IDenteq

Data Quality Assessment:

Rule Base

Version	Date	Author	Changes Made
1.0	28/02/2025	IDenteq	Initial version

IDenteq

Table of Contents

Overview

IDenteq is assessing CMOS addresses for accuracy by cross-referencing them with external data sources, including AddressBase Premium (ABP) and the Valuation Office Agency (VOA).

Each CMOS address undergoes a series of validation processes to account for exceptions, misspellings, and abbreviations.

- If a confident match is found in AddressBase Premium or the Valuation Office Agency, the SPID address is assigned the status of **"Verified."** The corresponding UPRN and/or VOA BA Reference is then compared with the CMOS UPRN and/or VOA BA Reference (if available) to determine accuracy.
- If a confident match cannot be established, the address is assigned the status of **"Of Concern."**
- A status of **"Unmatched"** is applied to SPID addresses that cannot be categorised as either Verified or Of Concern. This occurs when a supply address has no corresponding match in external data sources and is deemed **"non-addressable."**

Each SPID address, UPRN, and VOA BA Reference is assigned a status independently based on the verification findings.

The following outlines the potential categorisation combinations identified during the auditing process:

SPID Address Data - Status	SPID Premises Data - UPRN	SPID Premises Data - VOA
Verified	Verified	Verified
	Of Concern	Of Concern
	Verified	Of Concern
	Of Concern	Verified
Of Concern	Of Concern	Of Concern
Unmatched	N/A	N/A

Categorisation

SPID Address Data

IDenteq catagorises SPID Address Data into three categories:

1. **Verified**: The SPID address confidently matches an address in key external reference datasets, such as AddressBase Premium (ABP) and the Valuation Office Agency (VOA).
2. **Of Concern**: The SPID address does not confidently match any address in the key external reference datasets, AddressBase Premium (ABP) or the Valuation Office Agency (VOA).
3. **Unmatched**: The SPID address is not found in the reference data sources and is identified as a Non-Addressable Site (e.g., bin stores, public conveniences, allotments, or features like troughs and stand-pipes).

SPID Address Data - Status	Identified as a Non-Addressable Site	Address Base Match Found?	Valuation Office Agency Match Found?
Verified	☐	☑	☑
	☐	☑	☐
	☐	☐	☑
Of Concern	☐	☐	☐
Unmatched	☑	☐	☐

When a SPID address is matched to a record in AddressBase Premium (ABP) or the Valuation Office Agency (VOA), the corresponding UPRN/VOA BA reference is used to verify the accuracy of the SPID Premises Data.

Those that are identified as “Unmatched” do not progress further to checks on the SPID Premises Data.

SPID Premises Data

IDenteq categorises SPID Premises Data independently for UPRN and VOA BA references into two primary categories:

1. **Verified**: CMOS holds a UPRN/VOA BA reference which has been verified as correct against ABP/VOA datasets. When matching the CMOS address to

ABP/VOA, the matched ABP/VOA address returns from the same UPRN/VOA BA reference to the one held in CMOS.

2. **Of Concern**: A discrepancy exists between the CMOS UPRN/VOA BA reference and the matched UPRN/VOA BA reference, when matching the CMOS address to ABP/VOA.

The "Of Concern" category is further divided into the following subcategories:

1. **Conflicting** – The CMOS UPRN/VOA BA reference does not match to the CMOS address. The CMOS address has been 'Verified' to ABP/VOA, but the UPRN/VOA BA returned for that address does not match the one held in CMOS. A review is required to update the CMOS UPRN/VOA BA.
2. **Unconfirmed** – The CMOS UPRN/VOA BA reference does not match to the CMOS address. The CMOS address cannot be 'Verified' to ABP/VOA. It is not possible to confirm if the UPRN/VOA BA in CMOS is correct. A review is required to determine the correct address and corresponding reference(s).
3. **Appended** – There is no UPRN/VOA BA held in CMOS. The CMOS address has been 'Verified' and a corresponding reference has been found. A review is required to update the CMOS reference(s).
4. **None Found** – There is no UPRN/VOA BA held in CMOS. The CMOS address cannot be 'Verified', and therefore no corresponding UPRN/VOA BA can be proposed with confidence. A review is required to determine the correct address and corresponding reference(s).

Preparing the data for Address Matching

The received data file undergoes several preparatory processes before the address matching is carried out. This reduces the reliance on complex matching algorithms, thereby enhancing the speed and efficiency of the matching process. Additionally, these processes ensure consistency across all data sources, simplifying address comparison and search operations.

