BUILDING
INNOVATION
PARTNERSHIP

Technical Note:

Data Federation for Infrastructure 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

Information contained in this report has been obtained from sources believed to be reliable. However, neither the Quake Centre, its supporting partner organisations or the authors guarantee the accuracy or completeness of information published herein and neither the organisations or the authors shall be held responsible for any errors, omissions or damages arising out of use of this information. This report is published on the understanding that the authors are suppling information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

Technical Note –

Data Federation for Infrastructure Asset Data

Technical Note Summary

This technical note further develops on the concepts proposed in the previous Quake Centre publication 'Quake Centre Proof-of-Concept National Pipe Data Portal Report'. This previous work, from Quake Centre's Building Innovation Partnership (BIP), has developed the proof-of-concept for a National Pipe Data Portal. This technical note places a focus on the preparation and treatment of data in support of a federation process.

Four key challenges to data federation are described with technical approaches proposed to address each of these challenges. The four challenges are:

1. Data services and access to a platform for data federation.
2. Consistency of data hierarchies
3. Consistency of data and object attributes
4. Consistency of data and object values

This technical note also points to the opportunities to accelerate data federation across the sector. This can be achieved through activities which support the defining of standards and the creation of data-semantics which ensure there is a commonality of language when describing infrastructure in NZ. The next steps proposed in this technical note are:

- Defining the semantics of infrastructure objects – securing stakeholder agreement around what we call infrastructure objects in NZ (this can be achieved through the creation or adoption of a standard).
- Defining the semantics of object attributes - securing stakeholder agreement around what we call the attributes which are used to describe the characteristics of objects (this can be achieved through the creation or adoption of a standard).
- Securing stakeholder agreement on the key attributes required, or expected, for data federation
- Defining the format of values which are ascribed at an attribute level – this to include details of formats, units and content of pre-defined value lists (this can be achieved through the creation or adoption of a standard).

Note on terminology

Terminology can sometimes be confusing especially when discussing technical matters across different knowledge domains. Note that for the sake of this technical note, the term *asset* in the asset management domain is equivalent to the term *object* in the data science domain.

Introduction

This technical note provides a summary of key considerations to assist infrastructure asset owners prepare their data for federation. An example of asset data would include data which describes the waste water network, storm water network, road network, communication networks (cellular, fibre), electricity networks etc. The justification for federating data is not part of this technical note – it is assumed that the reader understands the benefits and opportunities which are enabled through data federation. However, it is worth noting a brief description of what is meant by data federation – to ensure there is suitable context for the contents of this technical note.

Data federation is the process of combining data so that it can be considered, or analysed, as a single dataset. Federation is different from the simple process of appending records to an existing dataset because the separate data sources remain separate through the federation process or, if they are combined, it is just in a temporary sense for the purpose of undertaking discrete analysis. Federation also means that a link is preserved to the authoritative source (or owner) of the data -so that when records in the source data are updated this flows-through to the federated environment. It is good to note that this flow-through doesn't necessarily need to happen instantaneously (although this is valuable) provided the frequency of updates to the federated environment is discoverable (so the user of the data knows the currency). Data federation is also different from the simple task of visually layering data which is common in many spatial applications – this process allows geospatial datasets to be displayed in a layered manner on a map background. Data layering, while providing geospatial context, does not provide a consistent link between the data contained within each layer. To summarise:

- Data federation combines data from different data owners;
- Data federation retains the link to the source data so that when data is updated 'at source' this flows through to the federation environment; and
- Data federation creates linkages between the attributes and values of one data source and another – it is more than the simple layering or unstructured appending of data.

Obviously, in order to achieve federation, the source data must be of a consistent type – meaning that the data must represent similar real-world objects (in the case of asset data). For example, federation would make sense between two datasets representing storm water pipes but it would not make sense between data of storm water pipes and data of storm water pump stations. Both datasets are representative of the storm water network but are obviously different types or different asset categories. Another way to consider or identify data which could be federated is simply a common-sense question of whether there is value is combining *these* data with *th*ose data. Combining data on pipes makes sense as it would provide insight into aspects such as the total length of pipe or quantity of pipe of a certain material. Combining data from pipes and pump stations does not yield any greater insight into the data than would have been obtained by analysing them separately. This highlights another important characteristic of data federation – that the source data share consistent attributes – that is, the separate data sources are representing similar physical characteristics and quantities in the real world.

There are challenges in a generalised federation approach which are the focus of this technical note.

