Request for Information: Core Market Data Cleanse

Version 1.00 | 19 October 2020

Contacts

Luke Austin
luke.austin@mosl.co.uk

Matthew Labrum
matt.labrum@mosl.co.uk

Milo Halford
milo.halford@mosl.co.uk

Hendriico Merila
hendriico.merila@mosl.co.uk

Table of Contents

1. Introduction 2

1.1 Purpose 2

1.2 Background 3

1.3 Summary 3

1.4 RFI questions 6

1.5 Next steps 7

2. Development of Data Cleanse Plan 8

2.1 Scope 8

2.2 Approach 8

2.3 Subject matter expertise 9

2.4 Core data items 10

2.5 Measuring data quality 11

3. Analysis of Core Data Items 14

3.1 Overview 14

3.2 Customer details and premises data 16

3.3 Meter location data 21

3.4 Meter details data 24

4. Conclusion 29

4.1 Recommendation 29

4.2 Next steps 29

Appendix I 31

A1. Data quality dashboards 31

1. Introduction

1.1 Purpose

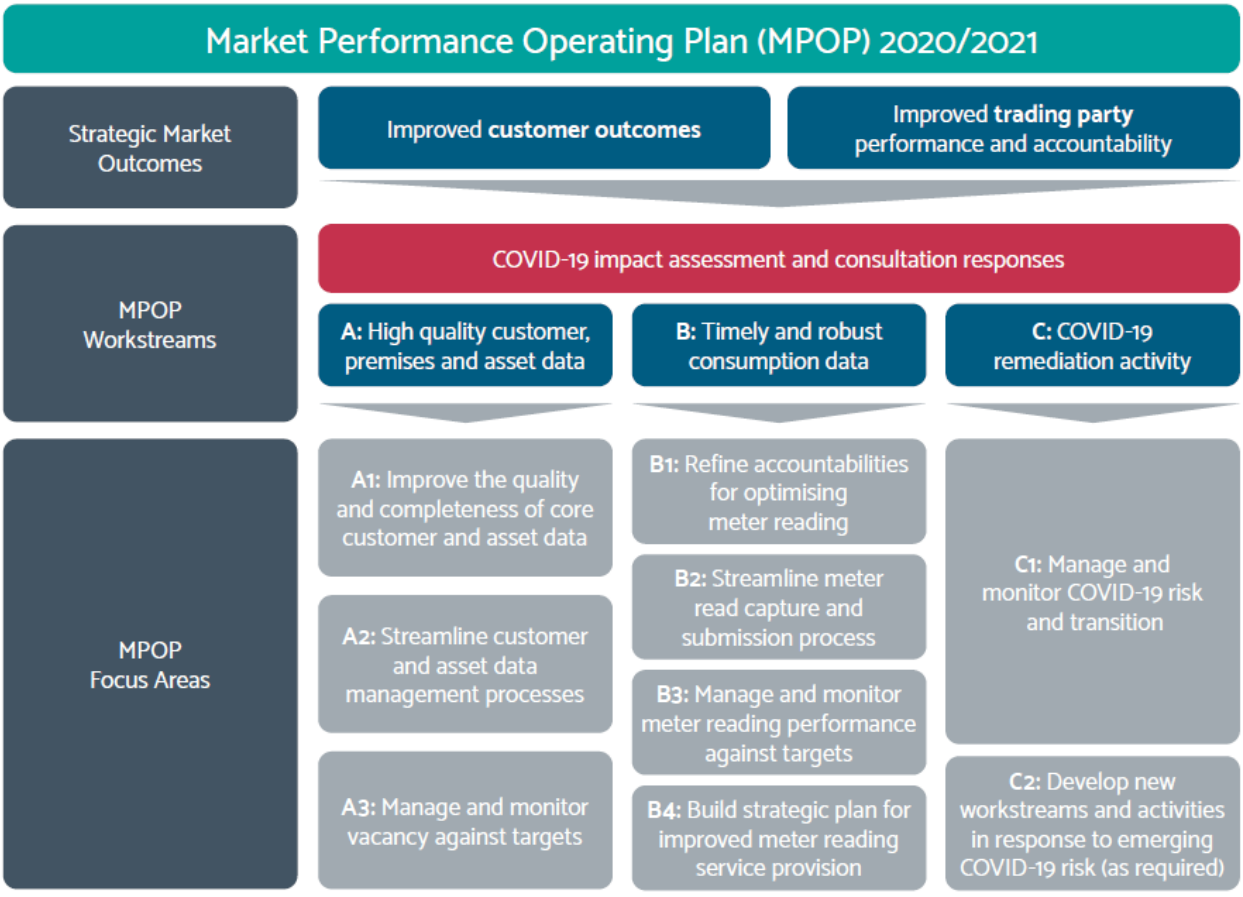
MOSL committed to delivering a data cleanse plan for core market data items as part of our Market Performance Operating Plan (MPOP) for 2020/21 (see Figure 1, Workstream A: 'High quality customer, premises and asset data'). This Request for Information (RFI) will inform the planning and prioritisation of our data cleanse activity.

We are looking for feedback on: (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 questions in the RFI cover three groups of data items:

- Customer details and premises data
- Meter location data
- Meter details data.

Figure 1: Overview of MPOP for 2020/2021



1.2 Background

Data quality issues cause significant operational challenges to the market, impacting customer outcomes and trading party performance.

Resolving data issues has been consistently highlighted in numerous industry consultations and market publications as essential for improving market performance, including:

- Ofwat's State of the Market 2019/20 report and RISE 2020 report
- PwC Market Audit Report for FY 2018/19 (excluding trading party specific items)
- CEO User Forum 2019 responses
- MOSL's Business Plan for 2020/21
- MOSL's 3-year Roadmap for evolving the Market Performance Framework (MPF)

The data issues most often cited since market opening include the following:

- **Legacy data:** Poor quality or missing data uploaded at market opening has led to a breakdown in operational processes creating additional manual workarounds and inefficiencies
- **Data ownership and incentives:** The current ownership and incentives structure of core data items impedes the correction of inaccurate or incomplete data items
- **Data consistency:** Inaccurate or missing meter location data impacts the ability of retailers to find meters and take reads. Mismatches of meter details between the market dataset and trading party systems also hinders the submission of meter reads into the market.

1.3 Summary

We have defined 'core market data' as the data that is needed to fulfil key market operations or functions, including: market entry; metering; asset maintenance and tendering and switching of customers.

Our approach to developing a data cleanse plan for core market data is based on defining, measuring and analysing data quality issues before consulting with the market to validate our recommended next steps.

We have used subject matter expertise across the industry to identify the main impacts of poor data quality, including inaccurate settlement and customer billing and increased cost-to-serve.

The core data items for each group, along with the impacts of poor data quality, have been summarised in Table 1.

Table 1: Impacts of poor data quality

Data Group	Data items	Impacts
Customer details and premises data	• Customer Name • Banner Name • UPRN • VOA BA Reference • Postcode	• Missing property reference data hinders the verification of premises address data, occupancy status, gap site eligibility and supply point eligibility. • Unreliable customer data obstructs retailers from reliably locating customer supply points within CMOS, creating costs and barriers for tendering and switching customers, as well as reputational damage.
Meter location data	• GISX / GISY • Free Descriptor • Location Code	• Difficulty finding and reading meters leads to increased costs for trading parties and could damage reputation. • Missing reads reduces the accuracy of settlement, leakage calculations and customer billing.
Meter details data	• Manufacturer • Serial Number • Meter Size • Number of Digits	• Increased meter read rejections reduces the accuracy of settlement, leakage calculations and customer billing. • Inaccurate meter details can lead to increased unplanned settlement runs and query requests, increasing trading party costs. • Incorrect meter size details can cause the wrong read frequency to be applied, leading to wasted effort and increased costs.

Our key findings and recommended next steps for each group of core data items are summarised as follows:

- **Customer and premises data**: Premises data include the Unique Property Reference Number (UPRN) and Valuation Office Agency Business Billing Authority reference (VOA reference). They are essential for the customer journey as they impact the tendering, switching and on-boarding of customers (particularly multi-site and national customers). They are also used for correctly determining occupancy status, gap site eligibility, supply point eligibility, linking together different datasets, identifying duplicate or missing supply points and linking water or sewerage supply to a particular building. The completeness of premises identifiers in the Central Market Operating System (CMOS) is low, with only 49 per cent of premises having at least one premises identifier. The codes require wholesalers to provide premises identifiers at supply point registration (CSD 0101 'Registration New Supply Points', section 2.2) and to ensure that it is corrected or updated going forward (CSD 0104 'Maintain SPID Data', section 5.1). We are therefore proposing to enforce these code obligations to support core retailer activities by using an additional performance indicator (API) based on the completeness of UPRN and VOA reference data. We will also use this data to link into external datasets (such as the VOA rating list or paid-for services) to validate and supplement market customer details data.

- **Meter location data:** Inaccurate meter location data hinders the taking of meter reads, which impacts the accuracy of settlement and customer billing. We have found that the inability of end-users, such as Meter Reading Service Providers (MRSPs), to either verify meter data, or to provide their own more accurate data, is the main cause of issues for metering data. This has led to end-users and retailers potentially having better data on their own systems and not having a mechanism or an incentive to provide it to the market. In the short-term we are proposing to coordinate a data sharing exercise with trading parties and MRSPs to cleanse current meter data. In the longer-term, MOSL is considering the development of new technology that will allow end-users to verify the data and to directly supplement the market dataset by providing more accurate information (for example, more accurate GIS coordinates). In addition, we will work with wholesalers to resolve known issues with GIS coordinates covering 202,540 meters, including those that are missing or duplicated.

