Manaaki Whenua
Landcare Research

A proof of concept (PoC) multi-indicator environmental data infrastructure

A MBIE Innovative Data Analysis programme white paper

Prepared for: Ministry for Business, Innovation and Employment

November 2018

A proof of concept (PoC) multi-indicator environmental data infrastructure

A MBIE Innovative Data Analysis programme white paper

Contract Report: LC3396

Alistair Ritchie, Juliana Osorio-Jaramillo, David Medyckyj-Scott, Michael Blaschek

Manaaki Whenua – Landcare Research

Reviewed by:	*Approved for release by:*
Anne-Gaelle Ausseil IDA Programme Leader Manaaki Whenua – Landcare Research	Sam Carrick Portfolio Leader – Characterising Land Resources Manaaki Whenua – Landcare Research

Disclaimer

This report has been prepared by Manaaki Whenua – Landcare Research for the Ministry of Business, Innovation and Employment (MBIE). If used by other parties, no warranty or representation is given as to its accuracy and no liability is accepted for loss or damage arising directly or indirectly from reliance on the information in it.

Contents

Glossary

ALA: Atlas of Living Australia, Australia's national biodiversity database.

API: application programme interface, a set of functions and procedures that allows the creation of applications that access the features or data of an operating system, application or other service. When used in the context of the internet, it often refers to the way a client and server interact in a standardised way. It is also called a web service, although the two are not equivalent.

Application schema: a conceptual schema for data required by one or more applications. A GML application schema is written in XML schema (ISO 19136:2007).

DwC-A: Darwin Core Archive (DwC-A) is a biodiversity informatics data standard that makes use of the Darwin Core terms to produce a single, self-contained dataset for species occurrence or taxonomic (species) data. It is the preferred format for publishing data to the Global Biodiversity Information Facility.

ELFIE: Environmental Linked Features Interoperability Experiment, a multi-agency Open Geospatial Consortium interoperability experiment undertaken in 2018.

EML: Ecological Metadata Language (EML), a metadata specification developed for the ecology discipline.

GBIF: Global Biodiversity Information Facility, an international organisation that focuses on making scientific data on biodiversity available via the internet using web services. New Zealand is a signatory and is responsible for a national GBIF node (which currently does not formally exist).

IDA: The MBIE Innovative Data Analysis Programme

LAWA: Land, Air, Water Aotearoa (LAWA), an environmental data and information repository, initially a collaboration between New Zealand’s 16 regional councils and unitary authorities.

Linked data: the use of the HTTP protocol for accessing, updating, creating and deleting resources from servers that expose their resources according to the following rules of linked data: use URIs as names for things; use HTTP uniform resource identifiers (URIs) so that people can look up those names; when someone looks up a URI, provide useful information using the standards (RDF, SPARQL); include links to other URIs so that people can discover more things.[1]

LRIS portal: the Land Resource Information Systems portal, Manaaki Whenua’s data repository of New Zealand land-related science data sets and information.

[1] https://www.w3.org/TR/ldp/

LRIS programme: the Land Resource Information Systems programme, a Manaaki Whenua – Landcare Research programme funded by MBIE through its Strategic Science Investment Fund Infrastructure Platform. The programme hosts and manages two of the 25 New Zealand Nationally Significant Collections and Databases (NSCDs) and a large number of related data sets.

LUC: The New Zealand Land Use Capability classification system.

LUNZ: Land Use of New Zealand dataset

MBIE: Ministry of Business, Innovation and Employment.

MfE: Ministry for the Environment.

NSDR: the National Soils Data Repository, a component of LRIS, is a versatile observation database that hosts the original National Soil Database and other full and partial soil profile descriptions, with associated laboratory analyses. Because of intellectual property and data privacy issues with respect to the soil data collected on private land (i.e. the data have commercial value for the land owner), many of the data cannot be released to the public.

NVS: the New Zealand National Vegetation Survey Databank

Observation: the act of measuring or otherwise determining the value of a property (ISO 19156:2011).

OGC: Open Geospatial Consortium, an international industry consortium of over 521 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards. OGC works closely with the International Organisation for Standardisation (ISO).

OWL: Web Ontology Language, a family of knowledge representation languages for authoring ontologies.

Profile: a subset of an application schema.

Property: a facet or attribute of an object referenced by a name (ISO 19156:2011).

PROV: a W3C standard that defines a data model, serialisations, and definitions to support the interchange of provenance information on the Web.

QGIS: open-source cross-platform GIS desktop software, formerly known as Quantum GIS.

RDF: Resource Description Framework, a standard model for data interchange on the Web.

RTF: Rich Text Format, a text file format used by Microsoft products such as Word and Office.

Resource: items of interest on the World Wide Web that are identified by global identifiers called uniform resource identifiers.

ReST: Representational State Transfer, an architectural style that defines how to deploy web services within the constraints of the HTTP protocol.

Sampling feature: a real-world feature, such as a station, transect, section or specimen, which is involved in making observations of an environmental feature.

SKOS: Simple Knowledge Organisation System, a W3C recommendation designed for representing thesauri, classification schemes, taxonomies, subject heading systems, or any other type of structured controlled vocabulary (source: Wikipedia).

URI: uniform resource identifier, a string that uniquely identifies a resource on the internet.

Web service: a service offered by a device (e.g. a web server) to another device (e.g. a web browser). In the context of this document, services provided include digital descriptions of a resource, and search facilities to help discover one or more resources.

WFS: Web Feature Service, a spatially enabled web service that supports requests for data about a real-world feature or resource (ISO 19142:2010).

WMS: Web Map Service, a spatially enabled web service that delivers pre-compiled maps or map layers to a client (ISO 19128:2005).

XSLT: XML Stylesheet Transformation Language, a language for transforming XML documents into other XML documents or formats (e.g. JSON or HTML).

1 Introduction

The Innovative Data Analysis (IDA) programme was an MBIE-funded research programme led by Manaaki Whenua – Landcare Research (MWLR) that ran for 4 years (2014–2018).

The aim of the programme was to research and develop processes to integrate and harmonise high-priority heterogeneous land resource and biodiversity data sets to support a step-change in the quality of environmental reporting. The aim was to support central and regional government to report on the state of the New Zealand environment in a standardised, statistically robust and transparent way.

The programme was aligned with key initiatives such as State of the Environment reporting, Environmental Monitoring and Reporting, and the National Science Challenges. It used next-generation data analysis techniques and worked with data custodians and end-users to develop statistical indicators for soil health, land use, and species occupancy. The programme focused on extracting knowledge and value from existing environmental data sets. Its focus was not research outcomes, but technical and social infrastructure outcomes.

The know-how and tools developed by the IDA programme were scientifically and technically significantly ahead of what others were doing in New Zealand. It is important, therefore, that what was learnt be shared with others, and this has been done through a series of white papers and presentations. This document summarises key aspects of the programme at a more technical level with respect to the implementation of a proof of concept (PoC) multi-indicator environmental data infrastructure. Following international standards and best technical practice, this PoC was used to generate indicators associating land-use and soil-quality databases. We argue that this PoC provides a design pattern that can be extended and reused for publishing and sharing other environmental indicators.

