BUILDING
INNOVATION
PARTNERSHIP

# Technical Note:

# Assessing Data Quality as a Stage in Federating Asset Data

## Better Investment Decisions (Theme 1)

This report is an output from the Quake Centre's Building Innovation Partnership programme (BIP), which is jointly funded by industry and the Ministry of Business Innovation and Employment (MBIE). This report provides a summary of the work undertaken in support of the Building Innovation Partnership project to develop a proof-of-concept National Pipe Data Portal.

Quake Centre
Delivering Industry Solutions

# Technical Note –

# Technical Note on Assessing Data Quality as a Stage in Federating Asset Data

**Technical Note Summary**

This technical note presents a short summary of the considerations for assessing the quality of infrastructure asset data which is intended for federating with other similar data sources. This note supplements the concepts proposed in the previous Quake Centre publication 'Quake Centre Proof-of-Concept National Pipe Data Portal Report' and the technical note 'Technical Note on Data Federation Jan 2020'.

There are many dimensions to consider when assessing the quality of a particular dataset. These dimensions are well researched and will be well understood by those who maintain datasets of infrastructure assets[1]. This short technical note places emphasis on the three questions which relate to all datasets representing infrastructure assets:

1. What is the quality of this data?
2. What is the quality of this data compared to data held by others? And;
3. What is the completeness of this data as anticipated by a standard?

This technical note proposes an approach to assessing the quality of infrastructure data – three principles are proposed:

- Standardise an approach to assessing the quality of a dataset;
- Create a standardised reporting template to readily compare one dataset to another (or compare one dataset with a baseline/federated dataset). And;
- Create a standardised reporting template which demonstrates the strength of alignment (completeness) of the data with respect to the standard.

The reason for having a standardised approach to assessing the quality of data is that it allows decision makers to determine the level of confidence they place in the decisions made from the data they have – and to determine the effort required to improve data quality and therefore confidence in evidence-based (data-driven) decision making. Standardised data quality assessment processes also allows for national-level understanding of asset infrastructure at a federated level – providing a basis for policy-setting, funding forecasts, assessment of infrastructure resilience and disaster recovery.

[1] https://en.wikipedia.org/wiki/Data_quality

## Discussion

One of the more significant barriers to federating infrastructure asset data is the reluctance of asset owners to publish their data and expose the presumed poor quality and completeness of their dataset. This is a mistake, that is to say that asset owners are mistaken in thinking that their data is too impoverished to publish – asset owners understand the value of data and so lack of quality cannot be attributed to ignorance or apathy - inaccurate or incomplete data is often simply a reflection on impoverished historic data collection due to limited data capture and storage systems. Data, even incomplete and unstructured data, is of value and provides a foundation for multi-year incremental improvement in the quality of the data. Also, improvements can be readily made to data once the data is exposed. The act of exposing and publishing data is often the first (and best) step to improving the quality of the data.

For this reason, it is important to have a standardised method of assessing the quality of a dataset which represents infrastructure assets. Once it is known, for example, that a dataset is missing important attribute data for a core asset class (object) then a plan can be drawn-up to remedy this. If we commit to decreasing missing attribute data then we can devise processes in support of this and measure progress year-on-year.

Once we understand the quality of the datasets we hold, the next question is how does this compare to the industry as a whole? If my data is missing 90% of a key attribute value, say the age of my storm water pipes, is this good or bad compared to other regions and asset owners? I might think that this suggests my data is very poor in terms of data completeness but without a method to benchmark the quality of a dataset I don't know if this is actually good or bad (compared to the rest of the sector). By creating an environment and process for data federation and by standardising the method of assessing data quality we can readily compare the quality of one dataset with another. To assist this, it is important to have standardised reporting templates which provide insight into the quality of datasets and how they compare to the federated baseline.

But, even if we can standardise the assessment of data quality and provide mechanisms to compare the quality of one dataset with a baseline of federated data how would we know whether we are improving the data we care about – the important data that provides the best return on the investment of collecting more data. We may not care that we're missing 90% of data compared to sector-leading datasets if the 10% we do have is data which allows us to effectively manage our local assets. This question demands that as a sector we agree on a data standard against which the quality (and completeness) of my data can be compared. This standard can highlight the asset classes (objects) and attributes which are 'core' – those we care about the most. Those objects and attributes which provide data from which useful insight and meaningful decisions can be made.

