Project Tide
Data Quality Assessment Headline Findings

CMOS data analysis

CMOS Data Validation	Insights
MOSL MARKET OPERATOR SERVICES LTD **4.5M Records**	**2.81M Supply Data Records** **1.64M Premise Data Records** The raw CMOS dataset provided by MOSL to Sagacity consisted of two files: • The Supply file, which contained SPIDs and the Retailer and Wholesaler information • The Premises file, which contained address and customer information The datasets were joined to enable analysis on one 'master' dataset
Sagacity	Sagacity performed a cycle of pre-cleansing and analytics to removed invalid data from the master dataset. Deregulated SPIDs, empty addresses and superfluous data was removed before making links between Water and Sewage SPIDs where addresses were the same. This reduced the volume of SPIDs by over 200K from the original 2.81M provided
Results	A total of 2.6M SPIDs were identified and included in the Data Quality Assessment. Using this dataset the number of premises and addresses within the market were able to be identified and are broken down below:
2.6M SPIDs	**1.38M Water SPIDs** **1.22M Sewage SPIDs** • A Premises is a location • A Premises can have 1 or more SPIDs • Each SPID has an Address → We have had to analyse 2.6M addresses attached to each SPID to identify discrepancies, duplicates and unique addresses
1.49M Premises	**1.11M with Water & Sewage SPIDs** **268K with Water SPID Only** **108K with Sewage SPID Only** The analysis conducted shows that a total of 376K Premises have either a Water SPID or a Sewage SPID attached, and not both → This is undergoing further investigation to understand why both Water & Sewage SPIDs are not attached to the same Premises e.g. Unpaired Supply Points
1.68M Addresses	**913K Premises with a matching W&S Address** **197K Premises with a differing W&S Address** **376k Water or Sewage Only Premise** Our analytics has identified 1.68M unique addresses that were required to be processed through Sagacity's Address Matching software. This could in part be a result of the addresses being captured by multiple Trading Parties for SPIDS ***Note**: There are 449 different combinations of Wholesalers and Retailers within the Non-Household Water market*

Key findings

Market Eligibility

1 **50k residential premises** and **35k demolished premises** in the non-household market
2000+ new commercial properties identified in last quarter (from New Properties DB) that are not in CMOS

Premise Accuracy

2 **870k SPIDs** are **missing a UPRN** (a further **610k** have issues)
1.34M SPIDs are **missing a VOA** (a further **705k** have issues)

Address Accuracy

3 Only **58 per cent** of CMOS supply point **addresses** are of **billable data quality.** 16 per cent (**415k**) do not match to any external data set

Trading Parties

4 **Data quality issues are widespread**. The top nine wholesalers (based on SPID volume) all had significant issues, with best performer achieving 70 per cent and the worst 58 per cent Data Quality score

Occupancy Status

5 **459k vacant SPIDs** in CMOS, but 45 per cent of these (**209k SPIDs**) show signs of active business

Customer Accuracy

6 **34 per cent** of **SPIDs (879k)** have no discernible customer name and **343k (20 per cent)** were deemed incorrect. A further **481k** require further validation.