Preparatory Processes (Data Standardisation):

1. Common Address Terms
2. Ordinal Number Conversations
3. Uniform Connecting Words or Characters
4. Remove Special Characters
5. Convert to Uppercase
6. Remove Unnecessary Customer Names
7. Remove Leading and Trailing Spaces

8. Format Date Numbers
9. Format Postcodes

Common Address Terms

Procedure: Replace the full version of a common address term (e.g., "road", "street", "avenue") with the standardised abbreviation for the corresponding address term (e.g., "rd", "st", "av").

Example: The address "123 Main Street" is processed to become "123 Main St" using the catalogue of abbreviations, ensuring that all addresses follow a consistent format.

Catalogue:

Full Term	Abbreviation
Road	Rd
Street	St
Avenue	Av
Ave	Av
Park	Pk
Grove	Gr
Drive	Dr
Close	Cl
Place	Pl
Court	Ct
Square	Sq
Lane	Ln
Head Quarters	HQ
Way	Wy
Terrace	Ter
Parade	Pde
Estate	Est
Green	Grn
Gardens	Gdns
Crescent	Cres
Boulevard	Blvd
Buildings	Bldgs
Apartment	Apt
Apartments	Apt
Units	Unit
Floor	Flr
Ground	Gnd
Grnd	Gnd
Part	Pt
Prt	Pt

Basement	Bst
Crt	Ct
Block	Blk
R/O	Rear Of
RO	Rear Of
W/Shop	Wshop
Workshop	Wshop
W/Shp	Wshop
O/S	Outside
Grge	Gge
Garage	Gge
Station	Stn
W/Hse	Whse
Warehouse	Whse
Village	Vlg
Industrial	Ind

Ordinal Number Conversations

Procedure: Replace abbreviated ordinal numbers with their spelled-out counterparts.

Example: The address **"The Barn, 1st St"** is processed to become "**The Barn, First St"** using the catalogue of abbreviations.

Catalogue:

Ordinal numbers	Word Form
1st	First
2nd	Second
3rd	Third
4th	Fourth
5th	Fifth
6th	Sixth
7th	Seventh
8th	Eighth
9th	Ninth
10th	Tenth
11th	Eleventh
12th	Twelfth

Uniform Connecting Words or Characters

Procedure: Replace various conjunctions and symbols commonly present in addresses with a hyphen (-) or the word AND.

Example: The address line "Units 1 & 2" is processed to become "Units 1 AND 2," while "1 TO 3" is transformed into "1-3," using the specified replacements.

Catalogue:

Conjunctions and Symbols	Replaced by
Ampersand (&)	AND
Plus sign (+)	
" TO "	Hyphen (-)
Hyphen with spaces (-)	
Forward slash (/)	
At sign (@)	at

Remove Special Characters

Procedure: Remove all symbols (all non-alphanumeric characters), except when the non-alphanumeric character appears between two digits. This ensures that characters between digits (e.g., "APT 4-1", "SUITE 5.8 FIFTH FLOOR") are retained.

Example: The address line **" 430 (Wheelock Station) Crewe Road"** is processed to become "**430 Wheelock Station Crewe Road "**

Convert to Uppercase

Procedure: Convert all characters in address lines to uppercase.

Example: The address line **"5 Forest way"** is processed to become **"5 FOREST WAY"** by converting all letters to uppercase.

Remove Unnecessary Customer Names

Procedure: Remove unnecessary customer names from the **Customer Banner Name** and **Customer Name** fields.

Catalogue:

Unnecessary Customer Names

THE OCCUPIER
NO CUSTOMER
NO OCCUPIER
NO-OCCUPIER
VACANT
THE TREASURER
EMPTY PROPERTY
VACANT PREMISES
TREASURER
THE HEAD TEACHER

Remove Leading and Trailing Spaces

Procedure: Remove any leading or trailing spaces from address lines.

Example: The address lines **" 123 The Grove "** & **"Holton at Sea "** is processed to become "**123 The Grove**" & "**Holton at Sea**" by removing the unnecessary spaces at the beginning and end of the string.

Format Date Numbers

Procedure: Reformat a house number string that may have been incorrectly interpreted as a date into a numeric range format. This procedure is only applied to CMOS address lines 1 and 2.

Example: The address line **"01-Oct Williams Way "** is processed to become "**1-10 Williams Way "** by parsing the numeric portion of the string and converting it to a range.