As an example of the power of this approach, during the QC project 'Quake Centre Proof-of-Concept (POC) National Pipe Data Portal' a team of data science students analysed asset data from a number of local councils. Each council had provided core asset data on stormwater and waste water pipes – with some councils providing additional data relating to capital replacements costs and remaining useful asset life. The students were able to derive estimates for asset replacement costs and useful asset life for those councils who did not

provide this data based on regression techniques which used the core data that was provided for all councils. Having a 'core' set of data for the important asset classes (objects – in this case water pipes) allowed us to leverage the power of data federation to derive valuable insight into the assets of other organisations where data is missing. Figure 1 illustrates the steps to assessing the quality of infrastructure asset data.

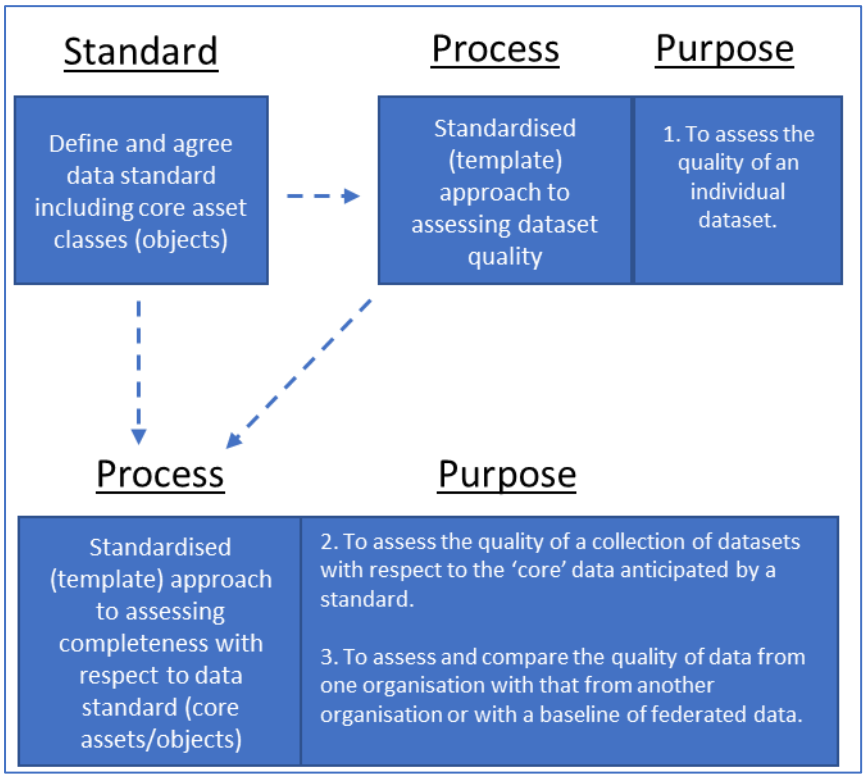



*Figure 1: Steps to assessing data quality*

**Example Data Quality Reporting**

Appendix A provides examples of proof-of-concept data quality templates undertaken by scholarship students in support of Quake Centre's Building Innovation Partnership.

**Where to next** – This technical note should be read together with the previous QC publications (Quake Centre Proof-of-Concept National Pipe Data Portal Report and Quake Centre Technical Note on Data Federation). Data quality is a key aspect of the data federation domain and there are readily implementable processes which would provide clarity and direction for the industry. These include:

- Defining standards – this is also a key recommendation from the 'Quake Centre Technical Note on Data Federation' with an additional requirement that 'core' asset classes (objects) are identified for data quality assessment.
- Design standardised reporting template and processes based on the exemplars in the appendix to this report.

Note prepared by: Angus Bargh, Open Plan Ltd

# Appendix A

Below are example data quality reports produced by scholarship students. These students focused on producing data quality reports for the data provided for the Quake Centre Proof-of-Concept National Pipe Data Portal – this was a proof-of-concept project completed in 2019.

The objectives and methodology for the production of these reports are also provided below – as a direct extract from the final report of the scholarship work.