Also, we will describe the importance of the social aspects that can underpin or threaten the future implementation and use of such an infrastructure. In doing so we seek to raise awareness that, while implementing technology for innovation can be challenging, the social aspects of complex information systems need to be fully considered and tackled if we want to ensure the successful establishment and operation of an environmental data infrastructure.

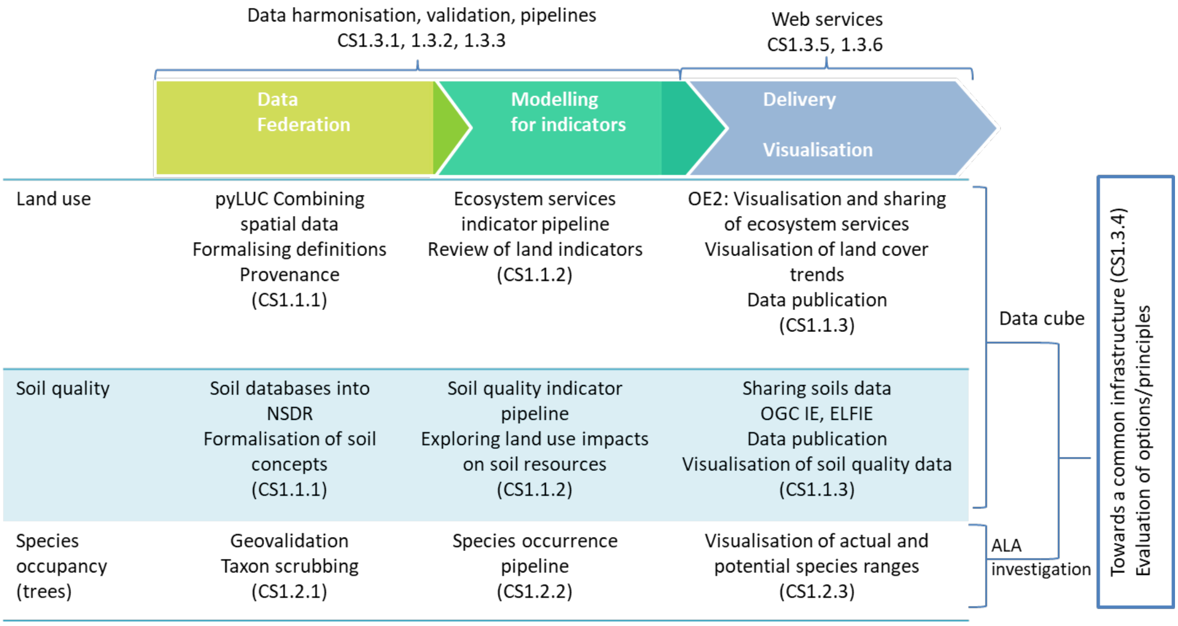
2 Background

The IDA programme had three domain components: land use, soil quality, and species occupancy (Figure 1). For each of these domains the key components of the programme included:

- data federation – bringing together heterogeneous spatial data from multiple sources to produce a suite of higher-value information products
- modelling indicators – developing indicators to respond to the pressure, state, and impacts framework for environmental reporting

- visualisation and delivery – enabling use by central and regional government and other agencies, via existing portals (e.g. the LRIS Portal, MWLR's national land information data portal) and via open standards-based web data services.

Underpinning these components, characterisation of provenance, quality, uncertainties, security, and workflows of the data and information was also explored to enable an auditable process.


Data harmonisation, validation, pipelines
CS1.3.1, 1.3.2, 1.3.3
Web services
CS1.3.5, 1.3.6
Data cube
ALA investigation
Towards a common infrastructure (CS1.3.4)
Evaluation of options/principles

	Data Federation	Modelling for indicators	Delivery Visualisation
Land use	pyLUC Combining spatial data Formalising definitions Provenance (CS1.1.1)	Ecosystem services indicator pipeline Review of land indicators (CS1.1.2)	OE2: Visualisation and sharing of ecosystem services Visualisation of land cover trends Data publication (CS1.1.3)
Soil quality	Soil databases into NSDR Formalisation of soil concepts (CS1.1.1)	Soil quality indicator pipeline Exploring land use impacts on soil resources (CS1.1.2)	Sharing soils data OGC IE, ELFIE Data publication Visualisation of soil quality data (CS1.1.3)
Species occupancy (trees)	Geovalidation Taxon scrubbing (CS1.2.1)	Species occurrence pipeline (CS1.2.2)	Visualisation of actual and potential species ranges (CS1.2.3)

Figure 1. The components of the IDA programme
CS = critical steps. For other abbreviations see Glossary.

2.1 Purpose of this document

The aim of the present document is to explain the technical and social aspects of some of the different technical components, models, systems and infrastructure designed and developed by MWLR in the context of the IDA programme[2].

A brief report of all the outcomes generated by the IDA programme was provided to MBIE in June 2018[3] for the review panel of the MBIE Targeted Appraisal of the IDA programme. The objective of the report was to provide the context and scope of all the deliverables built by MWLR without going into technical details.

[2] The reports, papers and some of the software developed by the programme will be available on the MWLR website early in 2019.

[3] *Innovative Data Analysis programme 2014–2018 – Brief report for the review panel*

One of the conclusions of the MBIE targeted appraisal was that the technical know-how and tools developed by the IDA programme were scientifically significantly ahead of what the end-users' own organisations were currently using, and that they looked to the IDA team to provide direction for their current and future work. One of the ways to do this would be to summarise key aspects of the programme in a more technical document.

The present document covers technical documentation of the single domain infrastructures (land use, soil quality and species occupancy) designed and implemented in the context of the IDA programme. We then use the knowledge learnt during the programme to describe the technical details of the implementation of a multi-indicator environmental data infrastructure. This is a proof of concept (PoC) that can be extended and reused to generate indicators associating land-use and soil-quality databases in line with international standards.

We will also describe the importance of the social aspects that can underpin or threaten the future implementation and use of such an infrastructure. We seek to raise awareness of the fact that although implementing technology for innovation can be challenging, the social aspects of complex information systems (the ones needed to describe, understand and manage our environment) need to be fully considered and tackled if we want to ensure the proper implementation of an infrastructure, its correct usage, and its sustainability and future evolution.

3 Domain-specific infrastructure

MWLR is the kaitiaki of seven of the 25 New Zealand Nationally Significant Collections and Databases[4] (NSCDs). The databases contain critical georeferenced and observational data for the biodiversity, land and soils domains. The data are kept in different repositories and have been supported by different computer infrastructures since the mid-1990s, when the NSCDs were transferred from the likes of DSIR, the Ministry of Works, and Forest Services to their successors, the Crown Research Institutes.

