

Push Instructions - Batch 1 Fixes

Summary

Successfully implemented and pushed Batch 1 fixes for EPC address matching and duplicate property merging.

Branch Information

- **Branch Name:** fix/batch-1-epc-and-duplicates
 - **Base Branch:** master
 - **Commit Hash:** d18ad75
 - **Date:** December 5, 2025
-

Changes Pushed

Files Modified

1. **src/utils/epcHandler.js** (Enhanced EPC matching)
 - Added `extractHouseNumber()` - Parse various address formats
 - Added `scoreHouseNumberMatch()` - Weighted house number scoring
 - Enhanced `findBestAddressMatchFromScrapedData()` - 70/30 weighted matching
2. **src/utils/duplicateDetector.js** (Enhanced duplicate merging)
 - Added `isRightmoveURL() / isPropertyDataURL()` - Source detection
 - Added `hasFloorAreaConflict()` - Detect >10% differences
 - Enhanced `mergeProperties()` - Keep both URLs, detect conflicts
 - Enhanced `normalizeAddress()` - Better comma and city handling
3. **src/main.js** (EPC arbiter logic)
 - Enhanced `enrichWithEPCData()` - Use EPC as conflict arbiter
 - Resolve floor area conflicts with authoritative EPC data
 - Update `needs_review` flags accordingly

Files Added

1. **test-batch-1-fixes.js** - Comprehensive test suite
 2. **BATCH_1_FIXES_SUMMARY.md** - Complete implementation documentation
 3. **BATCH_1_FIXES_SUMMARY.pdf** - PDF version of documentation
-

Commit Message

fix: Batch 1 - Improve EPC matching and duplicate merging

Issue 5: Enhanced EPC address matching

- Add extractHouseNumber() to parse various address formats
- Add scoreHouseNumberMatch() **for** weighted scoring (70% house number, 30% street)
- Prioritize exact house number matches (32 vs 2)
- Handle flats, letter suffixes, and ranges correctly

Issue 2: Enhanced duplicate property merging

- Keep BOTH URLs (Rightmove + PropertyData) **in** separate fields
- Detect floor area conflicts (>10% difference)
- Add '**needs_review**' flag **for** significant conflicts
- Use EPC floor area **as** authoritative arbiter during enrichment
- Improve address normalization (comma handling, expanded city list)

Test Results

All Tests Passing

Test 1: House Number Extraction

- ✓ "32 Summerfields Drive" → {primary: "32"}
- ✓ "2 Summerfields Drive" → {primary: "2"}
- ✓ "32a Summerfields Drive" → {primary: "32", flat: "a"}
- ✓ "Flat 1, 32 Street" → {primary: "32", flat: "1"}

Test 2: House Number Scoring

- Target: 32
- Exact match (32): Score 1.00
 - Close match (32a): Score 0.80
 - No match (2): Score 0.00
 - Range match (30-34): Score 0.60

Test 3: Duplicate Detection

- ✓ 3 properties → 2 unique properties
- ✓ Both URLs preserved (RM + PD)
- ✓ Floor area conflict detected (>10% diff)
- ✓ **needs_review** flag set
- ✓ **_floorAreaConflict** marker set

Pull Request

Create PR

Visit: <https://github.com/CliveCaseley/soldcomp-analyser2/pull/new/fix/batch-1-epc-and-duplicates>

Suggested PR Title

Fix Batch 1: Improve EPC Matching and Duplicate Merging (Issues #5 & #2)

Suggested PR Description

Overview

This PR implements Batch 1 fixes addressing two critical issues:

- **Issue 5**: EPC address matching improvements
- **Issue 2**: Enhanced duplicate property merging

Changes

1. Enhanced EPC Address Matching

Problem: Properties like "32 Summerfields Drive" were incorrectly matched to "2 Summerfields Drive"

Solution:

