MOSL
MARKET OPERATOR SERVICES LTD

Central Data Cleanse consultation

23/11/2022

Contents

Introduction

Purpose

This consultation sets out the work undertaken by MOSL to explore the value a centralised data cleanse service could deliver to the non-household (NHH) market, including:

- Details of the work undertaken by MOSL and Sagacity to assess market data quality
- The benefits case for a central data cleanse service supporting address and customer data cleansing
- The proposed solution and phasing of work
- The proposed funding model.

MOSL is seeking feedback from stakeholders to refine the final service proposal that will be presented in our draft 2023-26 Business Plan for consultation in January.

Written responses to the below questions are requested by email to matt.labrum@mosl.co.uk by 5pm on Wednesday 7 December. A template for the consultation submission is provided on page 23.

1. Do you agree that a centralised data improvement service (focused on address and customer data) would benefit customers and market participants?
2. Have we captured all material benefits? If not, what else should be considered?
3. Do you use any third-party services/products to improve or maintain your customer or address data quality? If so, what is the approximate annual spend related to the non-household market?
4. Do you agree with the proposed scope and phasing? If not, please propose an alternative.
5. Do you agree with the proposed funding model? If not, please propose an alternative.
6. Do you have any additional comments you would like to make?

We will publish a redacted, collated view of the answers received on the MOSL website after the consultation closes.

"All market participants need to improve market frictions – in particular on data quality. We are pleased to see MOSL drive progress here and urge all stakeholders to engage constructively in this consultation and to ensure improvements to market data are made quickly and effectively." – Georgina Mills, Director, Business Retail Market at Ofwat

Executive summary

This consultation sets out the following:

- **The case for a central cleanse** – General consensus from market participants on the current issues and the need for complete and accurate address and customer data, and the value of a central service (with conservative annualised benefits in excess of £8M)
- **The prioritisation of eligibility, premises, and address data** – Our engagement suggests trading parties believe removing non-eligible premises and improving address data are the immediate priorities
- **The initial phase being funded by wholesalers** – Following conversations with several trading parties, there is general consensus that Phases 1 - 3 should be funded by wholesalers (based on ownership of address data) and the simplest route would be through increased wholesaler Market Operator (MO) charges
- **The cost to deliver a central service is circa £750k - £1m for the first year** – This cost covers the service and MOSL resourcing only
- **Consideration of resource requirements, appropriate phasing and realistic timescales** – Trading parties have highlighted that resource requirements are likely to individually exceed the disaggregated cost of the service. We also recognise that this will extend to increased MOSL resource requirements.
- **Trading parties resource levels** – trading parties have said that timescales will be a challenge, stating the need for a multi-year programme with sustained central support.
- **Parallel process and code improvement requirements** – Alongside the central service, market processes and codes must be reviewed to ensure enduring improvements in data quality. Any changes or improvements taken forward will still retain the current trading party data ownership model.

This consultation recommends that a central data improvement service be provided, with the initial priority being to cleanse eligibility, premises, and address data. These address/premises data items are wholesaler owned and as such wholesalers are considered the most appropriate funding mechanism for the initial cleanse service at a cost of up to £1m per annum. Further information is set out in the consultation document.

Background

Accurate and reliable data is critical to the effective operation of the non-household (NHH) market and is key to operational efficiency, evidence-based improvement, and positive customer and environmental outcomes. However, data quality remains a principal market friction that is driving significant cost for trading parties and impacting the market's ability to deliver better outcomes.

Data ownership is complex due to a mismatch between those who own data fields (and bear the cost of maintaining them) and those who rely on or derive value from the data. As an example, wholesalers register and maintain the supply point address data fields, retailers rely on that data to be accurate and complete to determine whether those premises are occupied.

This complexity has led to market wide issues with completeness, validity, and accuracy of key data fields in the Central Market Operating System (CMOS). Data quality and the impact of poor data has been consistently highlighted in industry consultations and market publications over the past three years, including Ofwat's State of the Market reports.

A few examples of key points raised that underpin the need to improve market data quality are set out below. Further examples are available in Appendix one.

Review of incumbent company support for effective markets (RISE Report, Ofwat), August 2020

Ofwat's RISE report examines the improvements made by incumbent water companies in their support for effective markets. Within their findings, Ofwat identified poor data quality as a "significant market friction" requiring "urgent action from all trading parties". It highlighted that "poor quality data can undermine the achievement of improved outcomes for customers, society, and the environment in the following ways:

- Poor quality customer, asset and consumption data can lead to incorrect bills (increasing time costs for customers and undermining customer satisfaction)
- Poor quality data can undermine efforts to use water more efficiently (water efficiency will benefit the environment as well as customers and companies)
- Poor quality data can also lead to some customers paying too much or too little (for example where occupied sites are not charged or customers pay on the basis of inaccurate estimates)."

Request for Information: Core Market Data Cleanse (MOSL), October 2020

Following our commitment to deliver a data improvement plan for core market data items as part of our Market Performance Operating Plan (MPOP) 2020/21, MOSL published a request for information (RFI) to understand: (i) the cost-impact and benefits of data quality; (ii) the required activities for mitigating or maintaining data quality; and (iii) the recommended next steps for resolving data quality issues.

The focus areas were customer details and premises data, meter location data and meter details data. It was determined that:

- Eight out of 10 retailers and 10 out of 14 wholesalers said they are adversely impacted by poor quality customer and premises data

- The estimated minimum annual resource cost to the market from managing poor quality data was £10m; with customer and premises details and location data each accounting for approximately 40 per cent of this cost, and meter details data 20 per cent
- Trading parties are deploying resource to obtain and verify missing or inaccurate data items; administer bilateral requests; manage consistency and mismatches between central and internal systems; resolve inaccurate or incomplete supply point data; address high volumes of failed meter reads; and manage meter verification requests.