This single domain approach helps reduce the complexity of visualising, analysing and modelling our environment. Just 10 years ago processing and modelling using big amounts of data was challenging. Considering the inter-relationships among the different domains was also problematic. As a result, the evolution of data infrastructures and information systems has been led by the single domain view. This approach is not incorrect or unimportant, indeed it is the cornerstone for multi-dimensional, big data delivery and analysis.

In recognition of this status, the IDA programme has contributed to enhancing the capability and robustness of the single domain data and systems infrastructure for data delivery, collaboration improvement and facilitation of analysis and modelling.

[4] http://natsigdc.landcareresearch.co.nz/natsigdc_list.html

Although for each domain the infrastructure and standards used varied depending on national and international objectives and collaboration requirements, there is consistent agreement of the importance of standardised data services as one of the most versatile and widely used means of data sharing.

3.1 Biodiversity domain

In the biodiversity domain MWLR maintains, operates and develops six databases: the Allan Herbarium, the International Collection of Micro-organisms, the National New Zealand Flax Collection, the New Zealand Arthropod Collection, the New Zealand Fungal and Plant Disease Collection, and the National Vegetation Survey (NVS). All these databases, except NVS, are part of the Systematics Collection Data and can be accessed by the public via their respective websites[5]. It is possible to perform data searches in the five databases at the same time and retrieve the matching results. The individual results can be exported as Comma Separated Values (CSV) files. For the NVS databank, data can be queried and requested from its specific website[6].

Although these databases have been providing data to a wide variety of end-users for 5 years (mainly via the websites), a common access point to serve data that follows international standards was not available until last year. Figure 2 shows how the biodiversity data are served from the five different databases.

The infrastructure was designed and implemented to deliver data using the most widely accepted international data standard for biodiversity, the Darwin Core Archive (DwC-A).[7] This is the standard for the Global Biodiversity Information Facility (GBIF)[8]. The node developed by MWLR is a GBIF node, which provides data to the New Zealand Virtual Herbarium. The node hosts biodiversity data provided by different parties as well (Te Papa and Auckland Museum, Lincoln University, and MWLR, among others). The New Zealand Virtual Herbarium delivers data to the Australian Virtual Herbarium, which is part of the Atlas for Living Australia.

As seen in Figure 2, the infrastructure deployed by MWLR provides data using an integrated publishing toolkit[9], which is a fundamental tool in the GBIF ecosystem. This access point archives and serves the information as downloadable XML files. Metadata can also be downloaded from the same access point[10] in EML and RTF formats. The files are updated on a weekly basis. The data at the MWLR integrated publishing toolkit point are publicly available and follow the specimen occurrence specification of the DwC-A.

[5] https://scd.landcareresearch.co.nz

[6] https://nvs.landcareresearch.co.nz

[7] http://rs.tdwg.org/dwc/terms/

[8] https://www.gbif.org/

[9] https://www.gbif.org/ipt

[10] http://ipt.landcareresearch.co.nz/

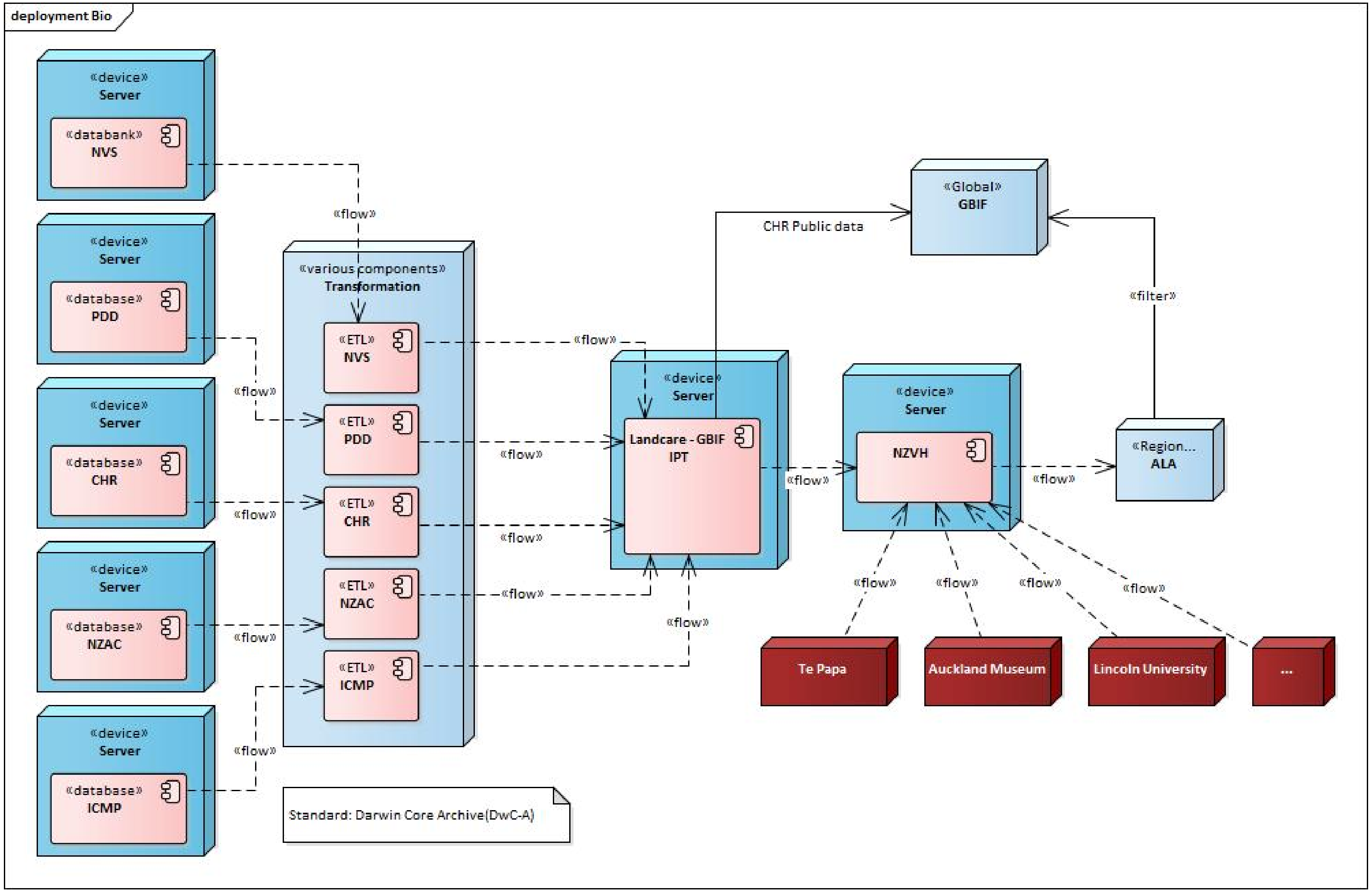


Figure 2. Biodiversity data infrastructure.

NVS = National Vegetation Survey; PDD = New Zealand Fungal and Plant Disease Collection; CHR = Allan Herbarium; NZAC = New Zealand Arthropod Collection; ICMP = International Collection of Micro-organisms; ETL = extract, transform, load; NZVH = New Zealand Virtual Herbarium; ALA = Atlas of Living Australia; GBIF = Global Biodiversity Information Facility.