Format Postcodes

Procedure: Clean and standardise postcode data to align with the MOSL format, as per the Data Assurance Service Premises and Addresses Good practice guide by:

1. **Replacing consecutive whitespace**: Substitute all consecutive whitespace characters (e.g., spaces, tabs) with a single space.
2. **Triming whitespace**: Remove any leading or trailing whitespace characters from the postcode.
3. **Correcting missing spaces**: Identify postcodes that lack a space between the area and district parts using a regular expression and insert the missing space.

Example: The postcode **"m12ab "** is processed to **"M1 2AB"** by standardising whitespace, trimming, converting to uppercase, and inserting the necessary space, ensuring the format adheres to the advised standard.

Remove Duplicated Address Line Data

Procedure: Identify and removing duplicate address components within each row, ensuring only the first instance is retained by:

Steps:

1. **Identify Duplicates:** Scan each address string for repeated address components (e.g., street name, locality, or building name).
2. **Remove Redundant Data:** Replace any duplicated address elements within the string with an empty space, keeping only the first occurrence.
3. **Trim Extra Spaces:** Remove any unnecessary spaces created by the removal process to maintain a clean format.

Example:

The address string **"10 High Street, High Street, London, London"** would be converted to "**10 High Street, London"**

Non-addressable Identification

Non-addressable sites are identified within the address matching process to support with the categorisation of “Unmatched”.

(***Unmatched****: The SPID address is not found in the reference data sources and is identified as a Non-Addressable Site (e.g., bin stores, public conveniences, allotments, or features* like *troughs and stand-pipes)*.

Procedure: Categorise addresses by identifying specific keywords and phrases within them, using these categories to determine if an address is a non-addressable site in accordance with the guidelines outlined in **Data Assurance Service Premises and Addresses Good practice guide**.

Additionally, to prevent valid addresses from being mistakenly flagged as non-addressable, a search is conducted for keywords or phrases that may resemble those associated with non-addressable categories, ensuring the address is recognised as deliverable.

Example:

Non-addressable

The address "F/SUPPLY, Meadow Lane, Springfield, SP1 2AB" is categorised as a **"Field Supply"** non-addressable site, as it contains the keyword **"F/SUPPLY".**

Deliverable Address

The address **"123 Field Lane, Mooresville, M44 1GH"** would not be categorised or flagged as a non-addressable site.

Catalogue:

Category	Exclusions
Toilets	CONVENIENCE, TOILET, TOILETS
Non-Delivery Point	ROUNDABOUT, TRAFFIC ISLAND, Feeder Pillar, Street Lighting, MOORING
Household Property	FLAT, APARTMENT, APT, FLT
Land	LAND ADJACENT, LAND AT, ADJ, Opp, NR, Rear, R/O, R-O, ADJACENT, LAND ADJOINING, LAND ADJOURNING, LAND NEXT TO, ALLOTMENT, ALLOTMENTS
Parking	CAR PARK, CAR SPACE, CAR PARKING, CAR SPACES, C/SPACE
Agricultural	SUPPLY, CDT, COMMUNAL OUTDOOR SUPPLY, COMPOUND SUPPLY, FARM SUPPLY, TRGH, TROUGH, YARD SUPPLY
Communal Supply	Bin, Building Supply, L/L, L/L SUPPLY, Landlord, LL SUPPLY, BINSTORE, Meter, MTR, landlords supply, COMMUNAL, L/LORD, LL SPY, L/LORDS SUPPLY
Temporary Premise	PLOT, SITE OFFICE, SITE SUPPLY, TEMP SUPPLY, TEMPORARY BUILDERS SUPPLY, TRADE SUPPLY, TEMP BUILD SUPPLY, TEMP BUILD SPLY, TBS, TEMP BUILD SUPPLY
Treatment Plant	TREATMENT PLANT
Water Supply	SEWAGE, SUB STATION, SUB STT, SUB-STATION, TREATMENT PLANT
Advertising	ADVERTISING HOARDING
Garage	THE GARAGES, GARAGE
Inactive	DERELICT, DEMOLISHED
Leisure	CARAVAN
Telecoms	CELL SITE, TELEPHONE EXCHANGE
Fountain	FOUNTAINS, FOUNTAIN
Field Supply	F/SUPPLY, FIELD OPPOSITE, FIELD OS, FIELD SPLY, FIELD SUP, FIELD SUPPLY, FIELD TANK, FLDS, HORSE PADDOCK, PLAYING FIELD, SPORTS FIELD, WATER POINT, WATER TANK

Address Matching

After the standardisation is complete, all addresses progress through the address matching process. This process takes the cleansed CMOS address, and compares them against addresses held within the ABP and VOA datasets.