There are four primary challenges with federating data. These are;

1. Service & access: Exposing data in a useful way so that it can be federated
2. Hierarchy: Accounting for different data hierarchies
3. Attributes: Accounting for varying attribution across datasets
4. Values: Accounting for, or normalising, permitted values

Service and access

Federation requires the source data to be discoverable, accessible and dynamic. This means, we know where to find it and if we don't know it exists we have somewhere to go to look to see if it might (a good analogue analogy of this would be the Yellow Pages ®). It means that once we've found it we can access it – it's a format we can use and we can learn lots about the data by looking at it (machine readable is often a term that's used and partly supports this). By dynamic it means that when the source data are updated then this flows-through to the federated environment and consumer of the data.

The challenge of service and access is partly addressed in previous QC work (Quake Centre Proof-of-Concept National Pipe Data Portal Report) and will be a focus of future work. Service and access are challenges which demand a systems-focused response – an enterprise level and authoritative approach for creating an environment which enables widespread service and access.

Hierarchy

Hierarchy within, and between, datasets is simple to understand. It is the simple groupings or categorisation of data. For example, if you consider an online auction site such as TradeMe ® or Ebay® there is a natural grouping of the items for sale. To find a lawnmower for sale you would select the categories of house & garden > garden > tools & equipment > lawnmowers. These groupings are the *hierarchy* associated with lawnmowers.

There are obvious challenges with hierarchy, if we are looking for shelves to buy on the online auction we may find them in the furniture category, office equipment category or garage equipment category. The same (or similar) object may be found in different parts of different hierarchies. The same applies to asset data. If we are looking for 'concrete pipe' we may find these objects with storm water data or waste water data or even maybe in a 'surplus stock' data category. It is not possible to anticipate all possible hierarchies when designing a system which would support comprehensive data federation – like that proposed by the National Data Infrastructure Model. And therefore, it is not practical to create a rule-set or attempt to codify a process to federate across all possible hierarchies. The approach proposed by this technical note to federating data, when the hierarchy in the individual data sources are different, is outlined below. The approach is based on a process which also considers the importance of *normalising attribution* – which is the next key challenge.

Attributes

Physical objects have obvious attributes – attributes are labels used to describe the object. When considering pipes attributes this will include the length of the pipe, material type, diameter, wall thickness, age, depth, location (geo-coordinates) etc. The general challenge associated with attributes when federating data is simply that there is no consistency between the attribute headings which are used by different owners of assets and the corresponding asset data. One owner may provide an attribute heading as 'pipe length' and another will simply use 'length' – federation demands that there is some specification that these attributes are the *same thing*. In this example the correspondence between pipe length and length may seem intuitive but this cannot be assumed. With attribute headings such as 'material type' and 'construction' the correspondence is less clear (these are real examples which are describing the same attributes from different data owners).

The production and adoption of standards is a common way that this mismatch between attributes is remedied across different data sources. The relative merits of standards and their adoption is partly addressed in previous QC work and while the production of standards are a high priority for QC they are only a small part of the federation process. *The standard* is less important than *a standard* – the process of meeting the challenge posed by mismatched attribution is outlined below.

Values

Values are the numbers and text providing the quantities and descriptions of the object. A pipe is 3.0m long. It has a 300mm diameter. It is 10 years old. It's lowest invert level is 2.5m. It is made from High density polyethylene. 3.0, 300, 10, 2.5 and HDPE are the entries in the record for this object – these are the values. The units (m, mm, years, m) will either be included as part of the entry or be described elsewhere (knowing where these are described and how to find them is a challenge for data federation). It is worth noting that some values will be only be permitted from pre-defined lists – pipe material values are an example of this where a data owner may have a list of 10 pre-defined material types to describe the object.

Challenges associated with federating data at a value level include the expectation of value type (a number entered when text is expected or vice versa), values of differing units between data sources (such as pipe length being in meters in one dataset and in millimetres in another) and pre-defined lists of permitted values being different in different datasets (one dataset having 10 material types and another having 15 where there are 6 or 7 unique titles in each dataset and no obvious correspondence – a real example is "CC" in one dataset and "ACM" in another, both referring to the same material type). Federation at a *value level* allows federated analysis to be completed – it's the final step in the federation process.

Meeting the Challenges.