- **Meter details data:** Inaccurate or inconsistent meter detail data hinders the submission of meter reads, which impacts the accuracy of settlement and customer billing and can also lead to additional costs to trading parties through unplanned settlement runs or query resolution. Meter detail mismatches account for 23 per cent of meter read rejections, with meter read manufacturer being the main cause. More than 25 per cent of meters have been assigned either an unrecognised meter manufacturer or they contain errors, misspellings or unnecessary information. Based on the findings of this RFI, we propose to investigate the removal of the validation requirement to provide a meter manufacturer when submitting a meter read into CMOS and to work with wholesalers to ensure that all manufacturers on CMOS correspond with a pre-defined list of manufacturers maintained by MOSL.

In addition, we will investigate next steps in the following areas:

- Standardise the interpretation of the codes with respect to customer name and customer banner name data

- Treatment of domestic level consumption non-standard or non-addressable premises (such as troughs, places of worship, public conveniences, stand-pipes, etc.)

- Standardise the customer name details for vacant premises, e.g. set customer name to ‘NULL’

- Establish best practice guidance for meter location descriptions.

1.4 RFI questions

The full Data Cleanse RFI survey contains additional information taken from this document to help you to answer the questions. Note that all answers containing financial information will be treated as confidential but could be aggregated with other responses to inform priorities and next steps.

The following is a summarised version of the questions:

1. **Are there any other impacts to your organisation of poor data quality aside from those already identified? For example, spillovers from poor data quality that impact profitability or hinder the development of your business.**

2. **What activities or types of activities do you undertake to mitigate the impacts of poor data quality? For example, the creation of a dedicated team to locate meters or the use of track-and-trace services. Please also provide an estimate of the resource or financial cost to your organisation of these activities measured either in terms of number of Full Time Employees (FTEs), including new or re-allocated FTEs, or in terms of expenditure.**

3. **What are the benefits to your organisation of good data quality? For example, customer satisfaction, market insight, improved efficiency or profitability, etc.**

4. **Do you use any paid for data services to manage or enrich data (e.g. Experian or Dun and Bradstreet)? Please provide further details, including what you use the services for (or why you do not use them), what limitations they may have and whether they have any disadvantages or advantages compared to free-to-use external datasets.**

5. **How do you ensure that good data quality is maintained for the CMOS data items for which your organisation is the owner?**

6. **Are there any differences between the core data on your own systems compared to CMOS? If so, please provide details of which data items, why they are different and whether you would be willing to share this data with the market.**

7. **Do you agree with MOSL implementing new technology to supplement the market dataset with end-user input, such as Meter Reading Service Providers? For example, this could be used to verify the accuracy of core market data or to provide more accurate data, such as meter location and meter details data. Please provide reasons for your answer.**

8. **What is your understanding of the proper use of the data items 'Customer Name' (D2027) and 'Customer Banner Name' (D2050) as defined in CSD 0104 ('Maintain SPID Data', section 4.1.1)? For example, would you enter individual, company or contact details in either of the fields?**

9. **Do you agree with removing the CMOS validation requirement to provide the meter manufacturer when submitting a meter read? Please provide further details for your answer.**

10. **Do you have any further comments regarding the data quality of core data items or the contents of the RFI document?**

1.5 Next steps

We will analyse the results of the RFI and incorporate the findings into our planning for data cleanse activity. We will consult with the market based on a draft plan before publishing the final version of the data cleanse plan to the market by the end of February 2021. Note that we may also undertake certain priority activities to improve the quality of market data before the final plan is published.

Our planning will balance the prioritisation of improvement activity based on the expected impact, cost and the amount of time required to implement. For example, developing an API and using performance rectification tools is relatively straightforward compared to developing a chargeable standard or implementing new technology—however, the latter may provide more impactful and enduring solutions.

The timelines for next steps have been summarised in Table 2.

Table 2: Timeline for publication of data cleanse plan

Milestone	Status	Target Completion
Agree high-level approach for data cleanse with Market Performance Committee (MPC)	Complete	01-Jul-20
Define scope, issues and objectives for data cleanse	Complete	29-Jul-20
Develop data quality metrics or indicators and highlight gaps	Complete	19-Aug-20
Analysis of core data items, identification of possible solutions and input from subject matter experts	Complete	09-Sep-20
Prepare RFI document for consultation and launch RFI	Complete	19-Oct-20
Complete RFI (three weeks)	On Track	06-Nov-20
Analysis of RFI responses and incorporate findings into data cleanse planning	On Track	27-Nov-20
Finalise draft data cleanse plan	On Track	30-Nov-20
Consultation on draft plan	On Track	30-Jan-21
Finalise data cleanse plan and publish to the market	On Track	26-Feb-21

2. Development of Data Cleanse Plan

2.1 Scope

The definition of 'core market data' is the data which is required to fulfil key market obligations or functions, including: market entry, metering, asset maintenance and customer switching.

The initial scope of the data cleanse will focus on improving the data quality of data items identified as 'core'. At a later phase we will review non-core data and whether there are data items that are underutilised or redundant.

The expected outcomes of this data cleanse plan are the following:

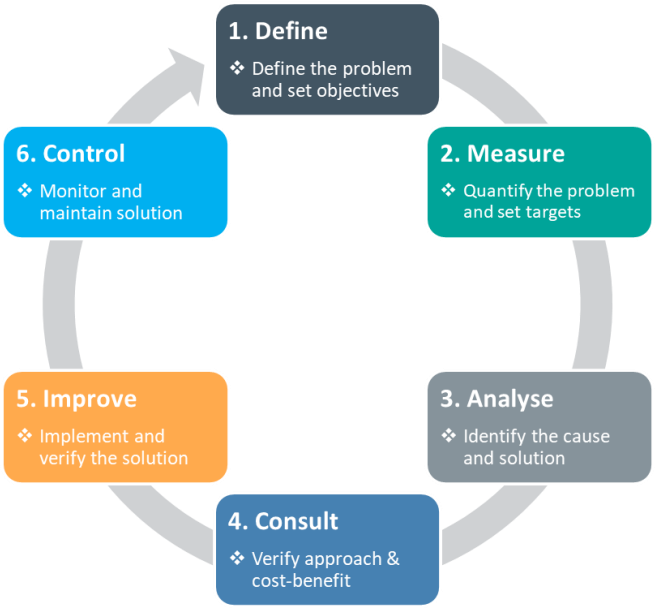
I. Retailers should be able to efficiently find, on-board, tender or switch a customer or customer supply point(s) using CMOS data

II. Trading parties should be able to efficiently find a meter, take a meter read, and submit a meter read into CMOS.

2.2 Approach

The development of the market data cleanse plan is based on six stages (summarised in Figure 2). This process is intended to ensure that the issues within each area of data cleanse have been clearly defined, measured (where possible) and evidenced. This is based on the best available information using desktop analysis of central market data (i.e. CMOS), input from subject-matter experts from various trading parties and a Request for Information (RFI) with the industry.

Figure 2: Overview of development the MOSL data cleanse plan



We expect data cleanse activity to be an ongoing, iterative 'lifecycle' requiring continuous efforts to monitor and improve the quality of data across the market. Certain 'priority change' activities will be initiated in the near-term; however, some longer-term, durable solutions may take several years to have an impact.

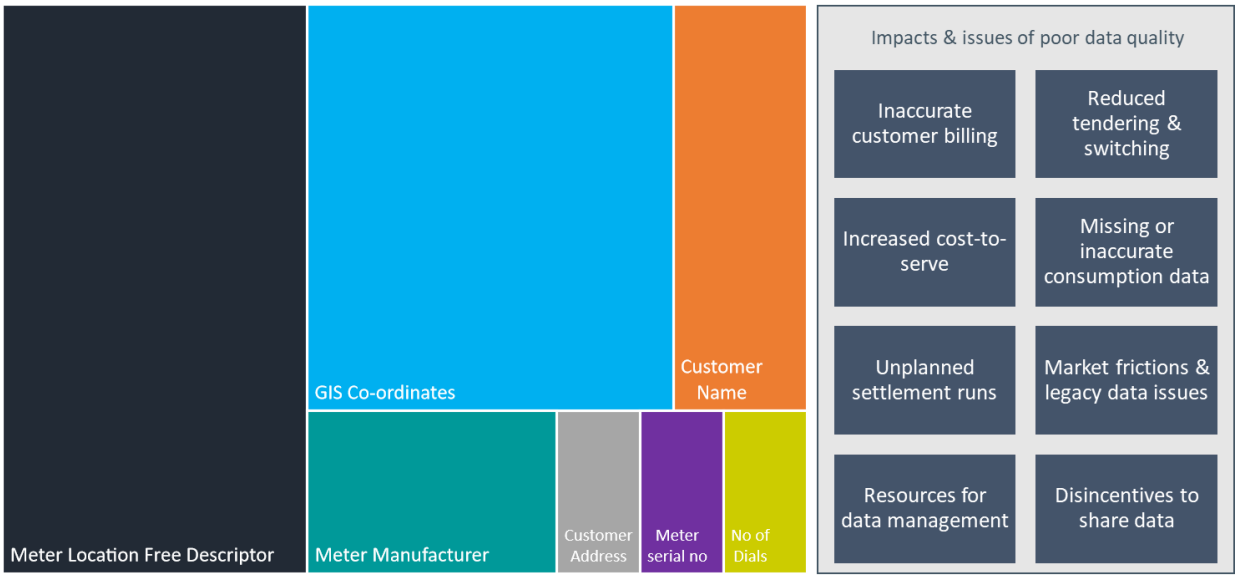
To achieve its outcomes the data cleanse plan will apply the following principles:

- **Proportional solutions**: The scale of data cleanse activity, along with its timing and prioritisation, will be proportionate to the impact and cost to the market. In addition, there will be no 'one-size-fits-all' approach: there may be some underlying or fundamental issues impacting all data items, but different solutions will be used according to the nature of the issues
- **Measurable improvement**: All data cleanse activity will be underpinned by a robust measurement of data quality to ensure the impact of cleanse activity is measurable and can be evaluated
- **Appropriate accountability**: We will ensure there are clearly assigned responsibilities and accountabilities based on data ownership and code obligations. Consideration will also be given to whether the current accountability or data ownership model is appropriate.