Initial Checks

#	Stage	Check Performed	Status	Outcome
1	**Postcode Validity**	Check if the postcode in the CMOS (client data) exists in the two external data sources (ABP and VOA).	PASS	Pass: Move to next stage
			FAIL	Fail: Marked as "Of Concern"
2	**Street-Postcode Match**	Verify if the street and postcode combination in the CMOS matches the combination in ABP or VOA.	No restriction: Move to next stage (regardless of match).	

Regardless of whether the CMOS address is flagged for a Street-Postcode Match failure, additional address-matching checks are conducted. This ensures that differences between datasets, such as minor textual discrepancy in street names, are accounted for.

Address Line Concatenations

For all addresses with valid postcode identified, the CMOS addresses undergo the following:

Step 1: Identifying Possible Matches

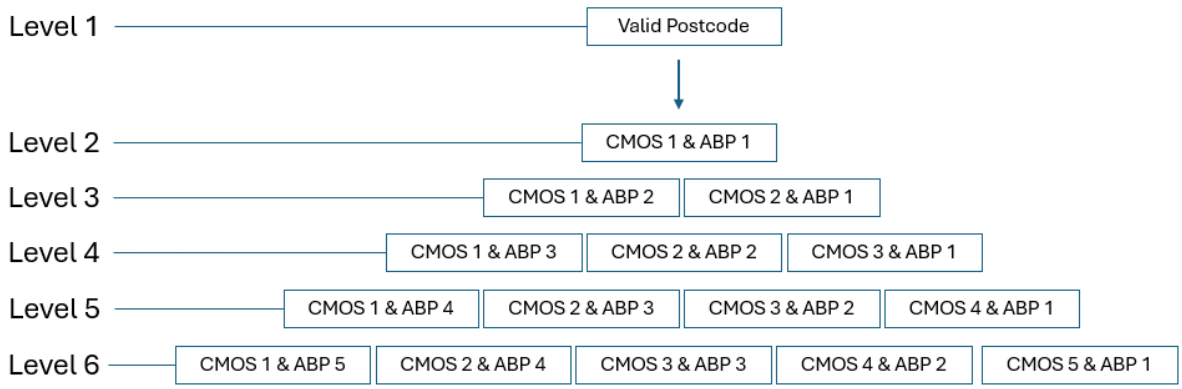
For each address in CMOS, the postcode is used to retrieve all possible matching addresses from the ABP and VOA datasets. These potential matches are extracted and stored for comparison.

Step 2: Create Address Variations

To improve matching accuracy, different address elements are combined in various ways using predefined concatenation catalogues. This process generates multiple address variations for each CMOS, ABP, and VOA address.

Step 3: Hierarchical Matching Approach

The concatenated address variations from CMOS are matched against those from ABP and VOA using a structured pyramid ordering scheme. This approach prioritises address concatenations that contain the most comprehensive address elements first before progressively expanding to broader comparisons. The pyramid structure follows this order:



This diagram illustrates a hierarchical address concatenation process used for matching CMOS and ABP addresses, with CMOS taking priority in the validation sequence. The same structured approach applies when matching CMOS addresses against VOA data.

Concept Overview:

- The process starts with a valid postcode at Level 1.
- Address concatenations are formed by progressively combining CMOS and ABP address components (or CMOS and VOA address components in the VOA process).
- CMOS addresses are prioritised, meaning they are always considered first when forming combinations.
- The hierarchy expands as more concatenation variations are tested to find a match.
- The pyramid structure continues until a match between addresses is identified, or all possible concatenation combinations have been exhausted, whichever occurs first.

- This hierarchical approach ensures that the most relevant matches are found first.

There are three versions of the concatenations generated, “A” (with spaces) and “B” (without spaces) and “C” (without catalogued stop words). The process follows this order:

- **CMOS 1A** is checked against **ABP 1A**
- **CMOS 1B** is checked against **ABP 1B**
- The process continues with the subsequent address lines in the same order, i.e., **CMOS 1A** with **ABP 2A**, followed by **CMOS 1B** with **ABP 2B**, and so on.
- **“C”** concatenations are then reviewed

Stop Words Catalogue:

Stop Words
NE_Landlord
NE_Nosewerage
NE_Nowater
Comms*
2nd Supply Unit
Linked
1 Mtr
1x
LL
Merged
See
Shop
WHCommServ
WORKSHOP OFFICES & PREMISES
WORKSHOP OFFICES
OFFICES
OFFICE
LTD
ARCH
PREMISES
&
AND
UNIT
AT
**

1@
SMC*
Temp*

Variation Checks

For each combination, a variety of further checks are performed to support with the accuracy of address verification. This is a structured approach that involves multiple validation techniques. These checks are designed to handle various inconsistencies, formatting differences, and potential errors in address records.