3.2 Soil domain – site observations

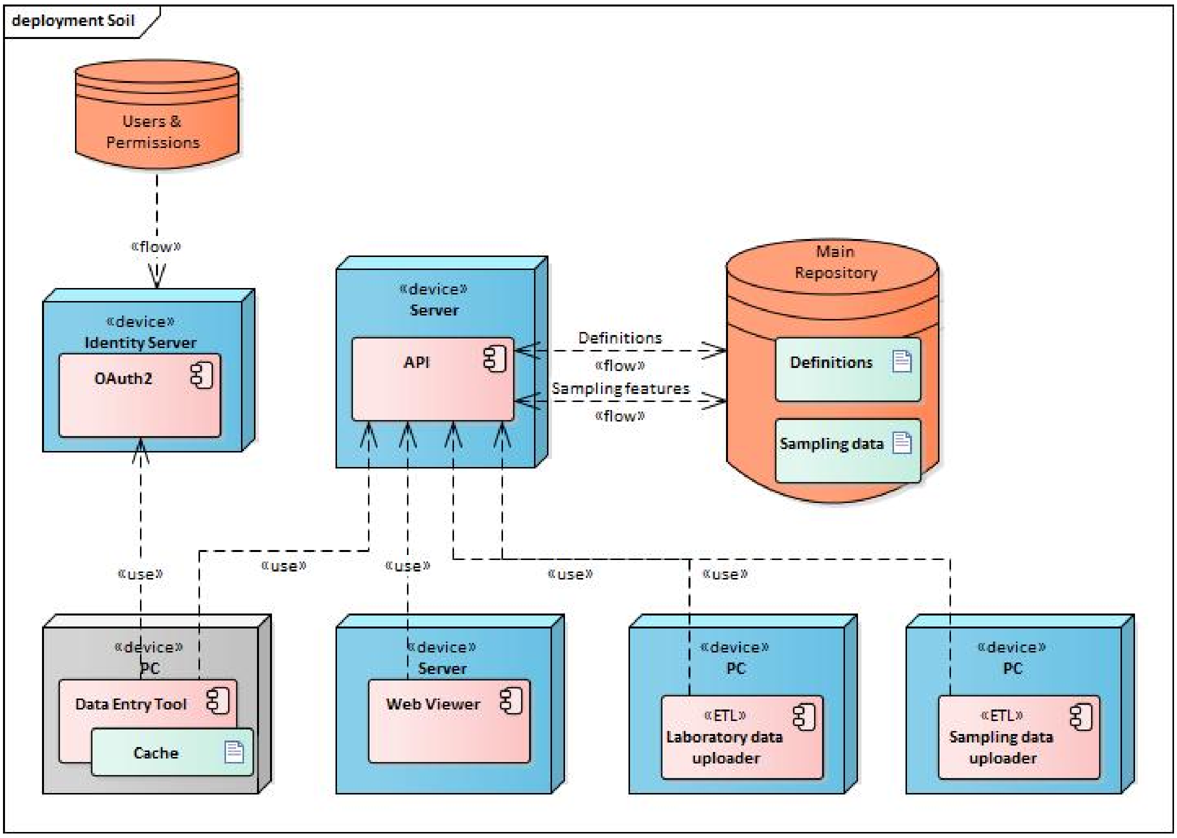


Figure 3. Soil data infrastructure.
OAuth2 = OAuth is an open standard for access delegation for websites and internet applications; API = Application Program Interface; ETL = extract, transform, load.

The soil data infrastructure, funded through the MBIE Strategic Science Investment Fund Infrastructure Platform, was implemented to archive, secure and provide access to soil point data of New Zealand. The design of the data model follows the principles of the observations and measurements conceptual model, which is one of the core standards in the Open Geospatial Consortium (OGC). MWLR works closely with OGC, helping with the definition of standards and the implementation of experiments to prove the real applicability of such standards, and to understand their limitations and evolve their development.

A diagram of the implemented infrastructure is shown in Figure 3. The design of the infrastructure is aligned with the following concepts:

- use new technologies that facilitate changes in the data model and the end-user requirements
- provide information in different formats and through different access points:
  - web portal
  - services/application program interface (API)
- keep data secured, with private or public access depending on specific contractual conditions or intellectual property

- allow the entry of new data using specific tools that validate and ensure the quality of the information
- facilitate the upload of historical and laboratory data.

The information flows from the data entry points (data entry tool, laboratory data uploader, and sampling data uploader) to the main repository via the API, and it is recovered from the data repository to be displayed in the web viewer and the data entry tool, again using the API. All the clients (points connected to the API) are secured and use OAuth2[11] as the protocol for authentication and authorisation.

The infrastructure provides access to soil point data for MWLR personnel and external users with private and public admission via the web viewer[12]. Only authorised access to the API is allowed. This means only the clients that are registered and configured with the security system can obtain and push data through the API. Hence new services that expose data to the general public must be developed if required.

In section 4, 'Multi-indicator infrastructure', we describe the set of services that were implemented in the IDA programme to allow the construction of multi-domain indicators. Some of these services take soil quality data from the soil domain infrastructure (Figure 3) and expose it using international standards (ANZSoilML) and widely used protocols (JSON, WMS), and allow the soil data to be merged with land-use information to facilitate more complete analysis and data visualisations (see Figure 4, below).