Non-household Retail Market Study (Economic Insight), April 2021

The UK Water Retailer Council (UKWRC) commissioned Economic Insight to develop a market assessment report to identify why the market may not be operating effectively, and to propose reforms to improve outcomes for customers. Amongst numerous challenges, the report concluded:

"We have some concerns regarding the quality of market data. The evidence suggests that these [market] frictions may be increasing operating costs and decreasing service quality for customers. Based on our review of the available information, we understand that the data issues currently apparent in the market are driven by the combination of poor-quality data at market opening, and ineffective action to remedy this. The updates and corrections to market data have not been satisfactory in fixing these issues. The current arrangement from the market operator is such that retailers and wholesalers both hold responsibility for maintaining and updating the market data. However, we understand that the efforts made by both have not been effective in fixing the issues completely. The data issues are complex and overlapping, as such it is difficult to fully assess the scale of the issues."

Review of the fifth year of the business retail water market 2021-22 (Ofwat), September 2022

In Ofwat's annual review of the non-household market, it provided an assessment of the industry's efforts towards improving market frictions, noting the following:

"Poor quality customer, consumption and asset data can significantly undermine the customer experience. Customers want timely and accurate bills, and this is simply not possible if the quality of market data is poor. Good quality data is also a crucial enabler of innovation, including in relation to water efficiency…Improving the quality of market data therefore remains a key and urgent priority for the market.

The business retail market is now five years old, and we urge all market participants to continue to drive progress aimed at getting these basics right, in particular to improve the quality of market data as well as trading party performance. We would like to see stronger incentives on trading parties to improve data quality included in the Market Performance Framework (MPF) to incentivise further improvements. Trading parties have code and licence obligations in relation to the provision and maintenance of good quality data, which

they must comply with, and financial incentives can be a useful means of further incentivising compliance with these requirements".

Background summary

These external reports, assessments and RFIs demonstrate that:

- Accurate data is essential to market functions and the delivery of improved outcomes for customers
- Poor quality data is causing friction between parties, and efforts to address have not been effective
- Poor quality data creates significant and increasing cost for trading parties
- The value of a centralised data cleanse has been evidenced by the Scottish Market (see Appendix One) and further supported by numerous conversations with trading parties
- Current data quality initiatives can only focus on completeness and are unable to assess or improve accuracy
- Data quality is a key enabler for the successful market outcomes identified in the Market Performance Framework (MPF) Reform
- Accurate data is a key enabler to identify outlier water usage to drive water efficiency. Accurate data enables accurate demand forecasting and is key to planning for secure, sustainable supply of water.

Project Transformation in Data Enrichment (TIDE)

Project TIDE was established in April 2022 by MOSL and Sagacity to look at how data quality could be improved with greater pace, consistency, and efficiency to deliver strategic market outcomes and mitigates risks and issues, by:

- Assessing the quality of premises, address, and customer data within CMOS
- Understanding how a central data cleanse service would allow trading parties to improve data
- Defining a case for change for a central service.

Rationale

Trading parties have questioned the logic and cost of working independently to fix core data fields, with the current rate of improvement being too slow for the intended benefits of the market to be realised.

This led MOSL to consider whether a centrally managed data cleanse and enrichment service could better facilitate data quality improvement at pace. The potential benefits of the service include:

- Quicker and more efficient data quality improvement
- Reduced cost-to-serve and administration for all trading parties
- Reduced third party data management services spend through economies of scale
- Increased revenue through occupancy verification and gap site identification
- Increased market competition and innovation through enhanced tendering capability
- More accurate insight that could drive improved water management and better outcomes for customers and the environment
- The ability to objectively assess trading party performance with respect to data accuracy
- Ability to identify market trends and patterns with certain data items or processes that may require market-level improvement.

We investigated a range of potential use cases and identified the following eight for further evaluation:

1. The ability to confirm retail market eligibility in CMOS (e.g., identify residential or demolished)
2. The ability to validate and identify premises reference data (review or find Unique Property Reference Number (UPRN) and Valuation Office Agency (VOA) reference data)
3. The ability to validate premises address data
4. The ability to confirm and track premises' occupancy
5. The ability to identify and maintain customer name
6. The ability to identify customer industry classification
7. The ability to identify gap sites
8. The ability to identify unmatched supply points / erroneous data.

Scope of Project TIDE

Following a tender process, we appointed data specialists Sagacity, to undertake a pilot.

We worked with Sagacity over a 12-week period to:

- Understand the key customer and address data items needed for successful market operation

- Assess the current data quality across all trading parties
- Validate cleansed data with a small pilot of trading parties
- Define the solution (process and technology) for a central cleanse service
- Develop a service proposal (case for change) that will inform a consultation.

Data Quality Assessment findings

Sagacity assessed all CMOS supply points, exploring the quality (accuracy and completeness) of customer and premises data. The key address and customer fields considered are listed below and outlined in more detail in Appendix two.

- UPRN
- UPRN Reason Code
- VOA Billing Authority (BA) Reference
- VOA BA
- Reference Reason Code
- Address Line 1-5
- Postcode
- Occupancy status
- Customer Name
- Customer Banner Name
- Standard Industry Classification (SIC) Code
- SIC Code type.

The assessment was completed by matching CMOS data against a range of reference data sets, including:

Google

A key source of accurate company name information and provides the operational status of many companies. Google also holds additional company status data, such the last time the company was reviewed by a customer or changed its information.

Address Base is a database provided by Ordnance Survey, which houses property data. Each property is assigned a UPRN which is maintained throughout its lifecycle. This allows the history of property changes to be tracked and maintained. Address Based Premium (ABP) also contains supplementary information, such as XY Coordinates and updates from the Royal Mail Postal Address File (PAF).

PAF informs Address Base (amongst others) and is the source of most address changes, as it is updated daily by postal workers. It also includes data from local authorities and property developers.