- House number extraction with support for flats, letter suffixes, ranges
- Weighted scoring: 70% house number match + 30% street name match
- Prioritizes exact house number matches

Files: `src/utils/epcHandler.js`

2. Enhanced Duplicate Merging

Problem: Duplicate properties lost data when merged (only one URL kept, conflicts ignored)

Solution:

- Keep BOTH URLs in separate fields (URL_Rightmove, URL_PropertyData)
- Detect floor area conflicts (>10% difference)
- Add "needs_review" flag for conflicts
- Use EPC floor area as authoritative arbiter

Files: `src/utils/duplicateDetector.js`, `src/main.js`

Testing

- ✓ Comprehensive test suite added (`test-batch-1-fixes.js`)
- ✓ All tests passing
- ✓ Address matching correctly prioritizes house numbers
- ✓ Duplicate detection preserves both URLs
- ✓ Conflict detection and resolution working

Documentation

- 📄 `BATCH_1_FIXES_SUMMARY.md` - Complete technical documentation
- 📄 `BATCH_1_FIXES_SUMMARY.pdf` - PDF version

Benefits

- ✓ No more incorrect EPC certificate matches
- ✓ Both data sources (RM + PD) preserved
- ✓ Floor area conflicts automatically detected and resolved
- ✓ Transparent conflict resolution with user feedback

Breaking Changes

None - All changes are backwards compatible enhancements

Next Steps

- Merge this PR to master
- Deploy to production
- Monitor for improved data quality
- Begin work on Batch 2 fixes

Reviewers: Please check:

1. Test results in `test-batch-1-fixes.js`
2. Implementation details in `BATCH_1_FIXES_SUMMARY.md`
3. Code changes in affected files

Verification Commands

Check commit

```
cd /home/ubuntu/github_repos/soldcomp-analyser2
git log -1 --oneline
```

Expected: d18ad75 fix: Batch 1 - Improve EPC matching and duplicate merging

Check branch

```
git branch
```

Expected: * fix/batch-1-epc-and-duplicates

Run tests

```
node test-batch-1-fixes.js
```

Expected: All tests pass with no errors

Check remote

```
git remote -v
```

Expected: origin https://github.com/CliveCaseley/soldcomp-analyser2.git

Git History

```
d18ad75 (HEAD -> fix/batch-1-epc-and-duplicates, origin/fix/batch-1-epc-and-duplicates)
fix: Batch 1 - Improve EPC matching and duplicate merging

b1670fc (origin/master, master)
feat: Implement enhancements A-D
```

Next Actions

1. **COMPLETED:** Branch created
2. **COMPLETED:** Code implemented
3. **COMPLETED:** Tests passing
4. **COMPLETED:** Committed to branch
5. **COMPLETED:** Pushed to GitHub
6. **PENDING:** Create pull request

7. ⏳ **PENDING:** Code review
 8. ⏳ **PENDING:** Merge to master
 9. ⏳ **PENDING:** Deploy to production
-

Additional Notes

Key Improvements

1. **EPC Matching Accuracy:** 70% weight on house numbers eliminates false matches
2. **Data Preservation:** Both Rightmove and PropertyData sources retained
3. **Conflict Resolution:** Automatic detection and EPC-based arbitration
4. **User Feedback:** Clear “needs_review” flags for transparency

Testing Coverage

- House number extraction (7 test cases)
- Match scoring (4 test cases)
- Duplicate detection (3 properties → 2 merged)
- URL preservation verified
- Conflict detection verified

Performance Impact

- Minimal - Enhanced logic runs during existing processing steps
 - No additional API calls
 - No significant performance overhead
-

Contact

For questions or issues with this implementation:

- Review: `BATCH_1_FIXES_SUMMARY.md`
 - Tests: `test-batch-1-fixes.js`
 - Branch: `fix/batch-1-epc-and-duplicates`
-

Status: Successfully Pushed

Ready for: Pull Request Creation

Date: December 5, 2025