3.3 Land-use domain

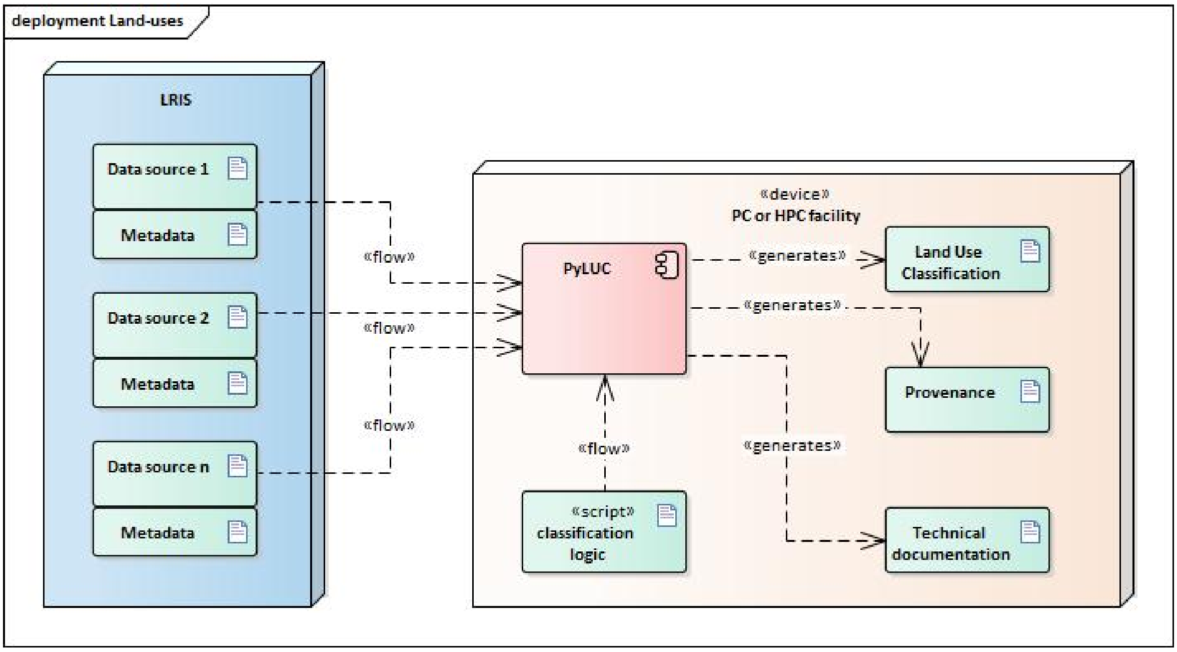


Figure 4. Land-use data infrastructure.

[11] https://oauth.net/2/

[12] https://viewer-nsdr.landcareresearch.co.nz/

Land use can be defined as the activities or socio-economic functions for which land is used. It differs from land cover, which describes the physical state of the land (Lesslie 2004). Land-use data provide information on the function and purpose for which land is currently used, and, when tracked over time, how land use changes (Young 1998). Land-use classification provides a framework to guide the collection of data and the creation of effective databases to ensure comparability and compatibility (Gong et al. 2009).

PyLUC is a framework written using the Python language and developed to help generate land-use classifications (LUCs) with self-documenting LUC definitions that could be processed to create both the spatial data set (the LUC) and supporting documentation, including provenance (Spiekermann et al. submitted). An essential output of pyLUC is the automated documentation and provenance information generated as part of a classification run (see Figure 4).

In the first version of pyLUC, input data sources are restricted to data hosted on the LRIS portal[13] (but could be expanded to any instances of portals provided by the Koordinates geospatial data platform, of which LRIS is one instance). This was done for three main reasons:

- many of New Zealand's authoritative geospatial data sources exist on this platform
- to encourage the use of original primary sources, with any modifications transparently 'baked into' the LUC definition
- because data layers stored on this platform are immutable, which ensures all referenced data sets will exist in their original form at any point in the future.

The last two reasons significantly contribute to the transparency and repeatability of the LUC process. This means it is possible for a user to recreate an LUC with essential technical documentation and provenance data given a pyLUC definition script (single text file) and an LRIS Portal account with appropriate permissions to access protected data sources.

When pyLUC ingests a definition script it must initialise an internal model of the algorithm(s) involved in creating an LUC to produce the classification data set, technical documentation and data provenance. The LUC output is a thematic raster, which consists of a raster attribute table (RAT) that can be vectorised to create a vector output.

Provenance is delivered by pyLUC as a PROV-N file (standard notation for W3C PROV model[14]), which, while not easily human-readable, can be ingested by other tools for analysis and visualisation.

The technical documentation is a human-readable Latex file, which is intended to be used as an annex in reports. It has information on raw code for each of the classification rules and tables showing the association between people and organisations, the list of inputs

[13] https://lris.scinfo.org.nz

[14] https://www.w3.org/TR/prov-overview/

(each data source from LRIS), the fields that were pulled from such data sources, and people attribution. Also, the documentation has a table that shows which rules are linked to which outputs, inputs and authors.

PyLUC was designed to be used in a high-performance computing environment or on a PC. This means an end-user on their PC can run the program and get the outputs. However, if the rules are complex and the estimated processing time is long, the user must use the high-performance computing facilities to accelerate the process.

3.4 Harmonising data

Harmonisation of data in each separate domain proved to be challenging, with different levels of difficulty depending on the structures, frameworks and institutions available per stream.

Biodiversity: The biodiversity domain is the most mature of the domains we analysed for this paper. This maturity can be defined in terms of having a central institution (GBIF), coordinated through its secretariat in Copenhagen, which acts as the central authority and repository; the adoption of specific international standards (DwC-A); and the provision of numerous tools for data validation and publication (integrated publishing toolkit, data validator and name parser, among others[15]). This ecosystem of tools, institutions and standards facilitated the implementation of new nodes in the context of the IDA programme (see Figure 2).

Similarly, the infrastructure that MWLR has provided for the custodianship of the nationally significant databases in the biodiversity domain facilitated the harmonisation of the data. Having these structured databases eased the implementation of transformations to publish the data in the required formats for the Atlas of Living Australia and GBIF.

Soil quality: The harmonisation of soil data had extra challenges compared with that for the biodiversity data. In the soils domain there are multiple data owners with their own repositories, with data in a variety of formats and technologies. Some of repositories are not even structured databases, but spreadsheets with different data structures to accommodate the requirements of each organisation or a survey.

MWLR is the kaitiaki of the National Soils Database (NSD), one of New Zealand's NSCDs. It works to collate soil data and related information from different agencies to store in the National Soils Data Repository (NSDR). This has been, and will continue to be, problematic if aspects such as a centralised site naming system, standardised methods, and the use of international data exchange standards for publishing data are not in place. (For more details on the harmonisation of soil data, see section 3.5).

Currently, the NSDR infrastructure (see Figure 3) complies with international standards that are under development and still evolving. It provides data collection tools that

[15] https://www.gbif.org/resource/search?contentType=tool

facilitate the harmonisation of data and methods. However, this infrastructure currently only provides services for data collection, management, and data access for use by MWLR employees. The challenges to collate data that are not generated by MWLR will remain if agreements and best practices are not put in place for national ongoing management of soils data (for more details on the social aspects of data infrastructures, see section 5).

Land use: Collation and harmonisation of data to generate land-use information required investigation of the available data sources, their quality, and their criteria for defining and categorising land in New Zealand (Manderson et al. 2017). The data sources were required to be raster or vector data, as these were the type of input expected by pyLUC (see section 3.3).

For the land-use domain the objective focused on guaranteeing replicability of the land-use models using a tool that could be fed by standard raster data. Finding data sources that meet this criterion was not challenging. However, understanding the set of rules and analysis behind some of the land definitions in the data sources and how the land-use models generated their data outputs required more analysis and investigation (Manderson et al. 2017).

3.5 Soil quality data – a case study of data harmonisation challenges

Soil quality data were chosen for a case study of data harmonisation challenges because they expose the range of needs, conditions and issues faced when aggregating data for analysis, as follows.

- They are a fundamental data set for State of the Environment reporting.
- They are collected, stored and maintained by a disparate set of agencies for both data management and analytical reasons.
- While stored and maintained separately, they are functionally a single, logical data set, with a need for consistent management this implies.
- There is no history of coordinated, nationally consistent, capture and management of data, but a widespread recognition of the need to do so.
- The basic set of technology and processes required to implement the case study should be appropriate to other environmental domains.