The VOA maintains the business rates data which is used by local councils to calculate the business rates of a property. This provides a high level of confidence around market eligibility, and whilst it does not contain all business properties, it does support eligibility decisions and is a verification tool in reviewing supply point addresses.

The full data assessment report (against a May 2022 CMOS extract) can be accessed here. Key findings are as follows:

1. 50,000 residential premises, 35,000 demolished premises and 20,000 duplicated premises were identified (with high confidence) in CMOS
2. 870,000 supply points are missing a UPRN and a further 610,000 supply points may have an incorrect UPRN
3. 1,340,000 supply points are missing a VOA number and a further 705,000 supply points may have an incorrect VOA

"UPRN and VOA population enables us to accurately align our SPIDs to the current list of eligible premises, which helps ensure that we charge retailers and customers accurately and fairly. We are better able to proactively identify potential gap sites, de-registrations, change-of-use and splits/mergers, reducing the reliance solely on the premises address, helping us to become a more data-driven wholesaler. VOA and UPRN population also provide confidence in alignment of household and non-household data sets, avoiding billing gaps or duplication. There are also longer-term benefits including the potential for data sharing between different organisations to validate occupancy, information on customer type to support water resource forecasting and the management of sensitive customers between utility companies." – Louise Rutherford, Project Lead, Business Market at United Utilities

4. Only 58 per cent (1,500,000) of supply points confidently match external sources (Google, VOA, UPRN and Royal Mail). 26 per cent (688,000 supply points) are a lower confidence match, due to poorer data quality. This results in difficulty matching or verifying the data across one or more of the external sources. 16 per cent (415,000 supply points) of CMOS supply points do not match to any sources

5. There are 459,000 vacant supply points in CMOS, but 45 per cent of these (209,000 supply points) show signs of business activity or occupation

6. Over 2,000 new commercial properties identified in Q1 2022 from the New Properties database register had not been registered in CMOS as at May 2022

7. 34 per cent of supply points (879,000 supply points) have no identifiable customer name and 20 per cent (343,000 supply points) were deemed incorrect. A further 19 per cent (481,000 supply points) require further validation

8. The top nine wholesalers (based on supply point volume) had significant issues with the quality of their premises and address data, with the best performer achieving a 70 per cent and the worst achieve a 58 per cent data quality score[1], indicating that data quality issues are widespread.

"[No data] is more important than premises and address data. These are the foundation data building blocks on which everything else is built… For us to function effectively and efficiently we need the triangulation of address, UPRN and VOA... These strong foundations help us establish which customer occupies the site and bring them into charge with confidence, resulting in a positive experience for the customer. With accurate address and premises data, we can also be certain of a vacant site and operate effective trace activity as well as identify duplicate supply points, gap sites and ineligible supply points. Furthermore, these data items also help us to establish if the services are correctly applied and if the meter details are accurate." – Trevor Nelson, Regulation and Compliance Manager at Business Stream

[1] To produce a data quality score each provider was graded on each address using a scale from an exact match being a full mark and an unmatched being no marks.

Proof of Concept (PoC) findings

We worked with three trading parties to review and validate a small sample of Sagacity's findings across each match level category (exact, definitive, probable, potential, unmatched), covering residential properties, demolished, UPRN, VOA and Address issues.

From the sample, the following data items were identified or corrected[2]:

- Either new UPRNs or VOAs were found or corrected across a third of sample cases
- 40 per cent of recommended address changes were updated
- 80 per cent of residential deregistration samples recommended by Sagacity were accepted by the trading party
- 100 per cent of demolished deregistration samples recommended by Sagacity were accepted by the trading party
- 50 per cent of duplicate deregistration samples recommended by Sagacity were accepted by the trading party
- 45 missing UPRNs were accepted out of 49 recommended by Sagacity to be added
- 50 incorrect UPRNs were corrected out of 56 recommended by Sagacity for correction
- 47 missing VOAs were accepted out of 53 recommended by Sagacity to be added.

In addition to validating the value of the central cleanse approach, the PoC highlighted additional considerations that should be explored in parallel:

- Address matching alone struggles with identifying non-addressable objects (e.g., taps, agricultural supplies, troughs), and there would be value in exploring alternative identifiers for this subset of supply points
- Address formats must be flexible enough to accommodate different premises types. For example, a high street business may have different address requirements to an office block or unit on an industrial estate
- Customer name can be an integral part of identifying/defining the address -businesses such as pubs and farms can be a key component of address verification.

[2] In some cases, Thames Water found an alternative UPRN/VOA from the one suggested by Sagacity or found a UPRN/VOA where none had been suggested from the matching process.

- The central cleanse service should provide a transparent view of matching confidence and underlying address data from the multiple sources where conflicts exist. This is key to allowing trading parties to act on the high confidence matches without significant further validation.

> *"Overall, the POC results were very positive, demonstrating that a central data cleanse function could potentially prove very useful for verifying and improving Premises data quality in the market"* - Mike Ward, Senior Data Analyst at Thames Water