Checks are performed for:

1. Exact Match
2. Range Match
3. Exact Match Without Dashes
4. Word-Only Match
5. Alpha Range Match
6. Levenshtein Match

Where an address is matched though any of the concatenations and variation checks, a further check is conducted to minimise the risk of false positives.

Address lines from CMOS data, ABP and VOA are compared and must meet the following criteria:

- The check digits, including any overlap in the range, should align.
- The terms “Rear,” “Flr,” and “BST” must be present in both the CMOS address and the corresponding matched address.
- The terms “Unit,” “Office,” and “Stall” should be checked first to see if they exist in either address. If no match is found for these terms, a match without them is still acceptable.

Parent Properties (ABP Only)

Parent properties are not given priority in the ABP matching logic. When all address concatenations and variations have been checked without finding a match, parent properties are then considered to determine if a match can be identified.

#	Stage	Check Performed	Status	Outcome	Match Status
1	**Exact Match**	Address lines from CMOS, ABP, and VOA are concatenated and compared for exact matches.	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_EXACT
			NO MATCH	Proceed to further checks	
2	**Range Match**	Check for overlapping numeric ranges in address lines, allowing for differences in pluralised words such as Units/Apartments. Expands to a fuzzy check on words, with a 85% match tolerance where no other match is identified.	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_RANGE
			NO MATCH	Proceed to further checks	
3	**Exact Match Without Dashes**	Replace dashes with spaces in all address lines and check for exact matches.	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_EXACT_W/O_DASH
			NO MATCH	Proceed to further checks	
4	**Word-Only Match**	Check for word-level matches, irrespective of word order (and after the removal of duplicated words)	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_WONLY
			NO MATCH	Proceed to further checks	
5	**Alpha Range Match**	Checks for overlapping alpha ranges in address lines where the words exactly match.	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_ALPHA
			NO MATCH	Proceed to further checks	

6	**Levenshtein Match**	Allows matches on address strings to pass where the string is greater or equal to 10 characters, with a character difference of up to 3 characters.	MATCH	Marked as "Verified"	CLIENT_CONCAT_#A_ABP_CONCAT_#A_LEVE
			NO MATCH	Mark as “Of Concern”	

Additional matches through comparison scripts

When an **address has been verified** due to a confident match with either ABP or VOA, two additional checks are performed to facilitate a match with the other dataset.

#	Stage	Check Performed	Status	Match Status
1	**Address Base - VOA mapping**	When an ABP match is identified, the corresponding VOA should be matched using the ABP-linked VOA (and vice versa)	MATCH	Matched_ABP_Compare
			NO MATCH	
2	**CMOS Ref Check**	When an ABP match is found and a valid CMOS VOA BA Reference exists, compare the address of the CMOS VOA BA Reference with the CMOS premises address (and vice versa)	MATCH	Matched_Client_Reference_Compare
			NO MATCH	

Match Insight Status & Sign-posting

Match Insight Status

When a match is identified, the match status is updated with the concatenation method used, and the additional variation check status is appended at the end.

If the match is determined through comparison script logic, the Match Insight Status will be one of the following:

- Matched_ABP_Compare
- Matched_Client_Reference_Compare

Sign-Posting for Unmatched Addresses

If no match is found, a Sign-Posting status is assigned to assist Wholesalers in identifying potential address discrepancies and determining the appropriate next steps.

No Match Found Status	Description
Invalid_Postcode	The address matching process could not identify a match because the postcode has been deemed invalid (it does not exist within the ABP/VOA datasets)

Invalid_Street_Postcode	The address matching process could not identify a match because the specified postcode and street combination does not exist in the ABP/VOA datasets
Invalid_Street_Start	The address matching process could not identify a match because the CMOS address starts with the street name, but no valid premise identifier is available to determine the property on the street
Reference_Number_Conflict	The address matching process could not identify a match and the previous three statuses have not been identified as failure reasons. Therefore, a comparison between the ABP/VOA reference addresses and the CMOS address has revealed discrepancies
Other	The address matching process could not identify a match. The reason is yet to be determined