2.3 Subject matter expertise

We have interviewed nine subject matter experts (SMEs) from four wholesalers and five retailers to understand the issues they face due to the quality of core market data. The main problem areas raised, alongside the associated issues, have been summarised in Figure 3. The size of the boxes shown are proportional to the number of issues raised by an SME.

Figure 3: Key problem areas associated with core market data highlighted by industry subject matter experts



Meter location data was raised by both wholesaler and retailer SME representatives as the biggest data quality issue in the market. In particular, the usefulness of the meter location free descriptor and the inaccuracy of GIS coordinates were highlighted as priorities. Both sets of trading party SMEs highlighted the time and cost required to rectify issues and locate the meter and pointed to the impact of long unread meters. Wholesaler SMEs were also concerned by the high volumes of meter verification requests caused by inaccurate meter location details as this adds to administrative costs.

Retailer SMEs specifically flagged issues with the uniformity and accuracy of customer name and unique premises identifiers such as UPRN and VOA references. They said that this impacts the ability of the retailer to identify customers within CMOS (particularly multi-site or national chains) and hinders the tendering and switching of potential customers.

Wholesaler SMEs highlighted mismatches caused by meter detail inconsistencies (i.e. meter manufacturer, meter serial number and number of digits) and the impact on settlement of missing meter reads. They said that poor record keeping of meter dials is leading to rollover errors which in turn leads to unplanned settlement runs.

Additional comments from one wholesaler SME included the suggestion that many data quality issues are legacy issues and not the fault of the retailers. They suggested that wholesalers should have more responsibility for metering as this would resolve a number of issues. Another respondent argued that complacency is the issue rather than data itself. A retailer SME argued that large retailers have insufficient focus and resources to manage large volumes of market data. Lastly, a different retailer SME suggested that many retailers have better quality data on their own systems but are either unable or even disincentivised (due to the risk of customers switching away) to upload this data to CMOS.

2.4 Core data items

Based on our analysis and the input we have received from subject matter experts across the industry, we have identified the following core data items for data cleanse activity, summarised in Table 3.

Table 3: Data items included within the scope of the market data cleanse

Group	Data Item ID	Description	Data Owner
Customer Details and Premises Data	D2027: Customer Name	The customer name associated with a given supply point.	Retailer
	D2050: Customer Banner Name	The trading name of the customer at a given eligible premises.	
	D2039: UPRN	Unique Property Reference Number (UPRN) as published in the National Land and Property Gazetteer (NLPG).	Wholesaler
	D2037: VOA BA Reference	Valuation Office Agency (VOA) Billing Authority Reference Number.	

Group	Data Item ID	Description	Data Owner
	D5009: Postcode (Premises)	Postcode for premises (with a space between the outcode and incode).	Wholesaler
Meter Location Data	D3017: GISX D3018: GISY	Specifies the X / Y coordinate of the location of the meter. The accuracy of the coordinates must be reasonable to facilitate finding the meter and should be within a range covering England and Wales.	Wholesaler
	D3019: Meter Location Free Descriptor	Free descriptor of the location of the meter.	Retailer
	D3025: Meter Location Code	Indicates whether the meter is inside or outside of a building.	Wholesaler
Meter Details Data	D3013: Meter Manufacturer	Specifies the make and/or manufacturer of a meter.	Wholesaler
	D3014: Manufacturer Meter Serial Number	Specifies the manufacturer's serial number of a meter.	
	D3003: Physical Meter Size	Nominal size of the meter in mm.	
	D3004: Number of Digits	The number of digits required to provide a reading in m^3. This is irrespective of the actual number of dials or digits on the meter, as meters may record volumes to a higher or lower resolution than $1m^3$. This data item is required with reference to m^3 for the purposes of rollover detection. This will also be the number of digits required for the maximum volume in m^3 that can be recorded by the meter.	

2.5 Measuring data quality

The quality of data can be measured in several ways with differing levels of importance and ease of measurement. We have limited our data quality metrics to four (summarised in Table 4). Note that measuring data quality is essential to data cleanse activity and improvement activity cannot be reliably tracked or evaluated without a metric.

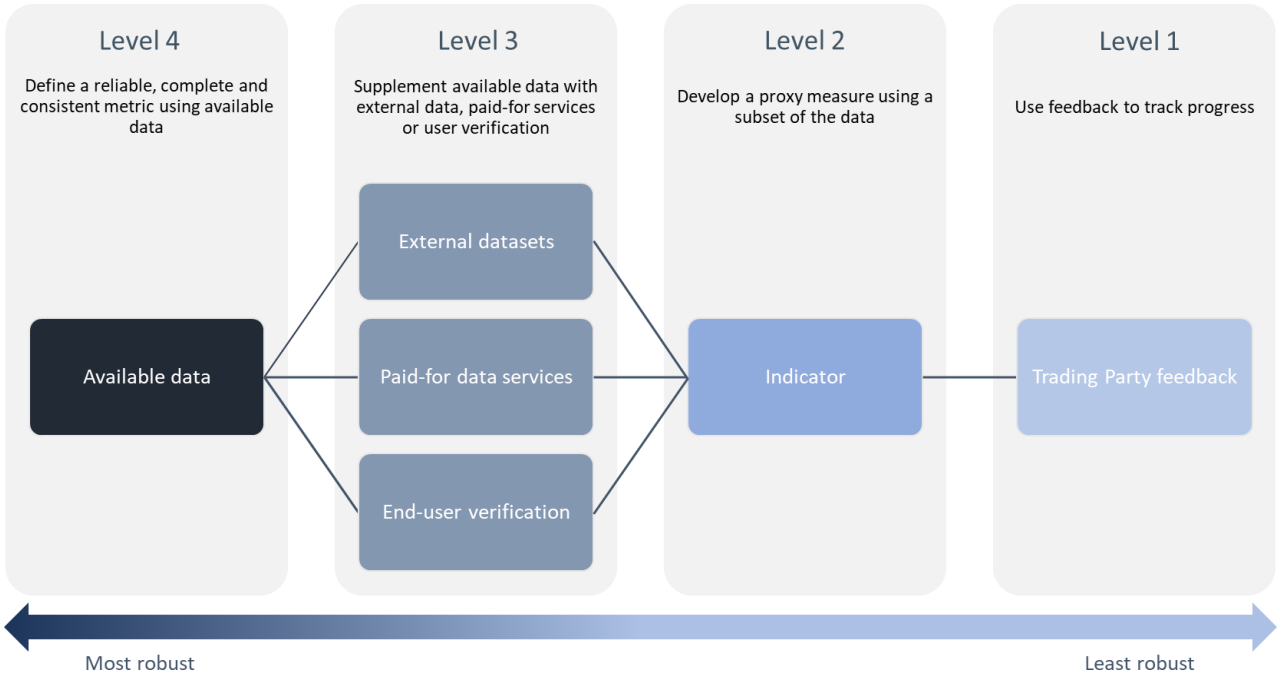
Table 4: Summary of main types of data quality metrics

Data Quality Measure	Definition	Measurement
Accuracy	Percentage of data fields that are correct compared to the expected or actual value.	No data items can be measured with currently available data and so this has been marked as undefined.
Completeness	Percentage of data fields that have non-NULL values.	All data items can be measured.
Validity	Percentage of data fields that meet the required format.	The majority of data items can be measured, where it does not apply it has been marked as N/A.
Timeliness	Percentage of data that has been maintained or validated within the last three years (i.e. since market opening).	All data items can be measured.

Where possible, we have defined a metric for each data item based on available data. In some cases, particularly for accuracy, it is not currently possible to define a metric. In some cases, we have been able to develop an indicator of *inaccuracy* based on analysis the data; however, this approach has significant limitations. In particular, there is no guarantee that resolving an issue associated with inaccuracy will result in a more accurate data item. For example, a GIS coordinate for a meter may be flagged as inaccurate because it suggests the meter is located in the sea or because it is improbably far from the post code centre; but without an accuracy metric there is no way of knowing that any new coordinates provided actually locate the meter.

Figure 4 describes the levels of data quality measurements based on what data is available. It ranges from the most robust measures that depend on available data, external data and user-verification to least robust that depend only on trading party feedback. We will endeavor to have metrics for data quality as a basis for measuring the outcome of data quality improvement activity. Trading party feedback is also important and will be used, although it could lack precision and require interpretation.

