. **Comprehensive Logging**
 - Input parameters (postcode, address)
 - Certificate count found
 - House number extraction details
 - Normalized text for comparison
 - Per-candidate scoring breakdown
 - Top 5 matches summary
 - Final selection with reasoning

3. Defensive Checks

- Validates certificates array isn't empty
- Checks for missing/empty target address
- Prevents crashes from null/undefined data
- Explicit fallback behavior with warnings

4. Score Threshold Adjustment

- Lowered from **0.4** to **0.3** for better matching
- Still prioritizes house number matches (70% weight)

Test Results

DN9 3BG Postcode Test (39 certificates)

PASSED - Correct certificate selected!

Results:

- Selected Certificate: 0587-3024-5207-2897-6200
- Selected Address: "32 Summerfields Drive, Blaxton, DONCASTER, DN9 3BG"
- Final Score: 1.000 (perfect match)

Scoring Breakdown:

House #	House Score	Street Score	Total Score	Selected
32	1.000	1.000	1.000	YES
2	0.000	1.000	0.300	NO
4	0.000	1.000	0.300	NO
Others	0.000	1.000	0.300	NO

Files Modified

1. `src/utils/epcHandler.js`

- Added `normalizeTextForMatching()` function
- Enhanced `findBestAddressMatchFromScrapedData()` with:
 - Defensive checks for empty data
 - Comprehensive emoji-based logging
 - Text normalization
 - Top 5 matches summary
 - Lower score threshold (0.3)
- Enhanced `getCertificateNumber()` with:
 - Detailed input validation
 - Step-by-step logging
 - Clear success/fallback messages
 - Updated module exports to include new helper function

2. `test-epc-matching-debug.js (new)`

- Comprehensive test script
- Tests all 39 certificates for DN9 3BG
- Validates house number extraction

- Shows detailed scoring breakdown
- Verifies correct certificate selection

3. EPC_MATCHING_DEBUG_FIX.md (new)

- Complete documentation of problem, solution, and test results
- Scoring algorithm explanation
- Recommendations for production debugging
- Next steps

How to Verify Locally

```
cd /home/ubuntu/github_repos/soldcomp-analyser2
git checkout fix/epc-matching-debug

# Run the test
node test-epc-matching-debug.js

# Expected output: ✓ SUCCESS: Correct certificate selected!
```

Creating Pull Request

1. Visit: <https://github.com/CliveCaseley/soldcomp-analyser2/pull/new/fix/epc-matching-debug>
2. Use this PR title:

```
fix: Enhanced EPC address matching with comprehensive logging
```

3. Use this PR description:

```
```markdown
```

```
Problem
```

EPC certificate matching was reportedly selecting wrong certificates in production.

Example:

- Input: "32, Summerfields Drive, Blaxton, Doncaster DN9 3BG"
- Expected: Certificate for "32 Summerfields Drive" (0587-3024-5207-2897-6200)
- Got: Certificate for "2 Summerfields Drive" (2510-0044-8002-0798-5706)

## Investigation

After thorough testing, the matching algorithm works correctly in isolation.

Issue likely due to:

- Missing/malformed address data in production
- Punctuation inconsistencies
- Fallback mechanism triggering

## Solution

1. **Text Normalization:** Added punctuation removal and whitespace handling
2. **Comprehensive Logging:** Detailed emoji-based logging for production debugging
3. **Defensive Checks:** Validation for empty/null data
4. **Score Threshold:** Lowered from 0.4 to 0.3 for better matching

## Test Results

- ✓ DN9 3BG test (39 certificates) - PASSED
- ✓ Correct certificate selected: 0587-3024-5207-2897-6200
- ✓ Perfect score: 1.000

```
Files Changed
- src/utils/epcHandler.js : Enhanced matching and logging
- test-epc-matching-debug.js : Comprehensive test script (new)
- EPC_MATCHING_DEBUG_FIX.md : Full documentation (new)
```

#### ## Next Steps

- Test in production with Apify actor
- Monitor logs for “⚠ Falling back” warnings
- Review upstream address data if issues persist

See `EPC_MATCHING_DEBUG_FIX.md` for complete documentation.

```

Production Debugging

If wrong certificates are still selected after this fix:

1. Check the logs for these indicators:

- ⚠ No target address provided - Address is empty/null
- ⚠ Falling back to first certificate - Score threshold not met
- ✗ No good match found - None of the candidates matched well

2. Verify input data:

- Log the exact address string being passed
- Check postcode extraction is correct
- Ensure address isn't empty

3. Review top 5 matches:

- Look for the expected certificate in top 5
- Check if house number extraction is working
- Verify street name matching

Scoring Algorithm

Weights

- **House Number Match:** 70%
- **Street Name Match:** 30%

Threshold

- **Score Threshold:** 0.3 (was 0.4)
- If best score > 0.3: Return best match
- If best score ≤ 0.3: Fall back to first certificate (with warning)

Contact

For questions about this fix, review the documentation:

- `EPC_MATCHING_DEBUG_FIX.md` - Complete technical documentation
- `test-epc-matching-debug.js` - Test script with examples
- GitHub PR comments

Status:  Ready for review and merge

Branch: fix/epc-matching-debug

Commit: 8822e69