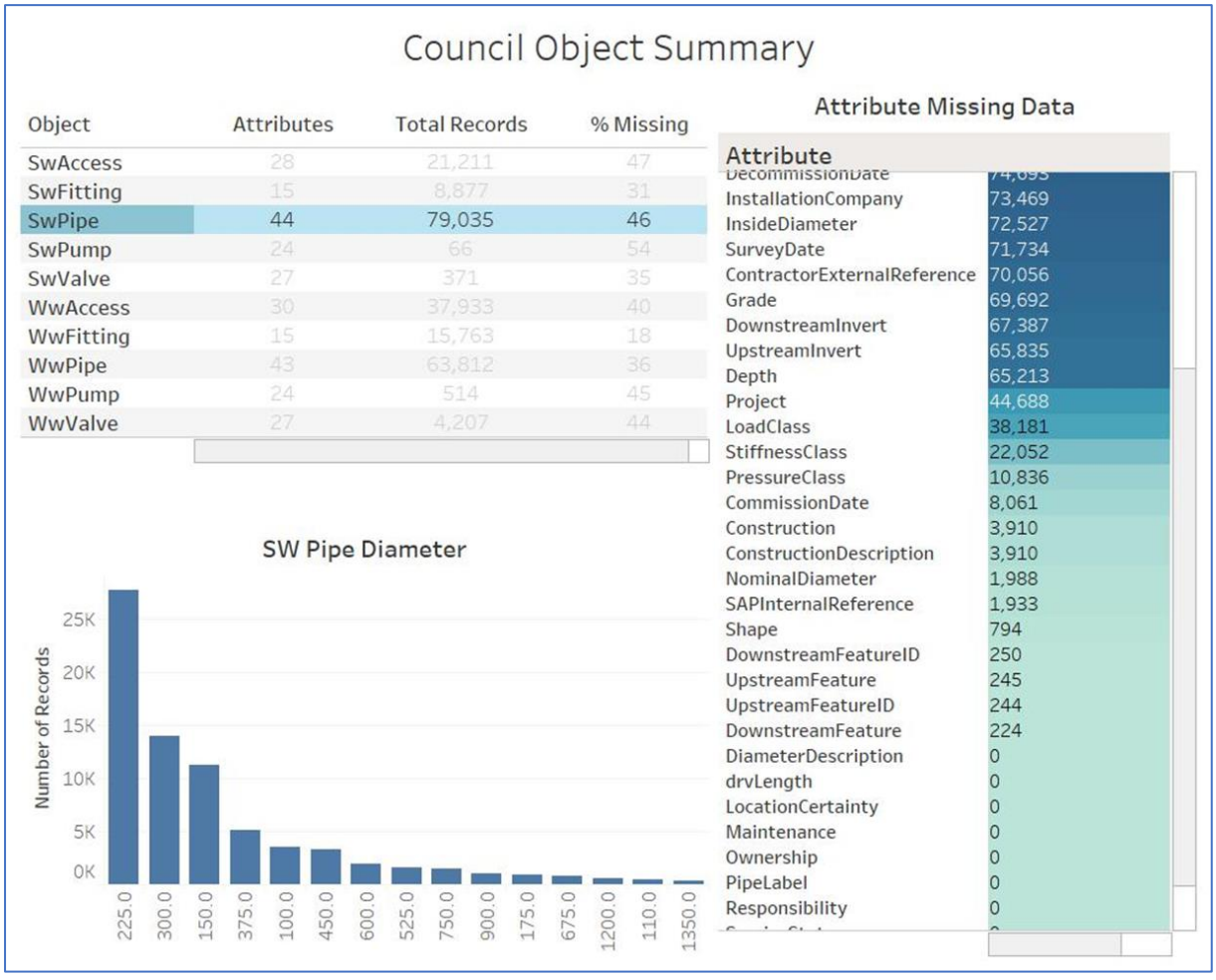



*Figure 2: Example Quality Report for asset object dataset*

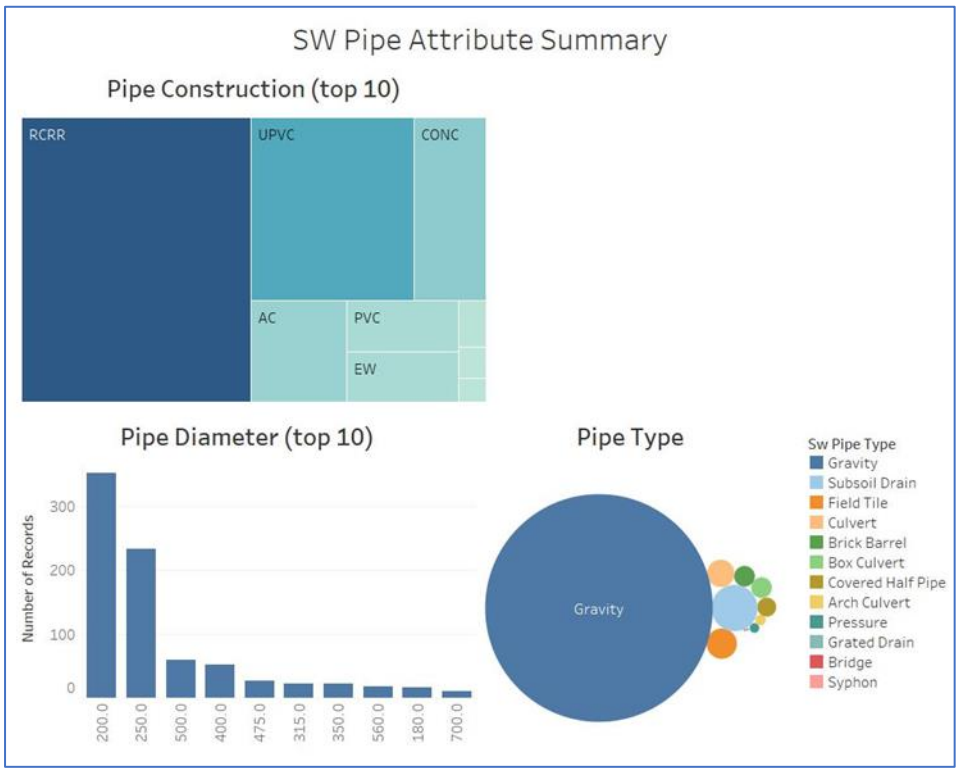
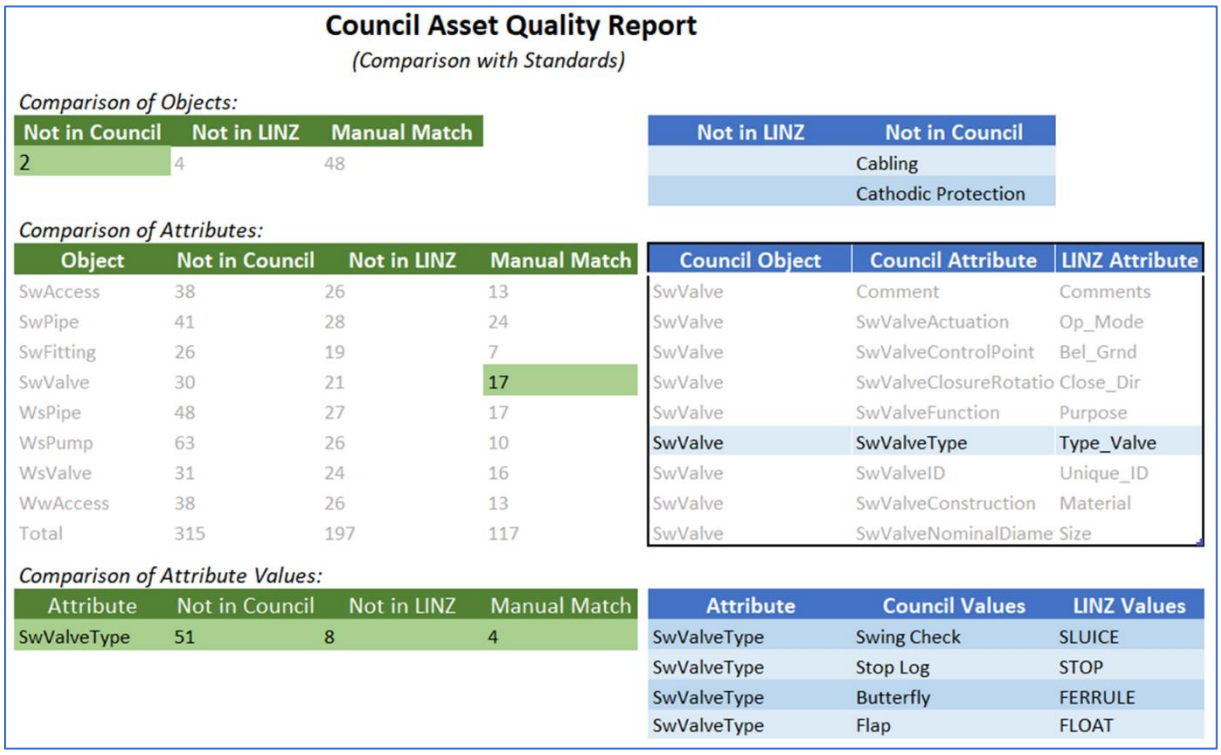



*Figure 3: Example Quality Report for asset object dataset*



**Council Asset Quality Report**

*(Comparison with Standards)*

*Comparison of Objects:*

| Not in Council | Not in LINZ | Manual Match |
|---|---|---|
| 2 | 4 | 48 |

| Not in LINZ | Not in Council |
|---|---|
| | Cabling |
| | Cathodic Protection |