Figure 4: Levels of data quality measurement



3. Analysis of Core Data Items

3.1 Overview

Through our analysis we have identified the following key issues contributing to poor data quality in the market:

- **Data verification**: There is a lack of end-user verification for the accuracy of market data. This applies to both the data uploaded by wholesalers at market opening (i.e. legacy data) and to newly registered supply points. CMOS provides limited validation which only considers whether a data item is required (i.e. completeness) or whether it complies with a certain format (i.e. validity). The inaccuracy of data is an often-cited concern by trading parties; however, due to the lack of verification of CMOS data, it is not possible define a robust accuracy metric for *any* of the core market data items. Certain external datasets (e.g. UPRN or VOA reference data) or paid for data services (such as Experian or Dun & Bradstreet) could allow us to partially verify the accuracy of customer details; whereas a mechanism to allow end-user verification would allow us to measure the accuracy of meter location and meter details data.

- **Market codes and disincentives**: There are ambiguities in the market codes (particularly for customer name and meter details data) and there are currently no reputational or financial incentives for trading parties to maintain the quality of their market data. This has meant that cleansing data stored in CMOS has been a lower priority for many trading parties compared to activities directly impacting settlement; and, in many cases, the costs of data cleanse have typically been borne by end-users rather than data owners. This issue has also potentially led to many retailers retaining better quality data on their own systems than on CMOS to reduce the risk of customers switching away—and therefore losing the investment in cleansing the data. This would mean that retailers are disincentivised to submit better quality data into CMOS. In addition, even if retailers have better quality data (e.g. more accurate meter location data), and be willing to update CMOS, they are not the data owners for many core data items. They therefore have no direct route to share this data with the market.

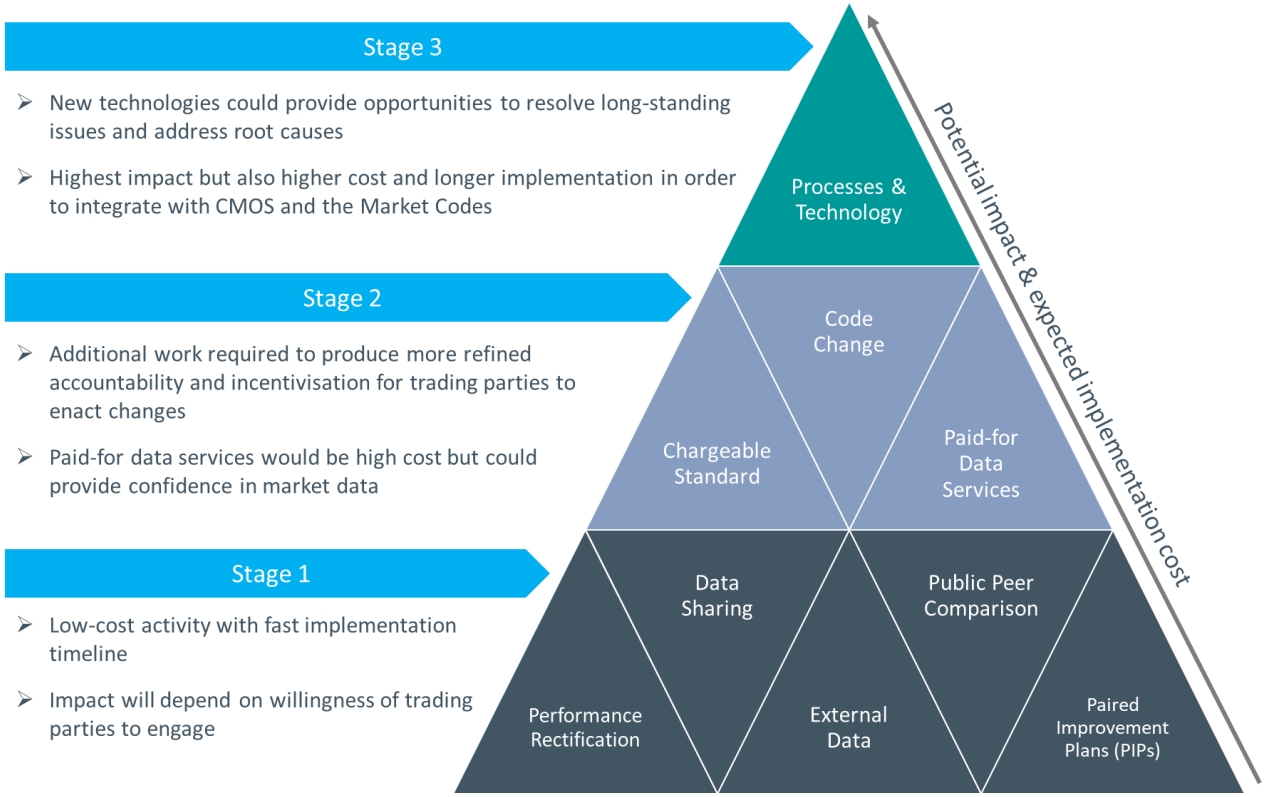
We have identified three different levels of intervention to improve data quality (summarised in Figure 5):

- **Stage one: Shorter-term initiatives** designed to have a more immediate impact, but which do not necessarily address underlying causes of poor data quality. This includes peer comparison, coordinated data sharing and performance rectification

- **Stage two: Market-based activity** that will address some of the obstacles to maintaining good quality data. This includes addressing inappropriate accountabilities and lack of incentivisation through code changes and financial or reputational incentives

- **Stage three: The highest level of intervention** focuses on addressing the root causes of poor data quality. This is done by ensuring the requisite processes and technology are in place to allow full verification of data and optimal data capture by those who are best placed to do so. This includes the need to develop a mechanism for end-user verification and data input. MOSL is currently working towards an Application Programming Interface that could allow end-users (such as retailers or meter reading service providers) to verify the accuracy of core market data and to provide more accurate data to supplement the market dataset.

Figure 5: Hierarchy of interventions to improve data quality



The rest of this section outlines the data quality metrics identified so far along with key observations and impacts of poor data quality, possible solutions and a discussion of the main issues. The analysis of data items is grouped into the following three areas:

- Customer details and premises data
- Meter location data
- Meter details data

3.2 Customer details and premises data

Observations

Premises data:

- The level of completeness for property references is low, with only 61 per cent of premises having an identifier (i.e. one of either a UPRN or VOA reference)
- Of those premises missing an identifier, 34 per cent are vacant
- In over 90 per cent of cases, wholesalers have provided non-specific reasons (i.e. 'Other') for not being able to provide a UPRN or VOA reference
- Only 10-14 per cent of UPRN and VOA references have been updated or corrected since market opening
- There is a risk that gap site incentive schemes cannot be properly administered without premises identifiers to avoid issuing duplicate incentives for gap sites already in the market
- 30,539 premises (2.34 per cent) have either an invalid or retired postcode.

Customer details:

- Accuracy of customer name data cannot be measured with currently available data but could be partially checked against VOA reference data
- 20 per cent of premises specify either a 'generic' (such as 'TREASURER' or 'BURSAR') or 'missing' (such as 'EMPTY, 'NO CUSTOMER', 'NO OCCUPIER') customer name reference. Of the 'missing', 90 per cent are vacant, and of these 43 per cent were vacant at market entry
- 35 per cent of customer name data has been changed since market opening
- The codes are unclear regarding usage of the customer name fields leading to inconsistent approaches by retailers. This includes retailers providing customer contact details in the customer name field or providing customer name in the customer address field
- We estimate that approximately 10 per cent of premises are disassociated from a valid customer name or address (such as troughs, public conveniences, stand-pipes, allotments, etc.) and there is no standardised approach in the market for dealing with them
- There is no mechanism for the data entered by the data-owner (the retailer) to be verified by the (potential) end-user (i.e. *other* retailers) until customers switch

- There is no incentive for retailers to maintain accurate customer details on CMOS and they could have more accurate data on their own systems.

Impact

- Missing property reference data hinders the verification of premises address data and supply point eligibility. This leads to increased costs to both retailers (e.g. due to utilisation of paid for services to verify data) and wholesalers (e.g. due to site verification requests). This also leads to increased vacancy at supply point registration
- The inability to locate customer supply points within CMOS creates costs and barriers for tendering and switching customers (particularly multi-site or national customers) thereby reducing market competition

Possible solutions

- Update CSD 0101 ('Registration New Supply Points'), CSD 0104 ('Maintain SPID data') and CSD0301 ('Data Catalogue') to strengthen the obligations for the entry and maintenance of the UPRN and VOA reference.
- Use the completeness metric as the basis for an API or chargeable standard to incentivise wholesalers to provide valid UPRN and VOA reference data either at supply point registration or when available
- Review the eligibility and guidance for non-business (i.e. customers with no UPRN or VOA reference) and other non-standard premises (i.e. troughs, places of worship, public conveniences, stand-pipes, etc.)
- Incentivise accurate customer name data by developing an accuracy metric based on external data (e.g. VOA reference data) or paid-for data services (e.g. Experian, Dun and Bradstreet, etc.) and implement an API or chargeable standard against it
- Update CSD 0104 ('Maintain SPID data') and CSD 0301 ('Data Catalogue') to clarify the definition of the customer name and banner name; and develop guidance to standardise the interpretation of the codes
- Standardise treatment of customer name for vacant premises, e.g. set value to 'NULL'. This could also include validation to ensure that occupied premises do not have this value
- Coordinated data sharing to validate CMOS data with data from wholesaler and retailer systems.

