# EPC Web Scraping Implementation Summary

## Overview

Successfully implemented a new approach to fetch EPC certificate numbers by scraping the postcode search page directly, replacing the previous API-based approach that was unreliable.

## Problem

The EPC API often doesn't return the `certificate-number` field in search results, making it impossible to generate direct certificate URLs. Even with a two-step approach (search API + individual certificate API), certificate numbers were frequently missing.

## Solution

Implemented web scraping to extract certificate numbers directly from the government's EPC postcode search page:

**URL Pattern**: `https://find-energy-certificate.service.gov.uk/find-a-certificate/search-by-postcode?postcode={postcode}`

**HTML Structure Scraped**:

```
<a class="govuk-link" href="/energy-certificate/2648-3961-7260-5043-7964">
  92a, The Quadrant, HULL, HU6 8NS
</a>
```

## Implementation Details

### New Functions

1. **`scrapeCertificateNumbersFromPostcode(postcode)`**
   - Scrapes the postcode search page
   - Extracts certificate numbers from href attributes
   - Returns array of certificate objects with:
     - `certificateNumber`: The certificate number (e.g., "2648-3961-7260-5043-7964")
     - `address`: Full property address
     - `href`: Complete certificate URL
     - `rating`: EPC rating (if available)
2. **`findBestAddressMatchFromScrapedData(certificates, targetAddress)`**
   - Matches scraped certificates to target address
   - Uses word-matching algorithm for fuzzy matching
   - Returns best match with score > 50%

### Updated Functions

1. **`getCertificateNumber(postcode, address, apiKey)`**
   - Now uses web scraping instead of API
   - Calls `scrapeCertificateNumbersFromPostcode()`
   - Matches by address using `findBestAddressMatchFromScrapedData()`
   - Returns certificate data with URL

2. **`fetchEPCDataViaAPI(postcode, address, apiKey)`**
   - Renamed but still used for backward compatibility
   - Now delegates to web scraping approach
   - Returns consistent data structure

### File Changes

**Modified**: `src/utils/epcHandler.js`
- Added web scraping implementation
- Updated version to v2.5
- Updated header comments to reflect new approach
- Kept API key parameter for compatibility but no longer required

**Added**: `test-epc-scraping.js`
- Test script to verify scraping functionality
- Tests postcode scraping and address matching

## Test Results

✅ **Test 1: Scraping certificates for postcode HU6 8NS**
- Successfully scraped 8 certificates
- Extracted certificate numbers from href attributes
- All certificate numbers validated successfully

✅ **Test 2: Address matching**
- Target: "92a, The Quadrant, HULL"
- Matched to certificate: `2648-3961-7260-5043-7964` ✓
- Generated URL: `https://find-energy-certificate.service.gov.uk/energy-certificate/2648-3961-7260-5043-7964`

## Benefits

1. **More Reliable**: Web scraping always provides certificate numbers
2. **No API Dependency**: No need for API keys or authentication
3. **Direct URLs**: Always generates working certificate URLs
4. **Better Matching**: Address matching ensures correct certificate selection
5. **Simpler Code**: Removed complex two-step API process

## Git Workflow

### Branch Created

- **Branch**: `fix/epc-scrape-certificate`
- **Base**: `master`

### Commit

- **Commit Hash**: `959eaa7`
- **Message**: “feat: Implement web scraping for EPC certificate numbers”

### Pull Request

- **PR Number**: #6
- **Title**: “Fix: Implement Web Scraping for EPC Certificate Numbers”
- **URL**: https://github.com/CliveCaseley/soldcomp-analyser2/pull/6
- **Status**: Open

### Previous PR

- **PR Number**: #5
- **Status**: Closed and Merged (already completed before this task)

## Testing Instructions

To test the implementation:

```
cd /home/ubuntu/github_repos/soldcomp-analyser2
node test-epc-scraping.js
```

## Next Steps

1. Review PR #6 in GitHub
2. Verify the implementation meets requirements
3. Merge PR #6 when ready
4. Deploy to production

## Files Modified

- `src/utils/epcHandler.js` - Main implementation
- `test-epc-scraping.js` - Test script (new file)

## Version History

- **v2.1**: Initial EPC API Integration
- **v2.3**: Certificate Number URL Format (using certificate-number not lmk-key)
- **v2.4**: Two-Step Certificate Number Retrieval (search API + individual certificate API)
- **v2.5**: Web Scraping Approach (current) ⭐