In this section a proposed solution is provided for three of the four challenges of data federation, which are described above. The challenges of 'service and access' is partly addressed in previous QC work (Quake Centre Proof-of-Concept National Pipe Data Portal Report) and will be a focus of future work. As has been discussed 'service and access' challenges demand a systems-focused response – an enterprise level and authoritative approach to creating an environment which enables widespread service and access. Or,

stated another way, there is little that an individual data owner can do to address the 'service and access' challenge – the enabling activities for 'service and access' has to come from an authoritative source which has strong connections with all potential data owners throughout the sector. The remaining challenges, by contrast, can be addressed (at least in part) through enabling tasks and activities which are completely within the control of the individual data owner.

Hierarchy – there is a simple approach to flattening a hierarchy which allows for universal federation of data. This is a two-step process, the first step is to define and agree the real world objects around which we would like to federate data – we call these entities or objects. Once we have defined and agreed what these objects are then the second step involves converting the hierarchy into attribute data for that object. To explain these two steps further:

> **Defining objects:** This is typically done when a data standard is created – but is the simple process of defining for what real world object data is going to be gathered and retained. An example would be a pipe, chamber or pump. Or to use the examples from the introduction, a lawnmower or shelf. These are objects. Once we have defined these then the whole industry or domain knows the object around which data will be collected and federated. This really is a simple process and while edge-cases do exist, where there may be uncertainty whether something is an object or whether it is really a component of a larger object, it is estimated that at least 90% of potential objects pertaining to asset data could be identified with no dispute or discussion (a pipe is a pipe is a pipe).[1]
>
> **Creating attributes from hierarchy:** This is the second step in addressing the hierarchy challenge and involves simply creating attribute data for each object based on the hierarchy data which has been provided. This, in most cases, can be an automated process – where data can be pre-processed prior to federation (or potentially dynamically processes during federation). Figure 1 shows this step in more detail. The approach is simply to take the list of categories which the object belongs to and include these categories as part of the data which describe that object (the object's attributes). In the example of a pipe object the attribute data would include length, material type, diameter etc. By including hierarchy data the attribute data for the pipe would be extended to include 'true' values for that pipe belonging to certain hierarchy categories. A pipe may therefore have a value of 'true' for the attribute 'storm water'. Using the examples from the introduction, a lawnmower would have 'true' values for the attributes *house & garden, garden and tools & equipment*. A shelf may also have 'true' values for the attributes *furniture, office equipment and garage equipment*. Note that in this example the object can hold attribute values for many categories – this is something that is not possible for standard cascading hierarchies. One of the significant benefits of attribute based

[1] QC would like to propose that significant value can be captured through the simple process of defining infrastructure asset objects – this therefore will be a key priority and focus.

categorisation is that hierarchies can be created on-the-fly if required by the user for a particular type of analysis or visualisation.