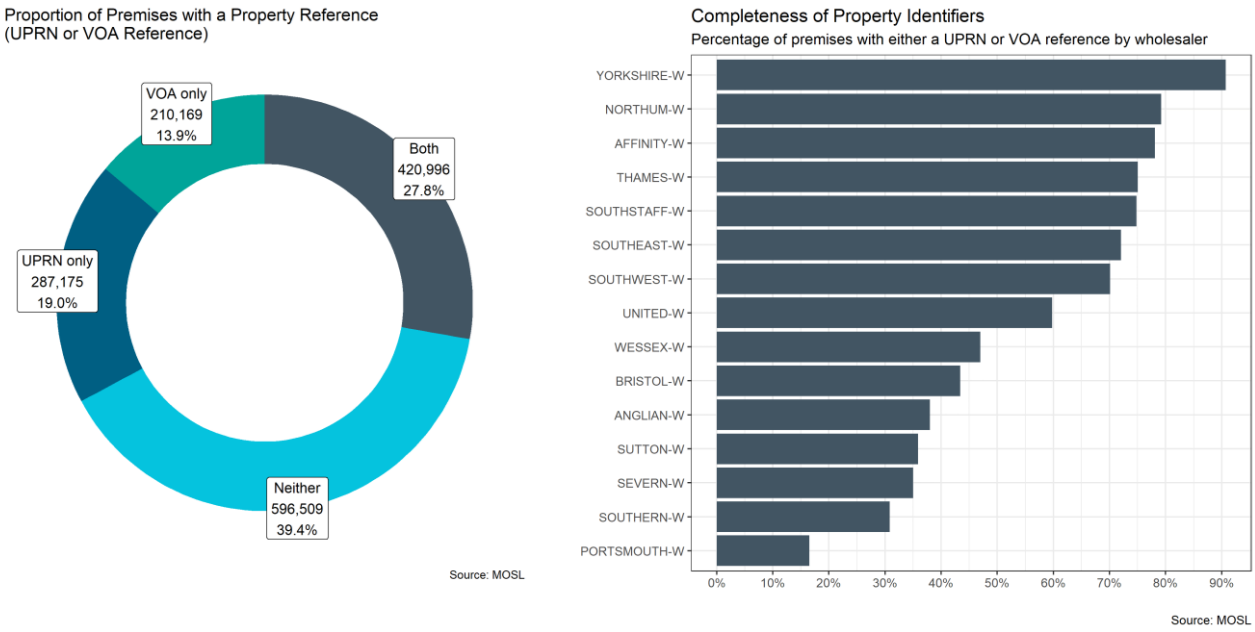
Discussion

CSD 0101 ('Registration New Supply Points', section 2.2) requires wholesalers to provide a UPRN or VOA reference at SPID registration, *or* to provide a valid reason for not providing the information. Additionally, CSD 0104 ('Maintain SPID Data', section 5.1) requires wholesalers to update premises details (including UPRN or VOA reference data) to correct or improve it.

The low completeness of premises details data (47 per cent for UPRN and 42 per cent for VOA reference) is compounded by wholesalers specifying, in most cases, 'Other' for not being able to provide the data at supply point creation (96.3 per cent of cases for a missing UPRN and 91.4 per cent for a missing VOA reference). This non-specificity suggests that either the reason codes are not fit to capture the challenges of obtaining a property identifier; or, that wholesalers have not made sufficient efforts to provide this data to the market at supply point registration or to seek to improve it at a later date.

Figure 6 highlights considerable variability amongst the completeness of property identifiers by wholesalers. The percentage of premises with at least one of either a UPRN or VOA reference ranges from less than 20 per cent (Portsmouth Water) to over 90 per cent (Yorkshire Water). There are seven wholesalers with less than 50 per cent completeness for property references.

Figure 6: Missing property references



The accuracy of customer name details cannot be measured with currently available market or public data. A lower bound for the accuracy of customer name details is indicated by matching customer names in CMOS with Companies House data: approximately 25 per cent of customer names in CMOS could be matched; however, the Companies House dataset is incomplete and does not include unregistered businesses.

A more reliable method for measuring the accuracy of customer name details is to use premises identifiers to check for matches with external datasets (e.g. VOA ratings list data) or paid-for data service (e.g. Experian or Dun and Bradstreet). There may be only partial coverage in some cases, and time lags between the external datasets and the market data; but we believe this would allow for customer details to be verified in most cases.

We therefore recommend prioritising improving the quality of premises details data as a precursor to improving customer details. A metric derived using external data or paid-for data service could then be used as the basis for an API for peer comparison and performance rectification activity.

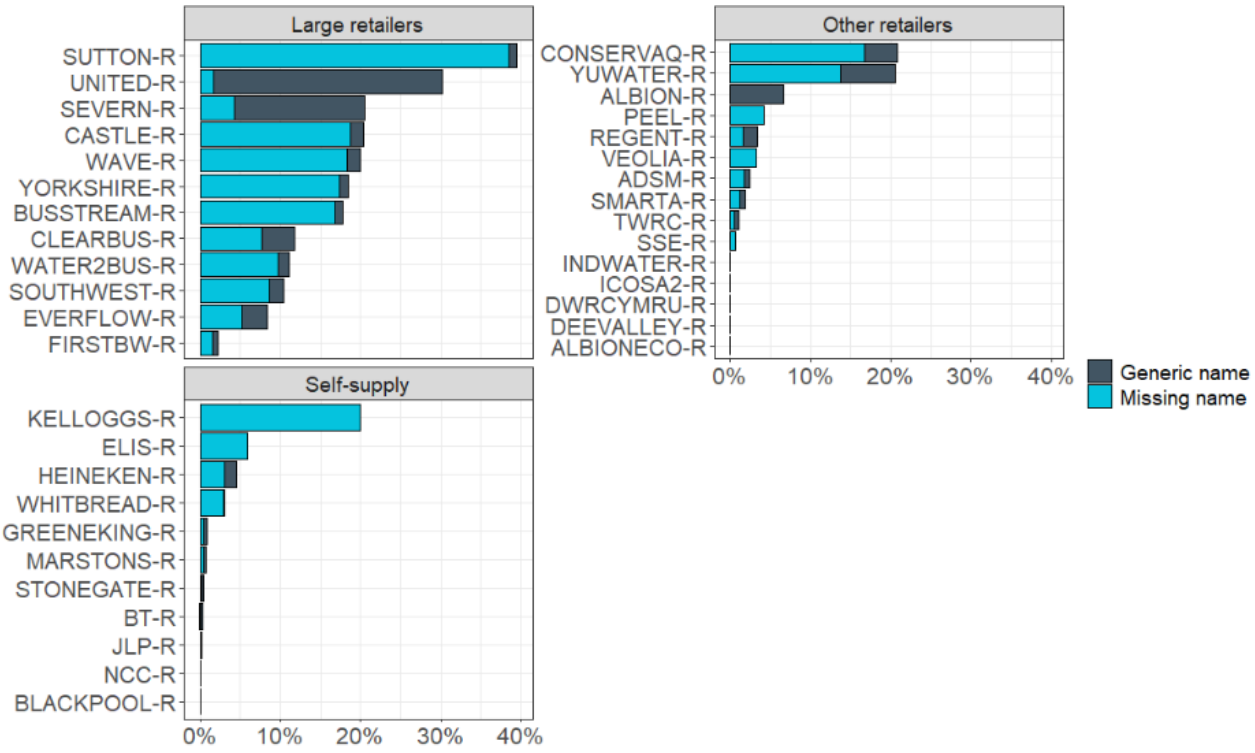
CSD 0104 ('Maintain SPID Data') specifies the following definitions for customer details data:

- **Customer Name (D2027)**: "The legal entity at the Eligible Premises contracting with the retailer for the provision of Water Services or Sewerage Service at the Supply Point."
- **Customer Banner Name (D2050):** "The trading name of the entity."

We believe that customer details data suffers from a lack of clarity in the codes. This leads to inconsistent approaches being taken by retailers that may hinder the identification of customers in CMOS. For example, in many cases, contact details have been provided rather than a customer name. In addition, we have identified 119,113 premises (8 per cent) with 'generic' customer names (e.g. 'BURSAR', 'TREASURER', 'HEADMASTER', etc.); which could restrict the ability to reliably identify the underlying customer entity in CMOS. We have also found that 184,969 premises (12 per cent) specify a 'missing' customer reference such as 'NO CUSTOMER', 'EMPTY', 'NO OCCUPIER', 'VACANT', etc. Of these, 154,179 (83 per cent) were vacant.

Figure 7 shows the level of generic and missing customer name details by retailer. Water Plus in particular (UNITED-R and SEVERN-R) have a high level of generic customer names; whilst there are several retailers with high levels of missing customer name details, suggesting that a common approach (e.g. assigning a 'NULL' value) to dealing with vacant premises would benefit the data quality of multiple retailers.