Case for change

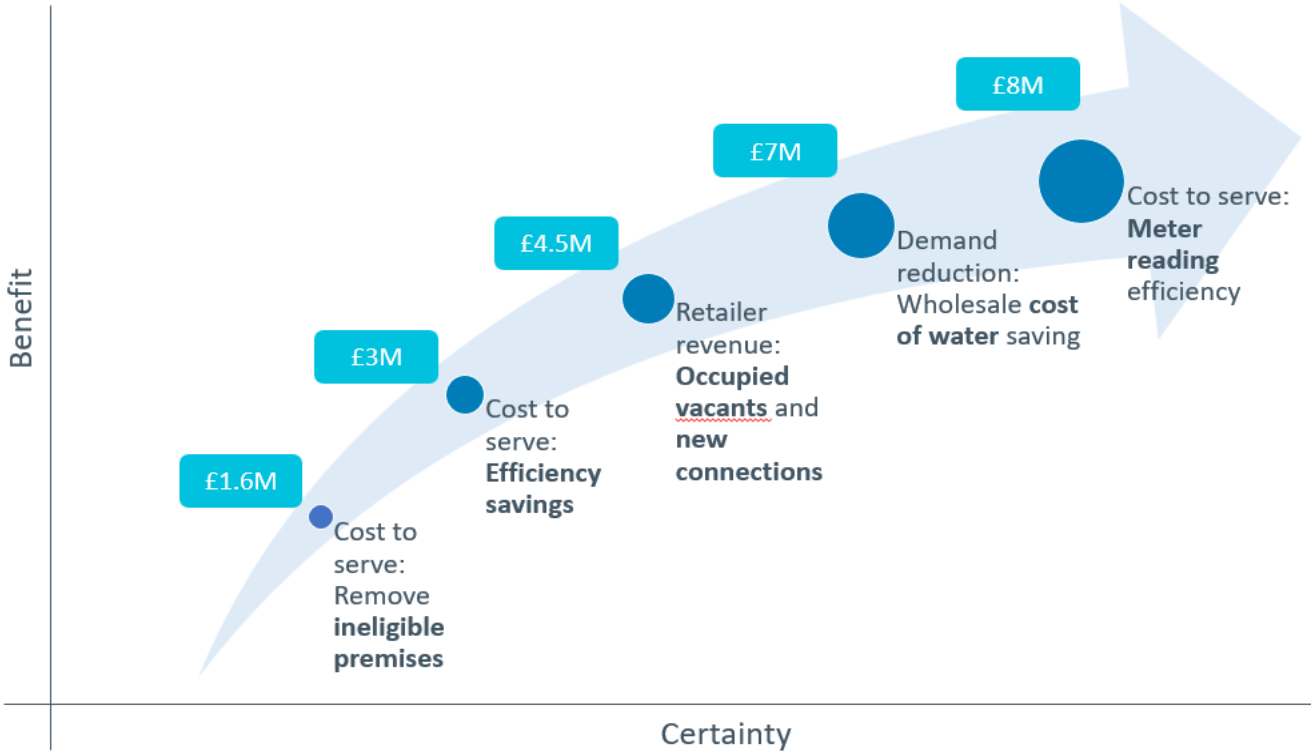
The primary benefits of a central data cleanse can be divided into three categories:

- **Reduced cost-to-serve:** Data quality improvement results in reduced administrative, resourcing or market costs, such as reduced meter reading costs, account administration or financial penalties
- **Improved revenue:** Data quality improvement results in identifying missing supply point identification or improved timings bringing supply points onto charge
- **Other:** Data quality improvement brings about numerous intangible benefits that may be difficult to measure such as water demand forecasting, improved customer experience and richer market insight.

Based on the outcome of the Data Quality Assessment and taking a conservative view of the potential impact, we have identified circa £8m of annual benefit to the market (although it is likely to be significantly higher). This excludes any Ofwat Outcome Delivery Incentive (ODI) upside for wholesalers.

We have taken a conservative view to also acknowledge the additional cost/effort trading parties may incur to correct this data at pace, for example, removing non-eligible premises could result in initial costs and administration for trading parties in the event of refunding retail charges to customers, refunding wholesale charges to retailers and deregistration of premises.

The graph on page 14 summarises the £8M annual benefit.



A summary of each benefit category is set out below, with examples of detailed calculations/assumptions.

Reduced cost to serve

- Remove SPID/Customer management costs for properties that shouldn't be in the market
- Reduced SPID/Customer management costs through "right first time" activities (registration, premises churn, occupancy, billing, deregistration, call handling, complaints handling and disputes)
- Reduced data management resourcing and data services costs
- Reduced meter reading costs (accurate address data)
- Reduced wholesaler cost to serve, accurate forecasting and effective demand reduction may mean wholesalers do not have to build additional infrastructure.

Revenue opportunity

- Additional retailer core revenue through identification of all active non-household customers (and reduced bad debt)
- Additional services revenue
- Additional wholesaler benefit through identification of all active non-household customers and reduction in assumed leakage (ODI) impact not included in graph above

- Foundation for innovation, enabling retailers to provide tailored and targeted services.

Other benefits

- Ability to monitor and incentivise trading party data quality performance
- Non-household demand reduction through enhanced targeted water efficiency intervention
- Improved non-household demand forecasting and planning
- Improved communications to customers during emergencies or unplanned outages
- In the event of a retailer failure, assisting the Interim Supply Allocation (ISA) process by ensuring customers receive a smooth transition
- Cross-utility (or cross-sector) data matching to provide more up to date occupancy information
- Fairer charging for customers (customers not subsidising gap site usage).