*Comparison of Attributes:*

| Object | Not in Council | Not in LINZ | Manual Match |
|---|---|---|---|
| SwAccess | 38 | 26 | 13 |
| SwPipe | 41 | 28 | 24 |
| SwFitting | 26 | 19 | 7 |
| SwValve | 30 | 21 | 17 |
| WsPipe | 48 | 27 | 17 |
| WsPump | 63 | 26 | 10 |
| WsValve | 31 | 24 | 16 |
| WwAccess | 38 | 26 | 13 |
| Total | 315 | 197 | 117 |

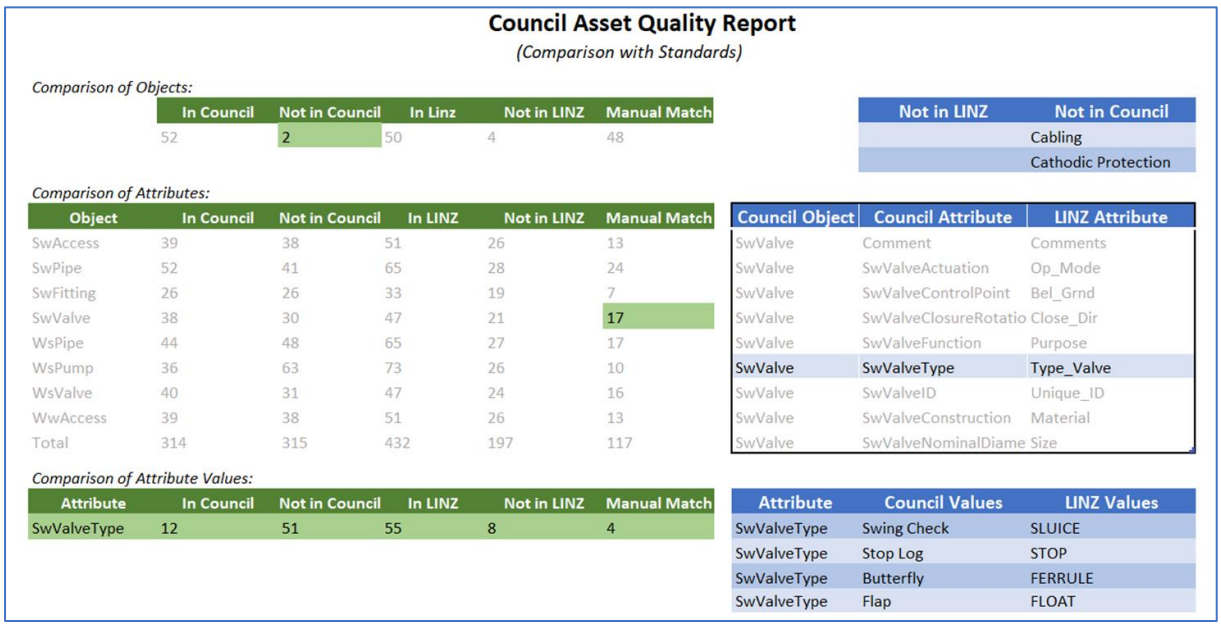
| Council Object | Council Attribute | LINZ Attribute |
|---|---|---|
| SwValve | Comment | Comments |
| SwValve | SwValveActuation | Op_Mode |
| SwValve | SwValveControlPoint | Bel_Grnd |
| SwValve | SwValveClosureRotatio | Close_Dir |
| SwValve | SwValveFunction | Purpose |
| SwValve | SwValveType | Type_Valve |
| SwValve | SwValveID | Unique_ID |
| SwValve | SwValveConstruction | Material |
| SwValve | SwValveNominalDiame | Size |

*Comparison of Attribute Values:*

| Attribute | Not in Council | Not in LINZ | Manual Match |
|---|---|---|---|
| SwValveType | 51 | 8 | 4 |

| Attribute | Council Values | LINZ Values |
|---|---|---|
| SwValveType | Swing Check | SLUICE |
| SwValveType | Stop Log | STOP |
| SwValveType | Butterfly | FERRULE |
| SwValveType | Flap | FLOAT |

*Figure 4: Example Quality Report for data quality assessment with respect to a standard*



**Council Asset Quality Report**

*(Comparison with Standards)*

*Comparison of Objects:*

| In Council | Not in Council | In Linz | Not in LINZ | Manual Match |
|---|---|---|---|---|
| 52 | 2 | 50 | 4 | 48 |

| Not in LINZ | Not in Council |
|---|---|
| | Cabling |
| | Cathodic Protection |