Figure 7: Retailers with non-specific or generic customer names



Source: MOSL

Recommended next steps

1. We will use the completeness metric for UPRN and VOA reference data as the basis for a wholesaler API to drive improvements to data quality. This could initially include peer comparison and performance rectification. We will seek to prioritise premises by consumption and occupancy. In the longer-term we will review the codes to strengthen the obligations for providing premises data and investigate the implementation of a chargeable standard against a completeness metric
2. We will investigate whether external or paid for data services could be used to verify the accuracy of customer details and therefore to develop an accuracy metric. If robust, the accuracy metric could then be used as the basis for an API to drive improvements through public peer comparison and, where appropriate, performance rectification. In parallel, we will investigate the need for a chargeable standard to maintain on-going data health
3. We will work with the RWG to develop guidance in the following areas: standardise the interpretation of the codes with respect to customer name details; treatment of domestic level consumption, non-business (i.e. customers with no UPRN or VOA reference) and other non-standard premises (i.e. troughs, public conveniences, stand-pipes, etc.); and standardise the customer name details for vacant premises, e.g. set customer name to 'NULL'
4. If necessary, we will seek to amend the codes to clarify the definition of customer name fields.

3.3 Meter location data

Observations

- 15.5 per cent of GIS coordinates have been flagged as inaccurate
- 17 per cent of GIS coordinates have at least one duplicate
- Trading parties and MRSPs could retain better meter location data than on CMOS and there is no mechanism for them to provide this data to the market
- 20 per cent of GIS coordinates have been updated since market opening
- There is no direct, financial incentive for wholesalers to maintain accurate GIS data, but it is in their interest that retailers submit meter reads and maintain settlement accuracy
- Accuracy cannot be defined for the meter location free descriptor, but it could be defined in terms of 'usefulness'.

Impact

- Inaccurate location data leads to additional time and costs for both retailers and wholesalers to find and read meters
- Poor quality GIS coordinates contribute to long unread meters and missing consumption data leading to inaccurate settlement and customer billing.

Possible solutions

- Wholesalers to review and correct the GIS coordinates that have been flagged as inaccurate
- MOSL to coordinate data sharing using data from wholesaler and retailer systems to improve the quality of CMOS meter location data
- Develop a mechanism for end-users to verify the accuracy of the GIS coordinates and usefulness of the location free descriptor
- Use third-party services to improve the identification and mapping of meter locations
- Develop industry standards and best practice guidance for the location free descriptor field, such as common abbreviations

Discussion

CSD 0301 ('Data Catalogue') states that the accuracy of GISY and GISX coordinates must be "reasonable to facilitate finding the meter". The accuracy of GIS coordinates cannot be verified using available data; and we are unaware of any paid for data service that could provide a reference to produce a reliable accuracy metric. This means that we currently have no method of verifying whether coordinates are sufficiently accurate to find the meter.

In the absence of an accuracy metric positively identifying accurate GIS coordinates, we have been able to identify a subset of GIS coordinates that we instead believe are inaccurate. In total, 202,540 meters were found to have at least one of the following issues:

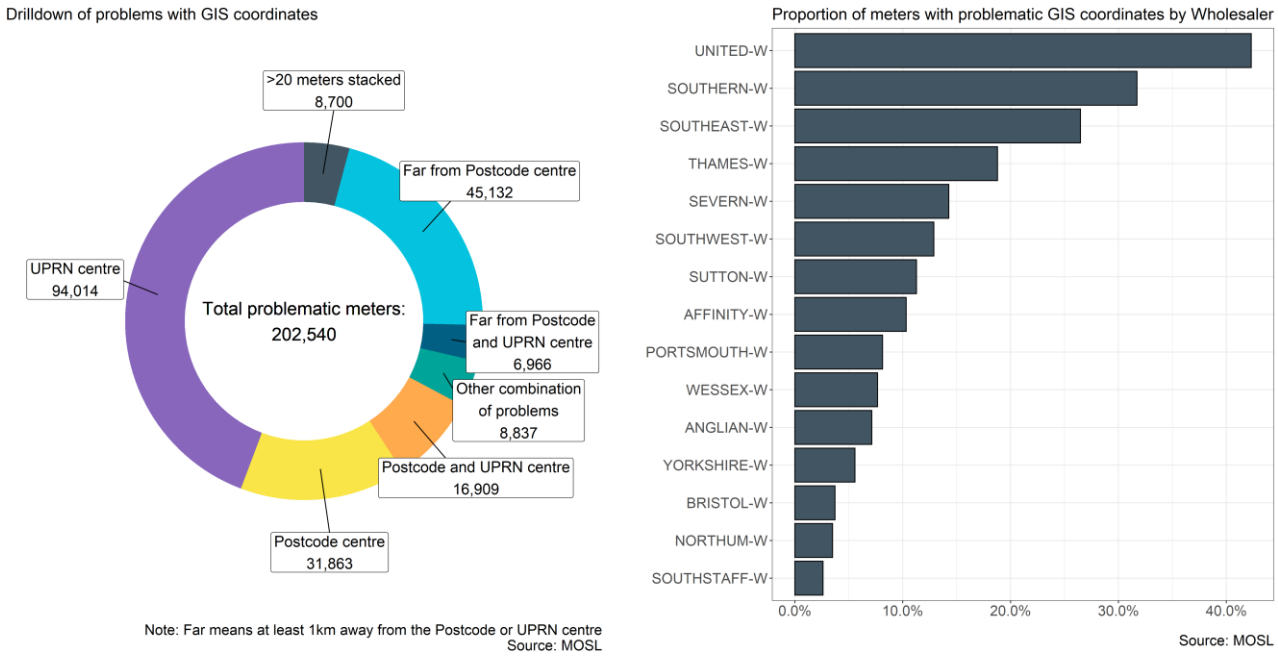
- Missing coordinates
- Coordinates assigned to the centre of a postcode
- Coordinates assigned to the centre of a UPRN, as given by the Ordnance Survey
- Coordinates suggest 20 or more meters are stacked within a 100cm2 area
- Coordinates suggest the meter is at least 1km away from the postcode centre
- Coordinates suggest the meter is at least 1km away from the UPRN centre.

The percentage of coordinates with at least one of these issues flagged can be used as an indicator of inaccurate GIS coordinates. Figure 8 shows the performance of wholesalers with respect to doubtful GIS coordinates. United Utilities has the highest proportion of problematic coordinates, with more than 40 per cent of GIS coordinates being flagged.

The indicator of problematic GIS coordinates could be used as the basis for an API, peer comparison and performance rectification activity to drive wholesalers to resolve these issues. We will share the list of flagged coordinates with each respective wholesaler.

It should be noted, however, that in the absence of a genuine accuracy metric we can only track whether the issue has been resolved, but not whether the updated GIS coordinate is accurate.

Figure 8: Wholesaler performance for quality of coordinates indicator



An alternative to directly measuring the accuracy of GIS coordinates would be to create a mechanism for end-users to verify the accuracy or usefulness of the data. For example, a MRSP could provide feedback to indicate whether the GIS coordinates were accurate. This end-user verification could then be used as the basis for an API or chargeable standard to ensure data quality or for the end-user to directly supplement the market dataset.

Lastly, based on SME feedback, we believe there could be more accurate meter location data stored on trading party and MRSP systems outside of CMOS. There is currently no mechanism for this to be shared with the market. This could be addressed in the near-term by MOSL coordinating a data sharing exercise with the market to identify inconsistencies and update the market dataset. In the longer-term, this will be addressed by providing end-users a mechanism for uploading a version of the data directly to the market which can then be accepted by the data owner.

Recommended next steps

1. We will use the indicator for inaccurate GIS coordinates as the basis for a wholesaler API to drive improvements to data quality. This could, initially, include peer comparison and performance rectification. We will prioritise premises by consumption and occupancy
2. We will continue to develop a mechanism for allowing end-user verification of data through an Application Programming Interface. This could also be used as a mechanism for the end-user to provide more accurate meter location data
3. We will work with the RWG to establish best practice guidance for meter location description.

3.4 Meter details data

Observations

- 25 per cent of meters have been assigned a manufacturer that does not directly correspond with a known manufacturer. This includes errors, misspellings and unnecessary additional information
- Meter detail mismatches accounts for 23 per cent of meter read rejections
- An accuracy metric cannot be defined with available data but could be verified by end-users
- We have identified approximately 8.4k meters with doubtful meter sizes or number of digits
- Retailers are not incentivised to provide accurate meter detail data in CMOS and could have more accurate data on their own systems
- 16 per cent of unplanned settlement runs were caused by changes to meter data with a total impact on settlement of £318 million in the financial year 2019/20
- Inaccurate meter details data (such as number of digits) has been the cause of 23 per cent of unplanned settlement runs, corresponding to a settlement impact of £765mln since market opening
- 652 meters (0.05 per cent) have been assigned a number of digits between 0 and 1
- Physical meter size is used to determine meter read frequency. Chargeable meter size underpins the tariff calculation; and where there is no read or YVE it is utilised for the industry level estimate (ILE). There may be differences between the physical and chargeable meter size, for example this will occur when there has been a change of use
- 3,177 meters (0.24 per cent) have been assigned a physical meter size of between 0-5mm.

Impact

- Invalid or inaccurate meter details can cause meter read rejections, which in turn leads to inaccuracies in settlement, leakage calculations and customer billing
- Additional costs to trading parties through unplanned settlement runs or query resolution
- Inaccurate physical meter size issues could lead to incorrect read frequency being applied potentially leading to increased meter reading costs.

Possible solutions

- Remove the requirement to provide the meter manufacturer from CMOS validation when submitting a meter read
- Peer comparison and performance rectification using indicators of inaccuracy to flag invalid meter manufacturer and meter serial number details
- Coordinated data sharing to update CMOS data with more accurate data from wholesaler and retailer systems
- Develop a mechanism for end-users to verify the accuracy of meter details
- Implement a standardised drop-down option within CMOS for the meter manufacturer and include stricter validation for the format of meter serial numbers.