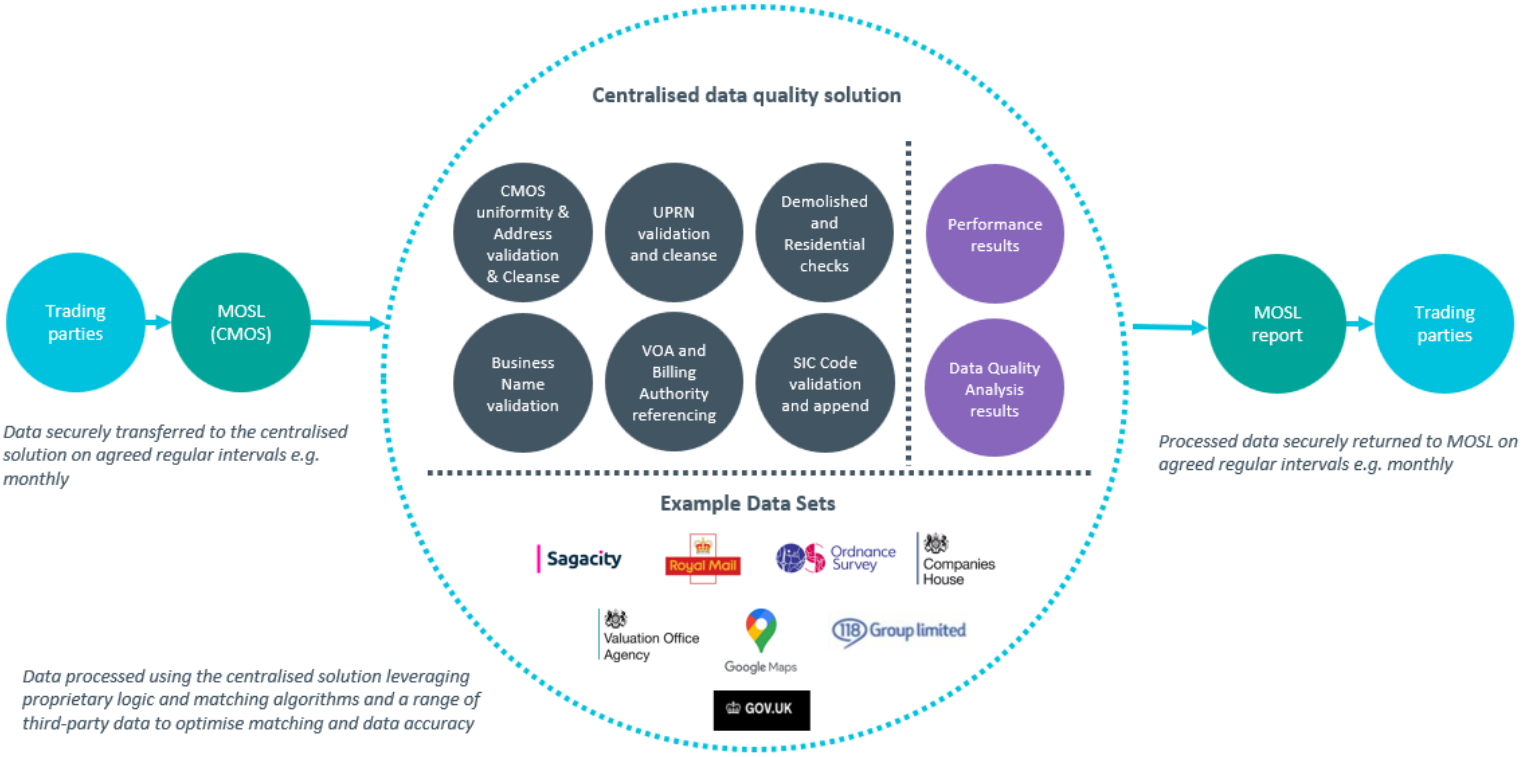
An example of a detailed calculation for "removing SPID/Customer management costs for properties that shouldn't be in the market" is provided below.

Issue	Benefit
• Over 100,000 SPIDs believed to be **Demolished** (35k), **Household** (50k) or **Vacant** (20k), with potential for further volume within lower confidence residential premise matches (118k) and the unmatched premises (415k) • These customers likely have higher cost to serve and higher bad debt exposure (wholesale charges already incurred) • Associated Market Performance Standard (MPS) charges.	• Assuming efficient cost to serve for retailers of £34 per supply point (REC23), this equates to circa £3.4M per annum. The rises to circa £4M based on upper quartile costs. • Assuming just **50 per cent** of these high-confidence indicators are **correct** (80 per cent from PoC) this equates to a cost-to-serve saving of **£1.65M** - **£2M** per annum • Removing the metering costs for these premises (assume £9 per annum for bi-annual reads) equates to **£450k.**

Proposed solution

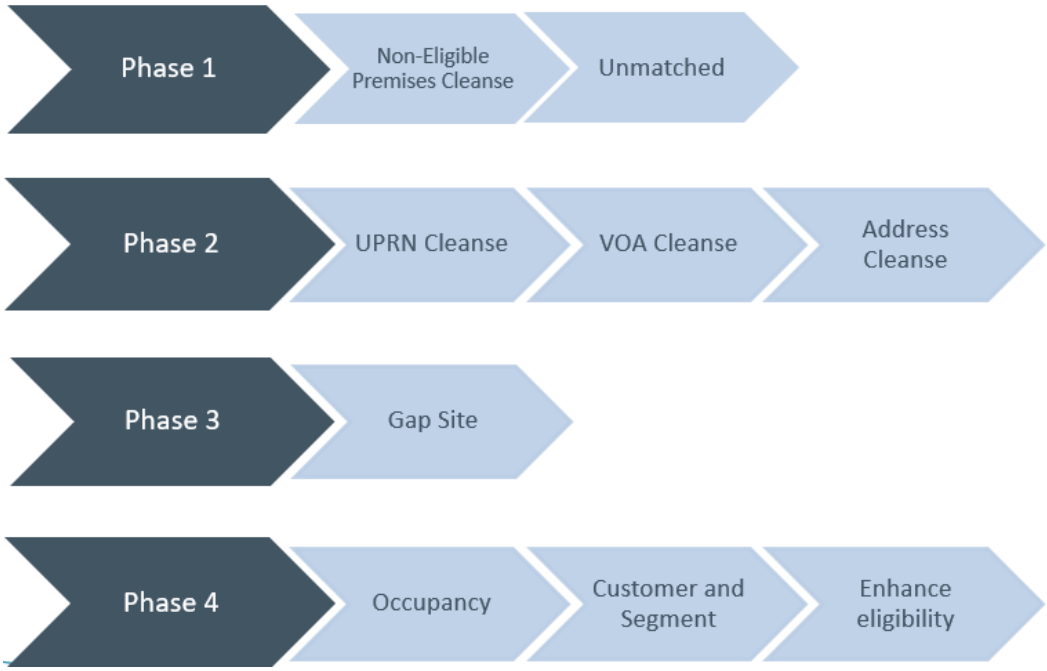
If a central cleanse service is established, trading parties will remain the data owners and responsible for making changes to data in CMOS. MOSL will be responsible for implementing, maintaining, and monitoring the service, incentivising data quality improvement, and ensuring data quality targets are being met.