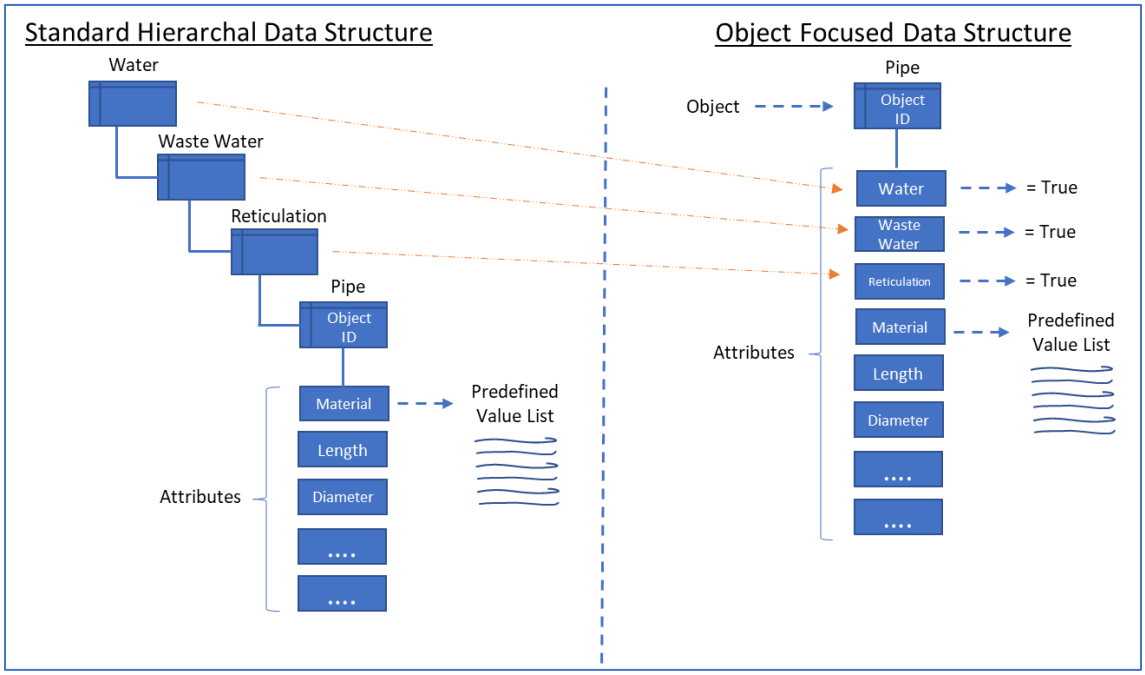

Figure 1: Changing data structure from hierarchal to object focused by creating attributes from hierarchy

Attributes – as discussed above, attributes are what describe the object. The key attribute-related challenge with federating data is that disparate data sources, which relate to the same types of object, will vary in the attributes that they prescribe to that object. This results in two deficiencies or barriers to federation – firstly attributes from different data sources have different names to describe the same thing and secondly, data sources will have different numbers of attributes from each other (one data owner may have 20 attribute for an object and another may only have 10 for the same type of object). We overcome these challenges by agreeing on a standard list of attributes including what we name each attribute (creating of a standard). We then agree what attributes are core (or essential) for each object for the purpose of federation.[2] At an attribute level we also need to consider the pragmatic approach to existing data where object names and attribute names are different from those given in the standard – this is addressed through correspondence tables.

Creating a standard (of objects and attributes): A standard would authoritatively declare 'objects' and the expected attributes for each object including the official name for each attribute. For example, 'pipe' would be declared an object and 'length' would be the official attribute label to describe (obviously) the length of the object. As noted above – rapidly creating a good standard will capture significant benefits and

[2] QC would like to propose that the process of defining authoritative attribute names and establishing core (or essential) attributes could be achieved through a rapid design process.

vastly outweighs the cost of waiting for a perfect standard (*the standard* is less important than *a standard*). In addition, previous QC work has shown that dynamic standards should be anticipated and welcomed – meaning that we can agree a standard and move forward confidently, knowing that the standard can be improved over time as practical implementation identifies potential improvements. Acknowledgement that standards are dynamic provides forward and backward compatibility as data and standards evolve over time. Additionally, the creation of a standard does not demand that data owners modify their data, it simply provides an authoritative reference for 'this is important and expected data' and 'this is what we should name this attribute'. Previous QC work has already shown that a standard is what enables data federation without requiring the data owner to modify their data.

Defining core attributes: Once there is a developed standard which defines objects and their attribute (including the official names of the objects and attributes) the next step is to determine the core set of attributes which are essential for federation – that is the minimum set of attribute data required for the object to be added to the federated environment. This overcomes the challenge that differing data sources have different numbers of attributes – provided there is an agreed set/list of core attributes then federation can be progressed on these. It is worth noting that any non-core attributes which are present in the dataset can also be used in the federation process – but are not mandated.

Authoritative Correspondence Tables: A standard provides clarity on what objects and attributes should be present in the data and what they should be called.[3] Defining a core set of essential attributes ensures that the federation process can anticipate a minimum level of attribute data and therefore a pre-determined level of value-capture as a result of the federation process. The creation of *authoritative correspondence tables* ensures that each individual source data set can be 'mapped' to the agreed standard – and therefore federated. As an example, if one data set includes a pipe attribute of 'construction type' where the standard anticipates that the attribute name will be 'material type' then correspondence table can be used as a reference (like a thesaurus for objects, attributes and value names) to inform the federation process that *this* attribute is actually the same as *that* attribute. Correspondence tables are closely linked to data and schema mapping processes (which are not described in this technical note) and are compiled and maintained by the authority who is responsible for service and access to the federation environment.

[3] Semantic data structures are the core of the modern web environment and are ubiquitous to the large platforms which dominate online activities. These are commonly referred to as graph data structures. Applying semantic data structure principles to infrastructure data is new but is increasingly compelling due to the desire to federate data – the web can be considered one giant data federation environment (enables through the http protocol). Data federation has never been demanded from infrastructure data previously – hence why old data structures have prevailed.