Discussion

Meter detail mismatches accounts for 23 per cent of rejections of meter read submissions, with a mismatched meter read manufacturer being the main cause. CSD 0301 ('Data Catalogue') states that the meter manufacturer (D3013) must "specify the make and/or manufacturer of a meter". CMOS, however, provides a free text field and does not ensure that the details entered correspond to a valid meter manufacturer. We have produced a list of 48 known meter manufacturers (see Table 5) in order to use the correspondence of manufacturers on CMOS with this list as a validity metric. Note that the correspondence is not case-sensitive, but misspellings or abbreviations of known meter manufacturers, such as 'SOC' instead of 'SOCAM', are not recognised as valid in this analysis.

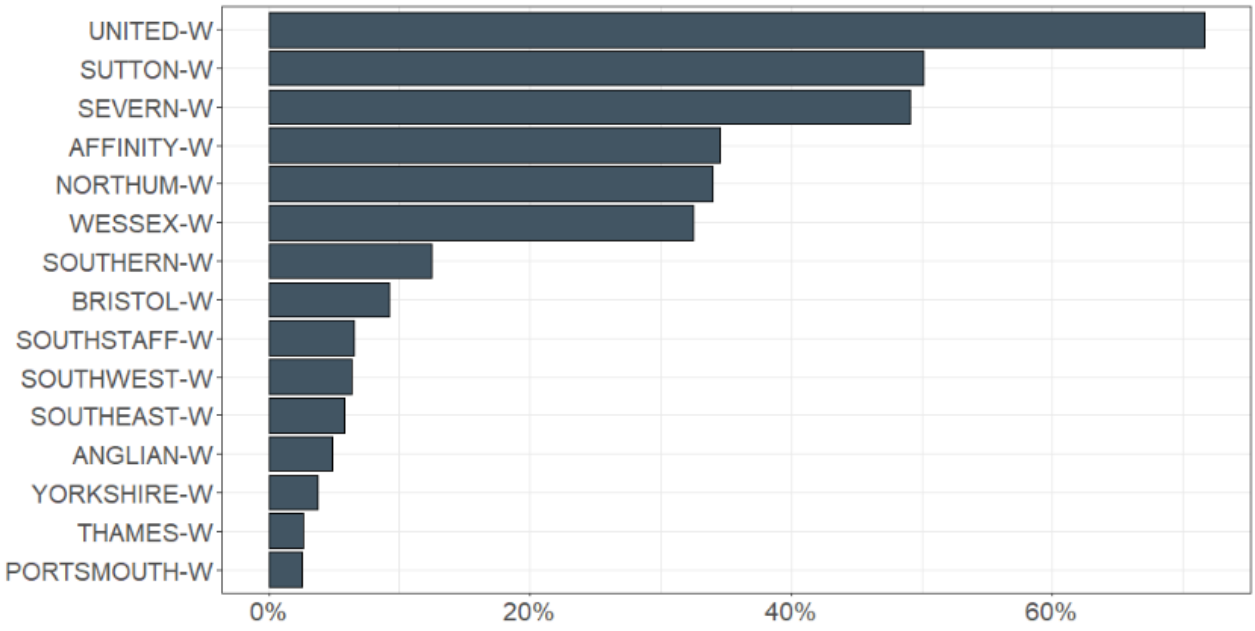
Table 5: Known meter manufacturers

ABB	BURKERT	HYCONTROL	NWM	SMARTSTORM
ACTARIS	DANFOSS	ITRON	PULSAR	SMC
AHS	DELTAFLOWTECH	KENT	RAMAR	SOCAM
APATORPOWOGAZ	DETECTRONIC	KROHNE	RELIANCE	SPARLING
AQUAMOTION	DIEHL	MADDALENA	SAPPEL	TAGUS
ARAD	ELSTER	MANCHESTER	SCHLUMBERGER	TELEDYNE ISCO
ARKON	ENDRESS & HAUSER	METRON	SENSUS	VEGA
BADGER	GENEBRE	MICRONICS	SGMLEKTRA	ZENNER
BELLFLOW	GEORGE FISCHER	NEPTUNE	SIEMENS	
BMETER	GWF	NORSTROM	SISMA	

Using the validity metric for meter manufacturers, we have found that 330,000 meters (25 per cent) have not been assigned a recognisable meter manufacturer. The performance of wholesalers against this metric is

summarised in Figure 9. Performance varies considerably, with some having close to zero invalid meter manufacturers (e.g. Thames Water and Portsmouth Water) and others having more than 70 per cent of meters with invalid details (United Utilities). In the case of United, approximately 74k of their meters (42 per cent) use the abbreviation 'SAPP' instead of 'SAPPEL'.

Figure 9: Proportion of meters with Invalid meter manufacturer details by wholesaler



Source: MOSL

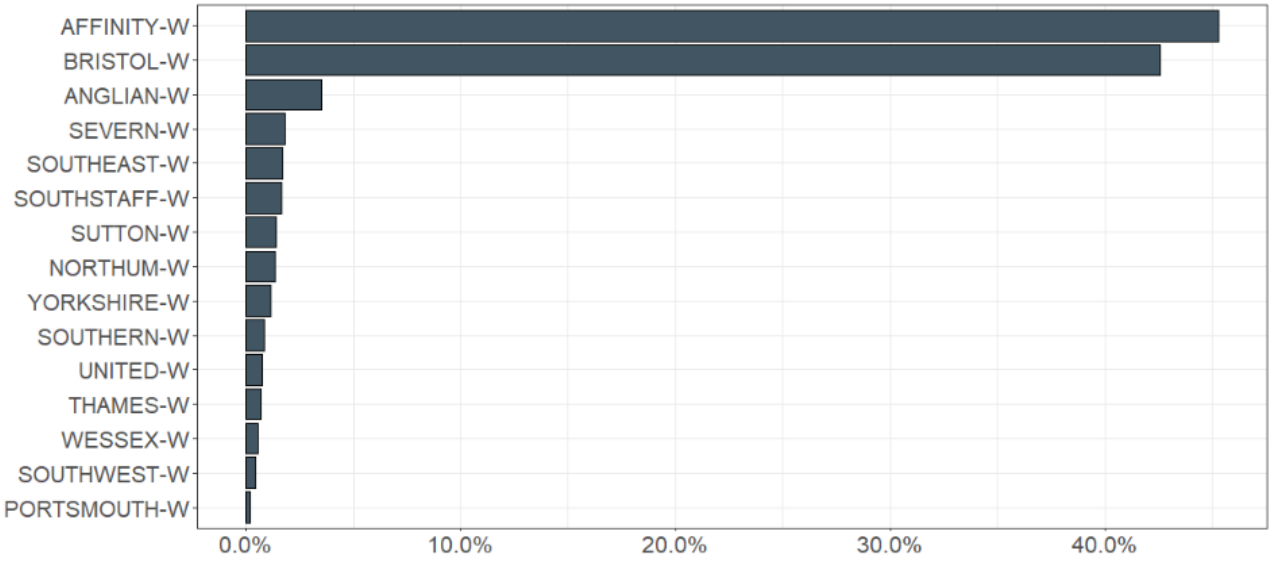
Based on the findings of this RFI, we propose to either remove the validation requirement to provide a meter manufacturer when submitting a meter read into CMOS or to work with wholesalers to ensure that all manufacturers on CMOS correspond with a pre-defined list of manufacturers maintained by MOSL.

We have also created a validity indicator for meter serial number to flag the following issues:

- Contains more than two consecutive letters
- Length is less than five characters
- Contains special characters.

Using this indicator, we have flagged 60,280 serial numbers (4.6 per cent) with at least one of these issues. The performance of wholesalers against this metric is summarised in Figure 10. The overall level of problematic meter serial numbers is low; and there are only two wholesalers that are particularly affected (Affinity Water and Bristol Water). We will work with these two wholesalers to reduce these issues using the indicator of problematic serial numbers as a basis.