The Resource Management Act 1991 requires regional councils to monitor and report on the state of the environment in their regions to help understand and manage human impact on the environment. Soil quality (or health) is one of the environmental performance indicators used in State of the Environment reporting in 2018.

Presently, soil quality monitoring in New Zealand is undertaken according to guidelines published by the Land Monitoring Forum (2009). These were the result of lessons from monitoring trials that began in 1955, culminating in the MWLR '500 Soils' project (1999–2001) (Sparling et al. 2004), and subsequent Land Monitoring Forum and regional council sampling. At present more than 1,000 sites are monitored by 13 of New Zealand's 16 regional and unitary councils (Cavanagh et al. 2017).

The fundamental soil-related properties measured at the sites focus on dynamic characteristics of the soil that describe the changing state of some of a soil's biological, chemical and physical components. Most properties are measured using equipment in soil chemistry or physics laboratories, but some characteristics (New Zealand Soil Classification (Hewitt 2010) soil order, land use) are established on-site or by reference to previous mapping. This limited, pragmatic set of laboratory and field measurements and sample metadata make for a simple data set to store and deliver.

Unfortunately, the nature of this storage varies. A survey of councils conducted by MWLR for an Envirolink Advice Grant (Cavanagh et al. 2017) found that the majority stored the monitoring data as spreadsheets (in one case in a document management system). A minority of councils used a database (the council system or Microsoft Access) or the Hilltop Software[16] system. Beyond the standardised structured used by the laboratories (MWLR and Hills), there was no consistent data structure or content model (shared vocabularies). Over time, copies of these data have been provided to MWLR as spreadsheets as source data for the generation of soil quality indicators. Numerous conflicting copies of these now exist, either as spreadsheets or in a Microsoft SQL Server database.

To understand what would be involved in the early collection stages of a soil quality data pipeline, the spreadsheets provided to MWLR were used as proxies for what was likely to currently be provided by councils. They were then imported into MWLR's National Soil Data Repository (NSDR, Figure 3).

While MWLR has succeeded in loading the observations into the NSDR, as a test of an automated process to import soil quality observation data, overall the process failed. The data as currently provided were not suitable for integration into a data capture pipeline: far too much human intervention, with associated costs, was required to consolidate the data into a single data set.

The data issues faced included:

- no single, authoritative source of monitoring data
- multiple copies of the same measurements held in different files (created for use in analysis or to correct errors in earlier sources)
- no globally unique identifiers for sites where data were collected and for samples to allow linking across laboratory data or time (site revisits)[17]
- undeclared changes of units of measure for analyses, often in the same column of a single spreadsheet
- missing, or ambiguous, laboratory method metadata

[16] http://www.hilltop.co.nz/

[17] In some cases, a weak form of identifier was used, which, while easily recognised by a human reader, could not be matched across spreadsheets due to inconsistent use of leading zeros, spaces, decimal points, dashes or underscores within the identifier string.

- missing site locations.

Put simply, there was much work to be done to get these sets of data into a state that supported interoperable exchange of data between monitoring agencies, research institutes and central government.

The Land Monitoring Forum's recommendations themselves note that a nationally consistent system is required. They recommended the establishment of

> a nationally centralised data management and electronic storage facility, with internet access available to interested parties [...] The exact location of the facility is irrelevant, but clearly security, long-term storage and accessibility need to be guaranteed. This could be a central government function, or delegated to a regional council, CRI or other suitable organization (LMF 2009 p51).

Since then regional councils have advocated a distributed network of databases exposed by standards-based web services such as those that support the Land, Air, Water Aotearoa (LAWA) website[18]. Any solution is likely to combine aspects of both centralised and distributed databases, and the following section describes a PoC multi-indicator environmental data infrastructure that tests data standards and web service protocols that would support such an environment. These standards and protocols can inform and constrain the capture, management, and delivery of soil-quality (and other environmental indicators) monitoring data.

4 Multi-indicator infrastructure

4.1 Introduction

State of Environment indicators describe important dimensions of the Earth's biological and physical environment – air, marine, fresh water, land, biodiversity, atmosphere and climate. They cannot be viewed in isolation because they reflect the interplay of these environmental systems and human activity. The data and information used or produced in their generation must therefore support multiple indicators and data from diverse domains and communities (e.g. soil, land-use and biodiversity data generated by researchers, consultants and government agencies). Logically, a unifying data infrastructure must also cater to these in an integrated and consistent way.

No such data infrastructure exists for New Zealand, so the IDA programme undertook to define one and test its viability with a simple PoC. Building a whole-of-environment system is an ambitious task, so the scope was tightly constrained to soil and land-use data. Further work must expand the scope to the other domains.

[18] https://www.lawa.org.nz/

The need for an environmental data infrastructure is not unique to State of Environment reporting, nor to New Zealand. Many members of the Open Geospatial Organisation (OGC), including MWLR, have similar motivations, so during the IDA programme the IDA team joined two OGC interoperability experiments that addressed necessary data standards and protocols. This provided an opportunity to work with international peers to experiment and gain experience to help understand what the technical design of a multi-indicator, multi-partner infrastructure might look like and the challenges faced in implementing such an infrastructure.

The Soil Data Interoperability Experiment (SoilIE[19]) initiated by the IDA programme.

> This was conducted under the auspices of the OGC Agriculture Domain Working Group in 2015. Soil data exchange and analysis is compromised by the lack of a widely agreed international standard for the exchange of data describing soils and the sampling and analytical activities relating to them. Previous modelling activities in Europe and Australasia have not yielded models that satisfy many of the data needs of global soil scientists, data custodians and users. This IE evaluated existing models and proposed a common core model, including a GML/XML schema, which was tested through the deployment of OGC web services and demonstration clients. (Ritchie et al 2016)

The Environmental Linked Features Interoperability Experiment (ELFIE[20]).

> This interoperability experiment explored OGC and W3C standards with the goal of establishing a best practice for exposing cross-domain links between environmental domain and sampling features. The experiment focused on encoding relationships between cross-domain features and linking available observations data to sampled domain features. '[The] approach [leveraged] the OGC service baseline, W3C data on the web best practices, and JSON-LD contexts…' (Blodgett et al, in prep.)

The infrastructure described in this document applies the findings of both experiments to create a standards-based system. **Both OGC interoperability experiment engineering reports, and their definitions of terms and normative references, are essential companion reading to this document**.

4.2 OGC collaboration

The IDA team chose to align infrastructure development work with the activities of the OGC for the following reasons.

1. The OGC provides the most advanced and capable standards for spatiotemporal data types and associated web services.

[19] https://github.com/opengeospatial/SoilDataIE/

[20] https://github.com/opengeospatial/ELFIE/

2 It provides clear mechanisms for the creation, governance and maintenance of data standards.

3 It has defined the Observation and Measurements specification for earth observation and sampling data (ISO 19156:2011[21]).