Values – The challenges to data federation at a values level are almost identical to those for attributes. A federation process will anticipate that the value entered for each object under a given attribute label will have certain properties and formats. An example would be that an anticipated value would be free text or a decimal number with units of meters. Alternatively, it could be a value from a pre-defined (formatted) list. We cannot meaningfully interrogate a federated dataset if there are not standardised values. To overcome this challenge, we consider three responses; firstly, by referencing a standard (including predefined lists), secondly, introducing the concept of function-mapping and lastly the process of manual validation.

> **Referencing a standard (including pre-defined lists of values):** As for objects and attributes, values require a standard so that there is an anticipated authoritative description of the format and properties for values under each attribute.
>
> **Function-mapping:** For the federation of a data from a wide range of data owners it is anticipated that there will be a low degree of alignment between properties of attribute and values of the source data and either the standard or other data sources. For attributes the use of correspondence tables partly addresses this challenge. For values, *function-mapping* is required. This is a process which is enabled during the federation process by the data owner or the agent undertaking federation. Function-mapping allows for standard operations (such as formatting, unit conversions, naming standardisation etc) to be applied to entire lists of values for given attributes. For example, a function could be implemented to convert millimetres to meters or to change a 'material type' value from ACM to AC. Function-mapping would be enabled by special tools which are part of the federation platform or system.
>
> **Manual Validation:** Much of the federation process outlined in this technical note (and previous QC work) can be automated through smart processes and tools. However, manual validation remains a key component of the federation process and particularly at the value level. Manual validation would be specifically required to ensure that the mapping of object names and attribute names has been successfully applied (especially if an automated correspondence look-up process is employed). Manual validation is particularly important at the value level as it ensures that values are federated in a like-for-like manner across all data sources. The process of functional-mapping requires manual intervention as this is difficult to automate.

Note on Null Values (missing data) – A null value means that a value does not exist in the record for a given attribute. Measuring and identifying null values is very important as it provides insight into the data collection effort that is required to improve the quality of the dataset. Null values are different from zero values – for example, an attribute may record the number of laterals associated with a pipe object. A value of zero indicates that there are no laterals attached to the pipe. A null value means that we don't know how many laterals are connected to the pipe – it could be many or it could be none.

Note on Metadata – Metadata is data which describes the information, qualities and attributes of the dataset as a whole (as compared to the data within the dataset). This includes information such as the owner organisation, the date the dataset was last updated, the file format, possibly the number of records, comments, version and version history etc.

Metadata is an essential source of information in the federation process.
This technical note has not attempted to define the approach to defining a metadata standard that best supports federation of infrastructure asset data – previous work by QC on the proof-of-concept for National Pipe Data Portal highlighted that the 'unpacking' of data for a data federation process has a strong service (human) component. This service component becomes the process through which the metadata is identified and defined. While a metadata standard, that can be universally adopted be all data owners for their datasets, is desirable the pragmatic reality is that this will be much more challenging to implement than the other data-preparation steps outlined in this technical note. It is worth noting that defining and (practically) implementing metadata standards is worthy of further investigation as the baseline of federated data expands across the sector. Metadata is very important but it's absence is not necessarily a barrier to progressing data federation.

Level of detail – this technical note has presented limited detail on the concept of 'level of detail'. This is a concept which is common in the vertical build sector particularly in regard to BIM models and design. 'Level of detail' is an expression used to define ascending categories of detail. For example, for water infrastructure it could be classified that a level of detail equal to 1 simply describes the alignment and location of pipes and pump stations (and nothing more) whereas a level of detail equal to 5 may describe the location of characteristics of the smallest components of the water system such as pump seals, bearings, motor casings, weld types etc. It is not intended that this concept is explored further as part of this technical note however, it is worth considering that the value-capture of data federation is maximised at relatively coarse levels of detail. The implication of this for this technical note is that the most significant advances can be made, and the most gains in value-capture, through the rapid development of simple standards and the agreement of core (essential) attributes. It is anticipated that the focus of standards should be at level of detail 1 or 2 as this would capture the most value in the shortest space of time.

Where to next – This technical note, together with the previous QC publication (Quake Centre Proof-of-Concept National Pipe Data Portal Report) concludes the data federation research that commenced in 2018. The challenges for data federation are well understood and, as evidenced by the content of this note, have practical solutions which can be readily implemented. There are clear easy-wins which would facilitate meaningful data federation around key infrastructure assets. These include:

- Defining standards – with the focus on naming objects, attributes as well as defining value level properties and formats.
- Determining the core (essential) attributes for key infrastructure objects.
  These two steps could be achieved quickly allowing wider federation activities to flow from this initial enabling work.

Note prepared by: Angus Bargh, Open Plan Ltd

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