MOSL will provide a managed service, in partnership with a data services provider. The proposed solution is summarised below. The detailed (or low level) technical design will be completed once funding is approved.



Data will be transferred to the third-party service provider, who will process and match against trusted third-party datasets to verify and assess the data quality across the core CMOS data fields. Data assessment findings will be returned to MOSL, these will then be distributed to the individual trading parties. MOSL will then track and incentivise trading party data quality improvement.

Proposed phasing



We propose a phased approach to implementing a central data cleanse, starting by removing ineligible premises and reviewing core address/premises data quality, before moving onto customer, occupancy, and gap site identification. This approach will allow the most efficient progress, whilst being mindful of the availability of trading party resources to implement the data field improvements assigned to them.

The length of each phase will be determined by the appetite for investment and the progress made by trading parties. The evolving scope of the enduring central data service will be agreed as part of MOSL's annual business plan, with year one (2023/24) being focused predominantly on Phases 1 and 2. It is expected that these phases will be delivered over a number of years, so we would expect Phases 3 and 4 to begin in parallel from year two (2024/25).

Phase 1: Non-eligible premises cleanse and unmatched supply points

The Data Quality Assessment identified many supply points that may need to be removed from the market. These supply points not only generate significant cost-to-serve for retailers, but also have a much higher propensity for common market issues (e.g., long unread meters, long term vacancy, meter location issues, address data quality).

It is also likely that a similar proportion of the unmatched premises will need to be removed from the market, driving further direct and indirect cost savings and improved customer experience. Phase 1 will therefore identify:

- Non-eligible residential supply points for market deregistration
- Non-eligible demolished supply points for market deregistration
- Non-eligible duplicate supply points for market deregistration.

Phase 2: Premises (UPRN and VOA) and Address

Once it is established that a supply point is eligible for the market, the next step is to ensure that identifiable and mappable reference data is provided so that its status and eligibility can be maintained and trading parties are able to fulfil their obligations to customers.

To achieve this, UPRN and VOA reference data needs to be complete and accurate. Phase 2 will focus on:

- Confirming supply point UPRN and VOA reference number accuracy
- Identifying and correct any incorrect UPRNs and VOA numbers
- Providing missing UPRNs and VOA numbers
- Identifying non-addressable premises (such as troughs, bin stores and standpipes)
- Confirming supply point address data accuracy based on external database matches
- Providing corrections for incorrect address data
- Passing any supply point address data that cannot be confidently matched or identified for correction back to the wholesaler to resolve.

Phase 3: Gap Sites

As we start to see material improvement in premises and address data quality, we can identify potential gap sites. Accurate and complete premises data would enable the market to cross-check data against the UPRN and VOA databases to identify gaps (including valid gap sites, such as mixed use and sub-metered premises). Phase 3 will look to:

- Identify gap sites through CMOS and UPRN / VOA database comparisons
- Work with wholesalers to identify mixed use, sub-metered and any special arrangement premises to create an offline database

- Identify eligible market gap sites and work collaboratively between wholesalers and retailers to register them into the market.

Phase 4: Occupancy Status, Customer and Segment

As we move towards a more accurate view of market data, focus will move to retailer obligations on occupancy status, customer name, and segment.

This phase will improve customer name accuracy, completeness, and visibility and subsequently accurate customer segmentation data. Some trading parties have an appetite to broaden the customer segmentation data to include the most recent classification data to allow (in conjunction with accurate and timely consumption data) more insight to be derived for water demand forecasting, leakage identification, communications, and sensitive customer identification. Therefore, Phase 4 will look to:

- Review supply point occupancy status accuracy, confirming where occupied or vacant premises are accurate and those deemed to be incorrect will be passed to the retailer
- Review customer name and customer banner name accuracy and completeness. Confirm accuracy and where the customer name is deemed incorrect pass back to the retailer to resolve
- Work with the market to ascertain the use and requirement for SIC codes and segmentation
- Offer a potential service that reviews and/or inputs missing SIC code classification. This can be done in tandem with approved Market Improvement Fund project, Project Discovery which is reviewing a standardised industry classification scheme.

Proposed process review and enduring service

In parallel to a large-scale data cleanse to address legacy data quality, it is equally important to establish an enduring solution that maintains the integrity of market data. This will include an ongoing review of associated market processes and incentives, including:

Eligibility: There are currently differing approaches to certain premises and customer types. Anecdotal evidence from CCW and specific housing associations highlight that different wholesalers approach the same type of premises differently, either managing them within the household or non-household market, i.e., care homes, communal halls, student halls etc. There are similar issues that are driving a lack of engagement elsewhere (see Gap Sites). If we have an agreed approach across all market participants, we can ensure that the right customers are in the market, the correct market size is maintained and that customers are receiving a consistent service.

New Connections (including change of use and temporary building supplies): If legacy data issues are resolved, this process should be the gatekeeper for CMOS data quality. We must determine and adopt best practice to encourage accurate and complete data entered into the market in a consistent format. Visibility must be maintained on developer sites, ensuring that any household premises are deregistered in a timely fashion. With adequate UPRN completion rates, the change of use process should be more efficient.

Gap Sites: Conversations with retailers have confirmed that they find it difficult to fully engage with gap site incentive schemes due to complexity and inconsistency across regions. Without adequate levels of UPRN completion and accuracy, the foundations for properly addressing gap sites are not in place. Visibility is required of all non-household premise types: those that will be subject to review, those that will not be accepted (mixed use or sub-metered customers) or have previously been rejected. This will afford market participants greater clarity, remove duplication and enable true gap sites to be the focus.