*Comparison of Attributes:*

| Object | In Council | Not in Council | In LINZ | Not in LINZ | Manual Match |
|---|---|---|---|---|---|
| SwAccess | 39 | 38 | 51 | 26 | 13 |
| SwPipe | 52 | 41 | 65 | 28 | 24 |
| SwFitting | 26 | 26 | 33 | 19 | 7 |
| SwValve | 38 | 30 | 47 | 21 | 17 |
| WsPipe | 44 | 48 | 65 | 27 | 17 |
| WsPump | 36 | 63 | 73 | 26 | 10 |
| WsValve | 40 | 31 | 47 | 24 | 16 |
| WwAccess | 39 | 38 | 51 | 26 | 13 |
| Total | 314 | 315 | 432 | 197 | 117 |

| Council Object | Council Attribute | LINZ Attribute |
|---|---|---|
| SwValve | Comment | Comments |
| SwValve | SwValveActuation | Op_Mode |
| SwValve | SwValveControlPoint | Bel_Grnd |
| SwValve | SwValveClosureRotatio | Close_Dir |
| SwValve | SwValveFunction | Purpose |
| SwValve | SwValveType | Type_Valve |
| SwValve | SwValveID | Unique_ID |
| SwValve | SwValveConstruction | Material |
| SwValve | SwValveNominalDiame | Size |

*Comparison of Attribute Values:*

| Attribute | In Council | Not in Council | In LINZ | Not in LINZ | Manual Match |
|---|---|---|---|---|---|
| SwValveType | 12 | 51 | 55 | 8 | 4 |

| Attribute | Council Values | LINZ Values |
|---|---|---|
| SwValveType | Swing Check | SLUICE |
| SwValveType | Stop Log | STOP |
| SwValveType | Butterfly | FERRULE |
| SwValveType | Flap | FLOAT |

*Figure 5: Example Quality Report for data quality assessment with respect to a standard*

**Scholarship Methodology Statement (from the student scholarship final report).**

## Background:

The Quake Centre is leading the development and implementation of National Metadata Standards for 3 Waters in NZ. As part of this, the Centre has been working with students from the Master of Applied Data Science to develop strategies for mapping data held within Local Authorities to a Beta version of a National Standard. This has been carried out for 3 local authorities to date. The Centre is also working with other agencies including the New Zealand Transport Agency (NZTA) to align national standards and roll them out across NZ. The eventual aim is to develop an integrated National Digital Infrastructure Model (NDIM). The work undertaken with CCC, Tauranga and Auckland Councils has proved the concept of the NDIM.

## Introduction:

Following the successful POC with 3 councils to develop National Infrastructure portal the next step is to roll-out mapping of other councils' data to the national standard across the country. To enable this there is a need to automate the mapping as much as possible. To this end it is proposed that a tool or tools are developed to undertake automatic mappings where possible.

The main aim of this research project was to develop a process flow for automation of the Data extraction and Mapping of Council data this would include the following steps.

1. Identify the objects in the council data
2. Extract object attributes from the Council data
3. Carry out an internal quality check for values such as total assets in council database, total number of attributes with NULL values, number of unique values, the maximum/minimum values or Range of values. Data errors and presence of metadata.

4. Carry out a quality check against standards which will compare the council data with the national standards at the class level followed by attributes and code list /values.
5. The final step in the process will be to develop a data dictionary to map the council attributes against the attributes and values in the national metadata standard.

As a part of the process development data from 6 councils was examined for comparison and understand the object identification process. This included data from the following councils.

1) Christchurch City Council

2) Tauranga City Council

3) Auckland City Council

4) Wellington City Council

5) Queenstown City Council

6) Bay of Plenty City council

It was also proposed to develop 2 sample dashboards as a part of the Quality check process.

1) Council Internal Quality check: This would provide a snapshot summary of the council's internal data with graphics which would serve to highlight the shortcomings in the council data. It should also be possible to provide a comparison of data from different councils

2) Council Quality check against standards: This would provide a snapshot summary of how the data from the council fared against the national metadata standards. This would give a summary at the Class level followed by Attribute and code lists and finally the code list values.

## Roadmap ahead

Post completion of the research project the next steps in the development of the infrastructure portal and mapping of council's data to the national standards would include the following steps.

1) Automate the Mapping process by identifying the objects corresponding to asset classes within the council data for e.g. Pipes, Valves, Fittings, Access Chambers etc.

2) Clearly define the objects within the National metadata standards to facilitate the mapping of council objects to them supported by a dynamic data dictionary for values.

3) Define the relevant metadata for the councils with timestamps to help validate the data from the councils

4) Define Quality metrics for the council data.

These steps could be undertaken as a continuation of the current project or as a separate project

# Data Extraction and Mapping Process

The information below presents the technical methodology for compiling the data quality checks and reporting. For further information and copies of the technical processing scripts an approach should be made to Quake Centre.