4 It hosts a variety of domain working groups that work on or towards mature and relevant specifications for environment-related domains: hydrology, geology, climate and agriculture data, which include soil data.

5 These working groups have a common membership and are motivated to work to harmonise the earth science data models.

6 MWLR is a long-standing member, with effective relationships with all the key member environmental data agencies.

The SoilIE built on research MWLR had done with CSIRO on a pan-New Zealand–Australian soil data exchange model and web services. This work had involved engagement with European agencies to reconcile Australia and New Zealand Soil Mark-up Language (ANZSoilML) with the ISO 28258 Soil Mark-up Language, but without much success. The SoilIE was initiated by MWLR and CSIRO to help progress development while considering other soil data models.

The interoperability experiment defined and implemented a simplified soil information model by consolidating core concepts and features from existing standards and tested the result against an agreed set of use cases for the exchange and analysis of soil data. Participants then deployed services and web clients that demonstrated the delivery and integration of soil sampling and sensor data. These were combined into a larger data set that included contributions from the participating agencies in Europe and North America.

The interoperability experiment was a partial success. It proved that the existing Observations and Measurements specification and its Timeseries Profile[22] could handle the delivery of most soil observation data, and it defined a simple scheme for soil descriptions. There were also clear signs of much common ground with other OGC environmental exchange models (especially the WaterML[23] suite of specifications and GeoSciML[24]).

However, from a technical perspective the results were unsatisfactory: the demonstration services made extensive use of XML as the data encoding and ageing service protocols (e.g. WFS 2.0[25]), which are not well supported by modern web developers. Members of the OGC, led by the US Geological Survey, recognised this and initiated another OGC interoperability experiment, ELFIE, to explore the use of modern techniques to publish

[21] http://portal.opengeospatial.org/files/?artifact_id=41579 and http://portal.opengeospatial.org/files/?artifact_id=41510

[22] http://docs.opengeospatial.org/is/15-043r3/15-043r3.html

[23] Specifically, WaterML 2.0, Part 4: GroundwaterML 2.0. http://www.opengeospatial.org/standards/gwml2.

[24] The GeoScience Mark-up Language. http://www.opengeospatial.org/standards/geosciml

[25] http://www.opengeospatial.org/standards/wfs

environmental data from a distributed set of data providers. The experiment found that the recommendations of the joint OGC/W3C Spatial Data on the Web Working Group could be applied to environmental data, particularly the use of Linked Data[26] principles and JSON for Linking Data (JSON-LD[27]).

Together the information models, web service protocols and data encodings tested by the SoilIE and ELFIE form the basis for an integrated soil/land-use data PoC infrastructure.

4.3 Proof of concept

As noted above, soil quality (or health) is one of the environmental performance indicators used in State of the Environment reporting. It is also of interest to many stakeholders. The use case for the PoC was a tool and related infrastructure that could inform a user about soil health in New Zealand. Soil health is evaluated based on seven soil quality indicators that best relate to the various aspects of a healthy soil. The user needed to be able to view soil quality data and how it varied by location, soil type and land use. The user also needed to see how a soil rated in relation to what are considered healthy soil conditions.

4.3.1 Data sets

The PoC used three types of data.

1. Soil observation data captured at sampling sites: while the IDA programme worked with the soil quality monitoring data, these were not available to the PoC due to restrictions on public access to that data. To ensure the PoC could be shared and demonstrated, a soil quality data set was fabricated from the publicly available National Soils Database (NSD) by filtering the data to provide only relevant soil quality properties and lab measurements. The soil sample and laboratory data were published using the OGC Observations and Measurements 2.0 (O&M) standard or its OGC/W3C semantic web derivative SOSA (Sensor, Observation, Sample, and Actuator) ontology[28]. Soil descriptions were provided using the SoilIE XML schema (SoilIEML).
2. Land-use data were sourced from the *in situ* observations made at the NSD soil sampling sites and from the national 2017 Land Use of New Zealand (LUNZ) model developed in IDA (Manderson et al. 2017). The NSD land-use observations were provided using Observations and Measurements. The LUNZ model was served as is: currently there is no standard for the delivery of land-use classification surfaces.
3. Soil and land-use vocabularies were collated into a vocabulary management system. Soil vocabularies were taken from the *New Zealand Soil Description*

[26] http://linkeddata.org/

[27] https://json-ld.org/

[28] https://www.w3.org/TR/vocab-ssn/

Handbook (Milne et al. 1995) and *New Zealand Soil Classification* (Hewitt 2010). Land-use vocabularies were taken from the *New Zealand Soil Description Handbook* and the 2017 version of LUNZ. Vocabulary data were provided according to the Simple Knowledge Organization System[29] (SKOS).

4.3.2 Architecture

The PoC architecture is based on that of the SoilIE, but it is complemented by the option to request data according to the formats and principles tested by ELFIE. Figure 5 shows the PoC architecture. Readers are referred to section 10 of the SoilIE engineering report (Ritchie et al 2016) for a description of the component OGC and linked data services. Tables 1 to 4 describe the specific function of the components in this PoC.

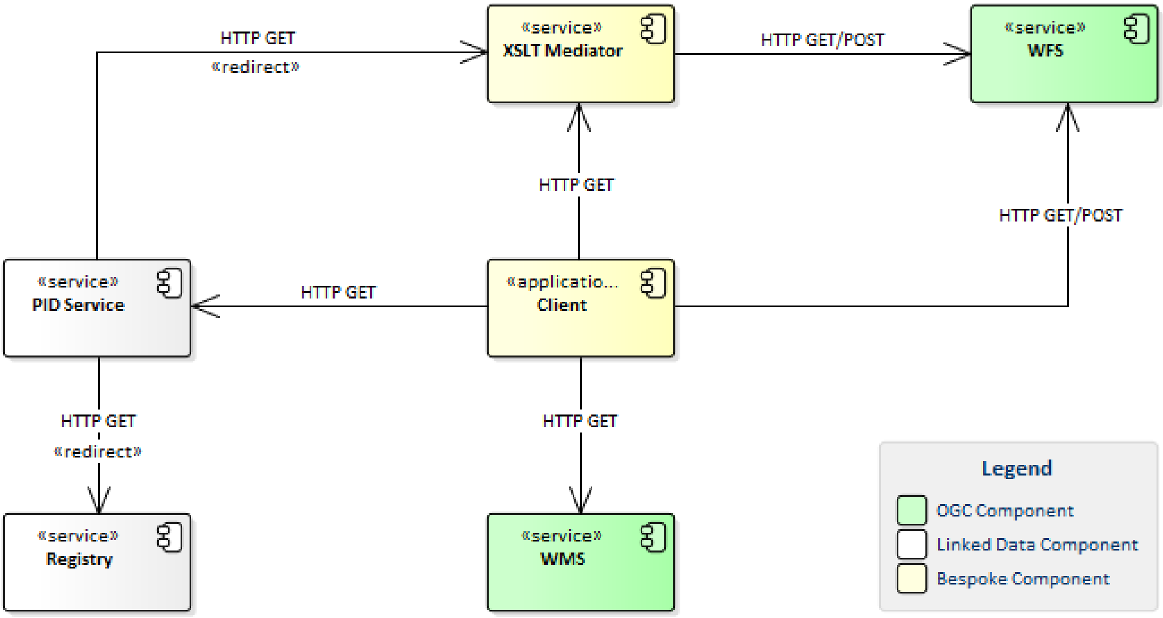


Figure 5. Overview of the IDA multi-indicator infrastructure proof of concept component architecture. The direction of associations is from the component that initiates the interaction between components. Source: Ritchie et al 2016, Figure 13.
WFS = Web Feature Service; PID = persistent identifier; WMS = Web Map Service.