Address: As highlighted in the POC findings, having a single format for address data is challenging when the optimal data differs depending on the premises type. This could be mitigated through a system change or good practice guidance.

Industry segmentation (SIC Codes): There is a growing pull from market participants to better understand the make-up of the NHH market to: drive effective communications during emergencies or unplanned events; assist with leakage detection and water demand forecasting; and to support tendering activity or water efficiency intervention with customers. SIC code is not mandatory in CMOS and most SIC codes are derived from the 1980 classification (used to determine whether water is vatable). The market should agree an appropriate mechanism for capturing and maintaining meaningful segmentation in CMOS.

Costs

In calculating the cost of the initial data cleanse (in particular the cost for the 2023/24 financial year), we considered two options in terms of scope and phasing. The first with a focus on eligibility, premises, and address and the second that also focuses on customer, segmentation, and occupancy data.

Detailed costing will be included in the final business plan, but indicative costs are outlined below.

Option 1: Eligibility, Premises and Address Cleanse – **Provisional cost of £450-650k**

Market eligibility

- Identifying potential residential premises for further investigation by trading parties
- Identifying inactive premises where they have been removed or demolished.

Premises accuracy

- Incorrect UPRNs and VOA Billing Authority (BA) references will be identified, the correct reference will be supplied, where available
- If there is no UPRN and/or BA reference, the address will be used to identify the correct UPRN or VOA BA, where available.

Address accuracy

- SPID addresses will be matched against multiple reference data sources to determine address accuracy
- Addresses matched to reference data sources with a high degree of confidence will be recommended for update
- Inaccurate addresses will have incorrect address elements highlighted for investigation by trading parties
- Non-delivery addresses and non-addressable objects (e.g., taps and troughs) will be highlighted and categorised for review.

Option 2: Eligibility, Premises, Address Cleanse and Customer – **Provisional cost of £750k to £1M**

The scope includes all activities specified in Option 1, plus the following:

Occupancy accuracy

- Deep dive analysis conducted to determine an accurate view of occupancy by utilising and combining evidence from multiple datasets
- SPIDs appearing as occupied but have evidence to suggest that they are vacant will be identified and provided for further investigation by trading parties
- Vacant SPIDs that have evidence of occupancy will be identified and processed through Sagacity's Data Quality for Business software to determine the correct occupier.

Customer accuracy

- SPIDs with an existing occupier will be validated against multiple data sets to determine accuracy
- Occupiers identified as correct will be confirmed to MOSL
- Occupiers identified as incorrect will be identified and highlighted to MOSL with an alternative occupier provided where available

- From the occupiers confirmed, SIC codes will be validated, corrected, appended, and provided to MOSL.

Proposed funding model

The Data Quality Assessment has identified widespread data quality issues, with limited variation in performance across wholesale regions. We therefore believe a simple, wholesaler-led funding model is appropriate for Phases 1 and 2 (with address and premises data being the primary focus), with consideration for more sophisticated "polluter pays" model when we transition to an enduring service and introduce Phases 3 and 4.

The proposal for Phases 1 and 2 is to fund the service through additional Market Operator (MO) charges levied solely on wholesalers. This provides a simple and existing mechanism that allows wholesalers to fund the improvement of address and premises data that they own.

If supported, this principle would be accompanied by a code change to allow for the levy of costs on wholesalers only.

Conclusion

The data assessment findings demonstrate that there are significant issues in the quality of data in the non-household market. The scale of the issues highlights the need for a new approach.

Current Additional Performance Indicators (APIs) are not adequate to incentivise individual trading parties to address issues surrounding data accuracy and maintenance. We believe that a central service that can identify data inaccuracy and incompleteness would be more efficient, enable CMOS data quality to be accelerated and deliver improved outcomes to customers and the environment sooner.

Consultation Questions

Please delete options as appropriate and provide reasoning for your response where applicable.

1. **Do you agree that a centralised data improvement service (focused on address and customer data) would benefit customers and market participants? Please provide reasoning for your response:**

Yes

No

Reasoning:

2. **Have we captured all material benefits? If not, what else should be considered?**

Yes

No

Reasoning:

Additional benefits to be considered:

3. **Do you use any third-party services/products to improve or maintain your customer or address data quality? If so, what is the approximate annual spend related to the non-household market?**

Yes

Annual spend details:

No

4. **Do you agree with the proposed scope and phasing? Please provide reasoning for your response. If not, please propose an alternative:**

Yes

No

Reasoning / alternative:

5. Do you agree with the proposed funding option? Please provide reasoning for your response. If not, please propose an alternative:

Yes

No

Reasoning / alternative:

6. Do you have any additional comments you would like to make?

Appendix one: Further examples outlining the cost of maintaining poor data quality and its impact on the effective operations of the market

New Connections RFI (MOSL), April 2021

Accurate and reliable data is fundamental in establishing new connections in the market. As part of MOSL's commitment to 'streamlining customer and asset management processes' as part of the MPOP 2020/21, we issued an RFI to understand whether the new connections process was working for trading parties. We found that:

- 75 per cent of respondents said that the new connections process does not work effectively
- Over 70 per cent of respondents are incurring additional costs through inefficiencies in the new connections process
- Retailers said that most supply points registered do not include a UPRN or a VOA reference
- Retailers stated that between 30 and 70 per cent of supply points end up vacant due to inadequate data, with performance varying significantly between wholesaler regions
- Some wholesalers raised concerns that sufficient data was being provided to retailers, but supply points were still ending up vacant.

The Scottish Market Assessors Data Project, 2013

The Scottish non-household water market conducted a central data cleanse to establish an accuracy baseline for the Competition and Markets Authority (CMA) database and improve market data accuracy and completeness.