## Native Data Structure

The format / hierarchy of Asset data received for CCC, Wellington, TCC, QCC, Auckland was inspected. The Asset data at each of the councils, is stored in geospatial data format (.gdb or .shp files) and is grouped by network (storm water, wastewater, and water supply). The information is further divided into multiple layers by asset classes (such as pipe, access chambers etc.), with **a separate layer for each specific asset class** within the network. Each layer contains the spatial data of the asset and its own attributes and attributes values. Data stored thus is input to the process for extracting attributes.

## Identify Objects

The information stored in the geospatial data base is in layers with a separate layer for each specific asset class. Objects are thus pre-identified within the geospatial data structure.

## Extract Object Attributes

This step would extract the network, geometry and attributes for each asset class. The extracted information looks like below:

| Columns | Description | Sample Value (SW Pipe) | Sample Value (WW Pipe) |
|---|---|---|---|
| Layer Name | Layer Name | vwOpenDataSwPipe | vwOpenDataWwPipe |
| Asset Id | Index starting at 1 | 1 | 1 |
| Geometry Type | Geometry type, which is defined by numbers | 5 | 5 |
| Geometry Name | Geometry name | MULTILINESTRING | MULTILINESTRING |
| GeoXLO | Downstream Longitude | 1579318 | 1580701 |
| GeoYLO | Downstream Latitude | 5178496 | 5176025 |
| GeoXHI | Upstream Longitude | 1579340 | 1580781 |
| GeoYHI | Upstream Latitude | 5178504 | 5175907 |
| Attributes extracted | Feature attributes | | |

## Checking

This step would inspect various extracted objects, their attributes and values.

### Internal Quality Check

In this step, certain statistical measures and consistency for the attribute values would be checked. For each object, **a file output** would be generated with the following measurements.

| Measurement | Description / Output | Sample | Notes |
|---|---|---|---|
| Total Count | Total number of records | | |
| Missing | Number of records with missing values | | |
| Complete | Number of records with values (number of non-null records) | | |
| % Complete | % of records with values | | |
| Data type | Data type – Integer, Float, Character, Date, Alphanumeric | | The program would infer the data type |
| Unique | Count of unique values (if character, alphanumeric, or | | |
| Maximum | Maximum attribute value | Only if Integer or float | |
| Minimum value | Minimum attribute value | Only if Integer or float | |
| Mean | Mean value | Only if Integer or float | |
| Mode | Mode | Only if data type is character | |
| Range | A Box-plot or summary table | If Integer or Float | Dashboard gives distribution of values. |
| Frequency by Value | Histogram or a density plot | If Integer or Float | Dashboard gives distribution of values. |

<table>
<tr><td>Zeroes</td><td>Count of attribute values that are zero</td><td></td><td></td></tr>
<tr><td>Negatives</td><td>Count of negative attribute values</td><td></td><td></td></tr>
<tr><td>Data Error 1</td><td>Count and list of attribute values with unexpected value (values different from a provided list)</td><td>E.g. for attribute ‘Pipe Material’ the possible values may be ‘Concrete’, ’Plastic’ and ‘Clay’. ‘Wood’ is not in this list.<br>list of unexpected values –<br><table><tr><td>unexpected value</td><td>count</td></tr><tr><td>Copper</td><td>5</td></tr></table></td><td>Can be done for a predefined set of attributes, if permissible values for it are available.<br><br>A list of code list values for object attributes is available for Christchurch and can be used.</td></tr>
<tr><td>Data Error 2</td><td>Count and list of mismatched values</td><td>12 mismatched values<br><table><tr><td>Permissible Value</td><td>Observed value</td><td>count</td></tr><tr><td>CONC</td><td>Concrete</td><td>10</td></tr><tr><td>CONC</td><td>Conc4</td><td>1</td></tr><tr><td>LDPE</td><td>LLDPE</td><td>1</td></tr></table></td><td>Same notes as above apply.<br><br>Auto check if textual match up to a certain threshold. Else, manually mismatched.</td></tr>
<tr><td>Is Meta Data Present?</td><td>Y / N if Metadata</td><td></td><td>What would this be checked against / council to provide?</td></tr>
<tr><td>Meta Data update</td><td>Last update date / (external) version number</td><td></td><td>Same as above</td></tr>
</table>

### *Quality Report (Internal Quality Check):*

Above output would be used to produce a dashboard / report shown earlier. This dashboard has been done in Tableau following below steps:

a. Outputs using scripts from earlier work that listed various statistic on objects were collated for various objects
b. The files were read into Tableau and a union of these files was done.
c. Columns were renamed, calculated measures were added for % missing, object name was extracted from file name.

d. Created a filter on object attribute to exclude reporting attributes such as attribute geometry and 8 such other attributes and shared across worksheets.
e. Created “ObjectSummary” worksheet to report Objects, the count of attributes, total and % of missing records for the object.
f. Created “AttributeData ” worksheet to show count of missing data for every object attribute.
g. Created a dashboard that combined these two worksheets and linked the two using action filters.
h. Created worksheets to visually depict the variation for a few other pipe attributes such as the Pipe diameter, Pipe construction and Pipe Type.