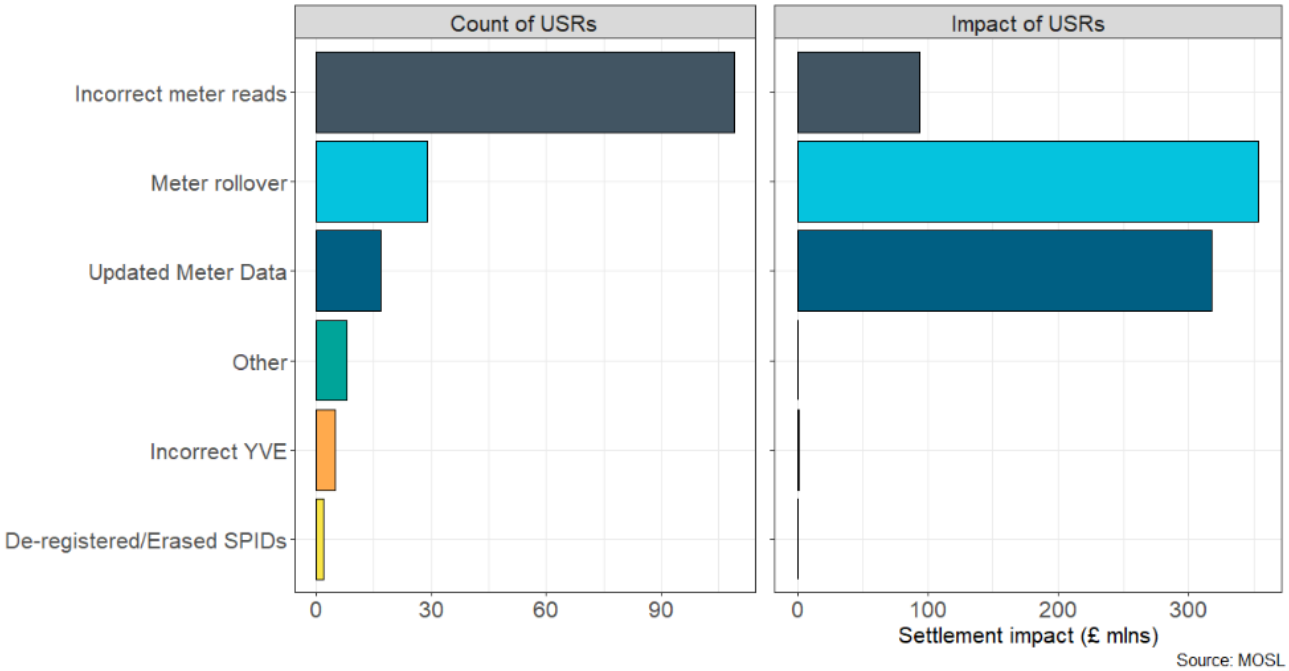
Figure 10: Proportion of meters with an invalid meter serial number by wholesaler



Source: MOSL

The accuracy of meter details, particular number of digits, has impacted the number of unplanned settlement runs. Figure 11 highlights that changes to meter data has been the reason given for 16 per cent of unplanned settlement runs with a total impact of £318 million (41 per cent of total impact on settlement) in the financial year 2019/20. However, we currently have no way to measure the accuracy of physical meter size or number of digits without end-user verification. Further work is required to see whether the occurrence of unplanned settlement runs can be reliably predicted.

Figure 11: Reasons and impact for unplanned settlement runs



Source: MOSL

Recommended next steps

1. We will investigate the possibility of removing the validation requirement to provide a meter manufacturer when submitting a meter read into CMOS. If this is not reasonable then we will use the validity metric for meter manufacturer details as the basis for a wholesaler API to ensure that all manufacturers on CMOS correspond with a pre-defined list of manufacturers maintained by MOSL. The latter could, initially, include peer comparison and performance rectification. We will also investigate whether CMOS could incorporate a standardised drop-down list for meter manufacturer details if it cannot be removed from the validation requirement.
2. We will use the validity metric for meter serial number as the basis for wholesaler APIs to ensure data quality is maintained. This could initially include peer comparison and performance rectification. We will prioritise premises by consumption and occupancy. In the longer-term we will seek to improve CMOS validation rules for meter serial numbers.
3. We will continue to develop a mechanism for allowing end-user verification of data through an Application Programming Interface. This could also be used as a mechanism for the end-user to provide more accurate meter details data.

4. Conclusion

4.1 Recommendation

We have found that missing premises identifiers hinders retailers from identifying premises and tendering and switching customers. We have proposed to use performance rectification and incentivisation to ensure that premises identifiers are being provided by wholesalers in accordance with the codes.

We have also found there is a lack of genuine verification of market data. This obstructs the measurement of the accuracy of core market data items and data cleanse activity. New technology and processes are required to facilitate end-users to verify data. In response to this need, we have said that we will develop an Application Programming Interface that would allow end-users (such as meter reading service providers) to supplement market data with more accurate meter location and meter details data.

Lastly, we have highlighted the need for a data sharing exercise as a first step to cleansing core market data and to allow non-data-owners to provide more accurate data into the market.

4.2 Next steps

We will analyse the results of the RFI and incorporate the findings into our planning for data cleanse activity. We will consult with the market based on a draft plan before publishing the final version of the data cleanse plan to the market by the end of February 2021. Note that we may also undertake certain priority activities to improve the quality of market data before the final plan is published.

Our planning will balance the prioritisation of improvement activity based on the expected impact, cost and the amount of time required to implement. For example, developing an API and using performance rectification tools is relatively straightforward compared to developing a chargeable standard or implementing new technology—however, the latter may provide more impactful and enduring solutions.

The timelines for next steps have been summarised in Table 6.

Table 6: Timeline for publication of data cleanse plan

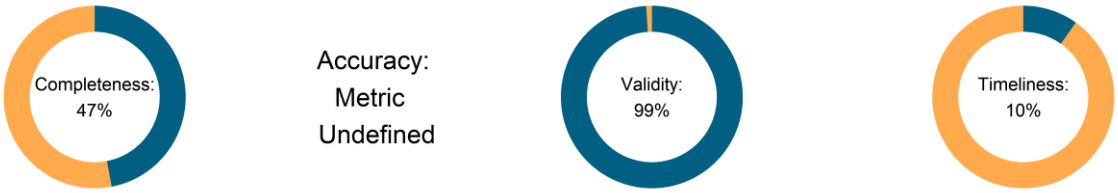
Milestone	Status	Target Completion
Agree high-level approach for data cleanse with Market Performance Committee (MPC)	Complete	01-Jul-20
Define scope, issues and objectives for data cleanse	Complete	29-Jul-20
Develop data quality metrics or indicators and highlight gaps	Complete	19-Aug-20
Analysis of core data items, identification of possible solutions and input from subject matter experts	Complete	09-Sep-20
Prepare RFI document for consultation and launch RFI	Complete	19-Oct-20
Complete RFI (3 weeks)	On Track	06-Nov-20
Analysis of RFI responses and incorporate findings into data cleanse planning	On Track	27-Nov-20
Finalise draft data cleanse plan	On Track	30-Nov-20
Consultation on draft plan	On Track	30-Jan-21
Finalise data cleanse plan and publish to the market	On Track	26-Feb-21

Appendix I

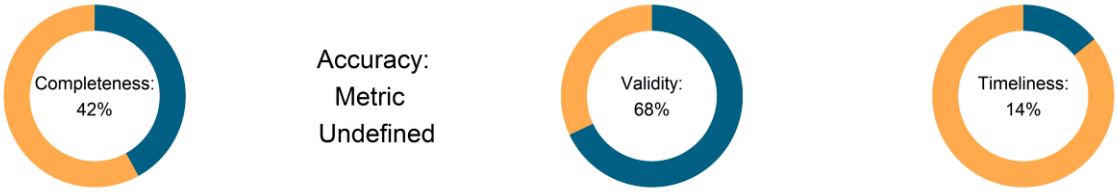
A1. Data quality dashboards

Customer details and premises data

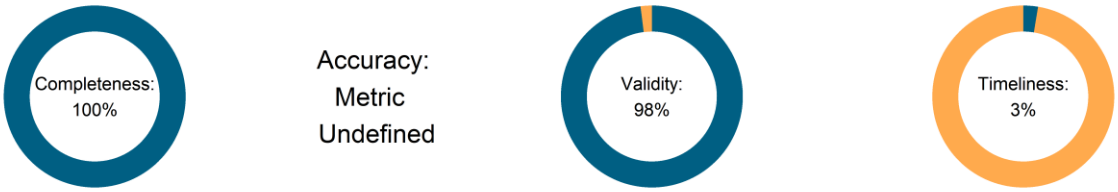
UPRN Code



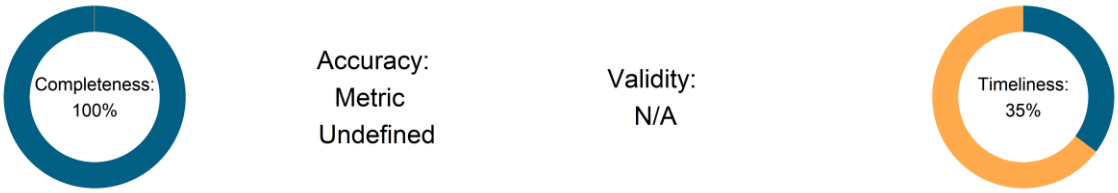
VOA BA Reference



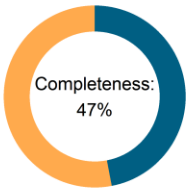
Postcode



Customer Name

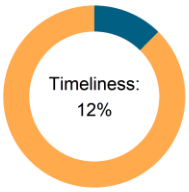


Customer Banner Name



Accuracy:
Metric
Undefined

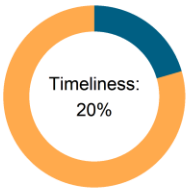
Validity:
N/A



Meter location data

GISX and GISY Coordinates

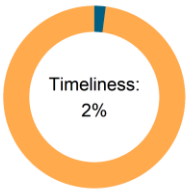
Accuracy:
Metric
Undefined



Meter Location Free Descriptor

Accuracy:
Metric
Undefined

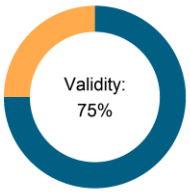
Validity:
N/A



Meter details data

Meter Manufacturer

Accuracy:
Metric
Undefined



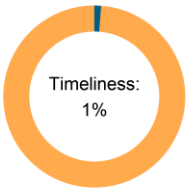
Meter Serial Number

Accuracy:
Metric
Undefined

Validity:
N/A

Physical Meter Size

Accuracy:
Metric
Undefined



Number of Digits

Accuracy:
Metric
Undefined