To create the baseline, supply points were aligned to the Scottish equivalent of the VOA, the Scottish Assessors Association (SAA). The project took place over 18-months and was able to identify non-eligible and gap sites, verify supply point rateable tariffs, and validate and correct premises/addresses.

The improvements and benefits generated from the project have now transitioned into enduring data monitoring and maintenance processes, which ensures that they have confidence in data, a consistent approach to maintenance and fit for purpose new supply point registration process.

Learnings from this programme of work have informed our proposed solution, phasing, and case for change. Scottish Water presented details and outcome of the project at the February 2022 User Forum.

Current Market Data Improvement Initiatives

The Core Market Data Improvement Plan, published in October 2020, set out activities to improve 'core market data' and fulfil associated market obligations, including market entry; metering; asset maintenance and customer switching.

The core data items identified for data cleanse activity were categorised as:

- Customer Details and Premises Data
- Meter Location Data
- Meter Details Data.

The plan included the creation of Additional Performance Indicators (APIs) for these data items to provide a measurable incentive for data owners to improve data quality.

In March 2021 the Market Performance Committee (MPC) approved three APIs, in the form of peer comparison tables, to incentivise trading parties to improve following premises and meter location data items.

Since these APIs were established, we have seen significant engagement with the wholesaler APIs, leading to the improvement of the completion levels of remises data, and the reduction of GIS issues, as seen in Table 1 below:

Table 1: API performance comparison

Additional Performance Indicator (API)	March 2021	November 2022
UPRN Completeness	50%	79.4%
VOA Completeness	42%	70.1%
Geographic Information System (GIS) data Issues	17%	6.1%

However, these improvements do not guarantee subsequent improvements to CMOS data quality. For example, whilst the UPRN completion rate has increased by 29 per cent, this does not mean that these UPRNs are accurate and guarantee they are maintained once entered into CMOS.

Market Performance Framework (MPF) Reform

The MPF Reform Programme consultation, published in October 2022, sets out the work undertaken to date on a 'root and branch' reform of the MPF. The programme is split into two phases, with phase one focusing on:

- Identifying the market activities that support delivery of desired market outcomes
- The risks and issues that could prevent the market delivering on its desired outcomes.

Market Activities

Driven by the Strategic Panel's priority market outcomes, the MPF Reform has defined the activities that are necessary to achieve these outcomes through collaboration with trading parties, PwC, Ofwat and MOSL. 11 (31 per cent) out of the 35 activities are directly underpinned by the need for good data quality and many more have an indirect reliance on this. Key examples are:

- CV1: Retailers to make sure that CMOS is regularly updated with good quality customer and consumption data
- CV3: Wholesaler to make sure that CMOS is regularly updated with good quality accurate asset and premises data including assessing eligibility for the market and timely deregistration of premises not eligible where appropriate

Market Risks and Issues

The Market Risks and Issues Tracker was created to identify future risks and current issues which prevent the market from delivering on desired outcomes. 14 (38 per cent) of the 36 risks and issues identified could be fully or partially mitigated by good data quality. Key examples are:

- CSE024: Wholesalers are not maintaining accurate or complete premises address data
- CSE025: Customers and/or properties are incorrectly included in the non-household market.
- Sagacity Data Assessment findings
- Activities for use in the MPF Reform programme.

Appendix two: Key central data cleanse items

Data Item No	Data Item name	Description and importance
D2039	Unique Property Reference Number (UPRN)	A unique alphanumeric reference number assigned to most addresses in Great Britain, enabling wholesalers to keep track of premises in their catchment throughout their lifecycle. This ensures premises and address data is kept up to date. Complete and accurate UPRN data is a key component in the identification of gap sites. This data enables retailers to cross match to other databases to track occupancy and customers.
D2040	UPRN Reason Code	Enables non-addressable supply point types to be flagged and tracked. These include bin stores, troughs, and public conveniences.
D2037	Valuation Office Agency (VOA) Billing Authority Reference	A unique alphanumeric reference number assigned to most addresses in England, that enables wholesalers to verify eligibility, separate businesses in multi-tenanted buildings and to segment customers. It enables retailers to track occupancy.
D2038	VOA Billing Authority Reference Reason Code	Enables non-addressable supply point types to be flagged and tracked. These include bin stores, troughs, and public conveniences.
D5004	Address Line 1	Premises address and postcode details, allowing wholesalers to determine where supplies and assets are maintained. Retailers use this data to determine occupancy. This, alongside the UPRN and VOA data, is key to enable the market to function. Without this data being maintained, eligibility, occupancy and customers cannot be validated.
D5005	Address Line 2	
D5006	Address Line 3	
D5007	Address Line 4	
D5008	Address Line 5	
D5009	Postcode	
D2015	Occupancy Status	Designates whether the premises is occupied and is maintained by the retailer. It is reliant on other data items (UPRN, VOA and address) to be accurate to enable verification.

D2027	Customer Name	Maintained by the retailer and designates the legal entity name of the customer. This, alongside banner name (below), allows for customer visibility within CMOS. This enables retailers to locate customers to tender their services to. This enables wholesalers to keep track of customers within their region in case of emergency, unplanned events, leakage detection and water demand management. It also aids the interim supply allocation process by ensuring visibility of customers particularly multi-site customers.
D2050	Customer Banner Name	Maintained by the retailer and designates the trading name of the customer, if different from the legal entity name.
D2008	Standard Industrial Classification (SIC) Code	A numerical code assigned to a limited business to designate the type of business or industry sector (see below). This is a non-mandatory data item provided by retailers.
D2092	SIC Code Type	Designates the sector or business type the customer operates in. This aligned to customer and consumption data is a key field in providing trading parties with insight into usage, assists wholesalers in demand forecasting and leak detection. Customer type insight is also beneficial in unplanned event planning, such as identifying businesses that were forced to cease operations during COVID lockdowns.