### Reference files:

a. *Dashboard_InternalQualityReport (Tableau Packaged work book)*
b. *Files_InternalQualityReport*

### Checking Against Standards

The step would compare the extracted objects, their attributes and values with the standards. A file output with below information would be generated:

#### *Comparison of Objects*

A file output listing all objects within the council and the standards would be generated. Below is a sample output. The mapping may be done manually if easier.

| Council Object | Standards Object | Match Type | Notes |
|---|---|---|---|
| SwValve | Valve | Manual Match | Matching object present within standards |
| | Cabling | Not in Council | Object present in standards but not in council |
| SwGrill | | Not in Standard | Object present in council but not in standards |

*Table 1: Comparison of Objects between standard and council*

#### *Comparison of Attributes*

A file output comparing the attributes for every council object with the standards would be generated. Below is a sample output. The mapping may be done manually if easier.

| Council Object | Council Attribute | Standards Object | Standards Attribute | Match Type | Notes |
|---|---|---|---|---|---|
| SwValve | SwValveID | Valve | Unique_ID | Manual Match | Matching attribute present |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | within standards |
| SwValve | | Valve | Purpose | Not in Council | Attribute present in standards but not in council |
| SwValve | SwValveResponsibility | Valve | | Not in Standard | Attribute present in council but not in standards |

*Table 2: Comparison of Objects attributes between standard and council*

### *Comparison of Attribute Values:*

A file output comparing the values for an attribute of the council object, with the values available within the matching object in standards would be generated. Below is a sample output for attribute SwValveType of SwValve object. This would be done for a council object having a matching object in the standards and if the council object has an associated list of value (code list is present). The mapping may be done manually if easier.

| Council Object | Council Attribute | Council Value | Standards Object | Standards Attribute | Standards Value | Match Type | Notes |
|---|---|---|---|---|---|---|---|
| SwValve | SwValveType | Butterfly | Valve | Type_Valve | FERRULE | Manual Match | Matching value present within standards |
| SwValve | SwValveType | | Valve | Type_Valve | WOOFF | Not in Council | Attribute value present in standards but not in council |
| SwValve | SwValveType | Check | Valve | Type_Valve | | Not in Standard | Attribute value present in council but not in standards |

*Table 3: Comparison of attribute code list values between standard and council*

### *Quality Report (Comparison with Standards):*

The above outputs would then be used to produce a dashboard / report shown earlier. The dashboard has been done in Excel following below steps.

a. Listed the objects listed in standards and from council and compared those manually.
b. Outputs generated using scripts from earlier work were collated for various objects to produce output as seen in table 2.
c. Created pivot table for object attributes and pivot details for a chosen object.
d. Generated output as shown in table 3 comparing the attribute values for all attributes for an object chosen in above step.
e. Created pivot table for the values on all attributes and pivot details for a chosen attribute.

These workings and their outputs are shown in the referred file below "*Report - Checking with Standards.xlsx".*

## Reference files:

a. *GISAssetModels.xlsx*
b. *LINZStandards_3Waters.xlsx*
c. *Output files in folder "LINZ auto mapping" generated using CCC_LINZ_auto_mapping.py*
d. *Report - Checking with Standards.xlsx*


Holmes


BMC
BATCHELAR
McDOUGALL
CONSULTING
STRUCTURAL & CIVIL ENGINEERS


ENGEO
Expect Excellence


MINISTRY OF BUSINESS,
INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI


Trust
power


EQC
EARTHQUAKE COMMISSION
Kōmihana Rūwhenua


OPUS


genesis


C. LUND & SON
MAIN CONTRACTORS & BUILDERS


Mercury


UC
UNIVERSITY OF
CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND


Auckland
Council
Te Kaunihera o Tāmaki Makaurau


Fletcher
Building


Fulton
Hogan


Christchurch
City Council


Beca


meridian


Absolutely Positively
Wellington City Council
Me Heke Ki Pōneke


Powell
Fenwick


JJS
JOHN JONES STEEL


Holcim