Table 1. Proof of concept Web Map Service

Web Map Service (WMS)	
Version	1.1.0 [OGC 09-110r4]
Implementation	Geoserver
Role	Provision of IDA LUNZ land-use maps.

[29] https://www.w3.org/2004/02/skos/

Data standards	None

Table 2. Proof of concept Web Feature Service

Web Feature Service (WFS)	
Version	2.0 [OGC 09-025r1]
Implementation	Snowflake Go Publisher WFS
Role	Provision of features describing site registration, and soil description, sampling and observations for evaluation of soil health
Data standards	Observations and Measurements 2.0 [OGC 10-004r3 (abstract); OGC 10-025r1 (XML)] Soil data interoperability experiment XML schema [OGC 16-088r1]

Table 3. Proof of concept Persistent Identifier Service

Persistent Identifier Service (PID Service)	
Version	1.1.137 [Auscope PID]
Implementation	AuScope SISS PIDService
Role	Manages the resolution of URIs identifying GML Features (soil and sampling data) and SKOS Concepts (term and property definitions). Links embedded in responses from the WMS, WFS and XSLT mediator are directed through this service. Agents dereferencing the URIs can request the appropriate media type (e.g. JSON-LD or RDF) using HTTP Content Negotiation (specifying the media type in the HTTP Accept header) or by appending an appropriate well-known extension (e.g. .json or .ttl).
Data standards	n/a

Table 4. Proof of concept XSLT Mediator service

XSLT Mediator	
Version	1.0
Implementation	Bespoke java servlet (MWLR)
Role	A proxy service used to translate GML output from the WFS to JSON-LD, RDF (XML and Turtle) and HTML.
Data standards	Semantic Sensor Network Ontology / Sensor, Observation, Sample, and Actuator (SSN/SOSA) [OGC 16-079; W3C TR/vocab-ssn]: JSON-LD representation of O&M 2.0 features as SOSA. Simple Knowledge Organisation System [W3C TR/skos-reference]: JSON-LD and RDF representation of vocabularies and terms. GeoJSON [IETF RFC 7946]: spatially located representations of soil quality sites and samples no formal model for property definitions.

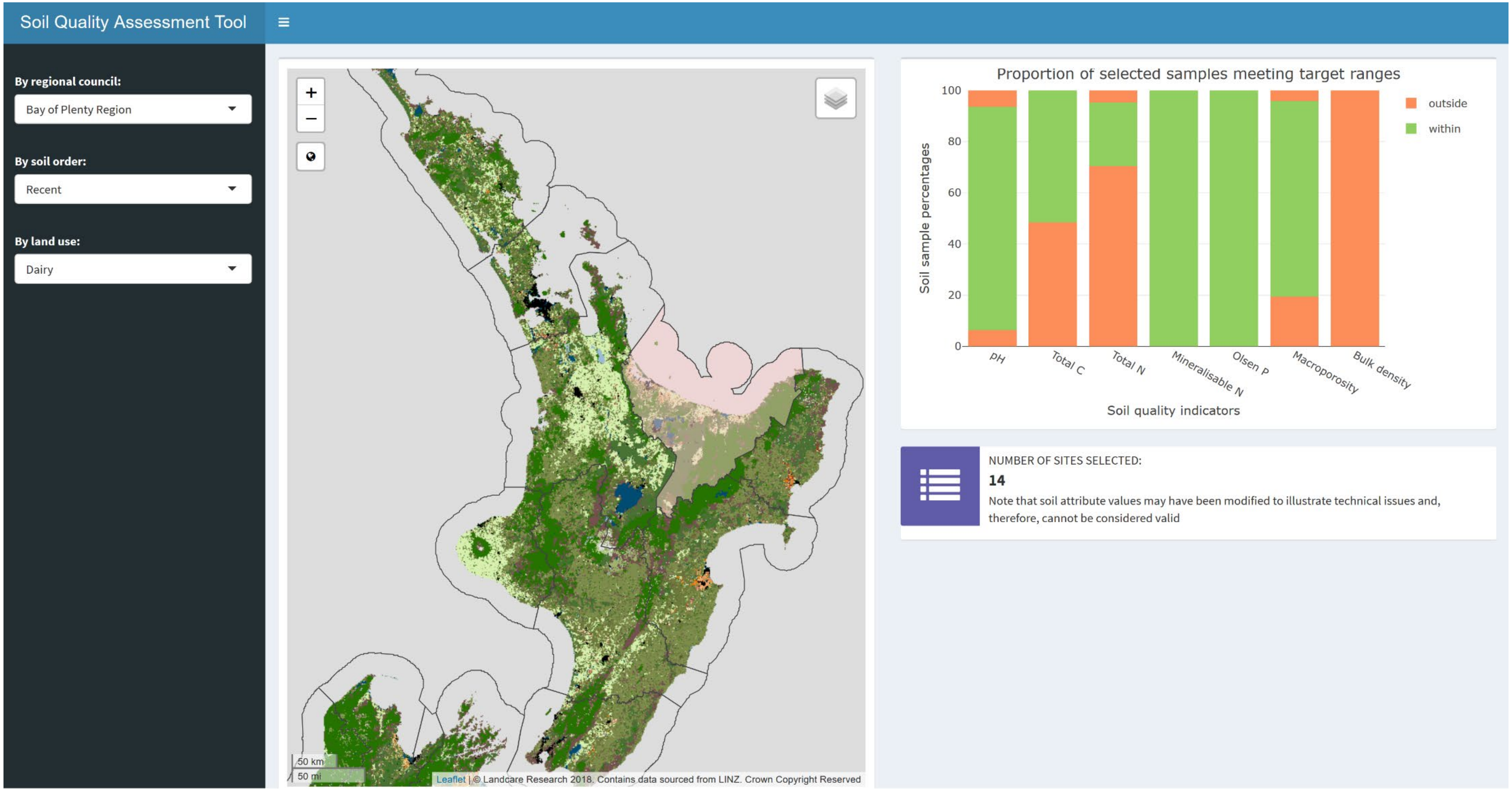
4.3.3 Demonstration client

The Soil Quality Assessment Tool (Figure 6) is a prototype application created for the PoC to provide information about soil health in New Zealand using data delivered using the PoC architecture. Soil health is evaluated based on the seven soil quality indicators that best relate to the various aspects of a healthy soil: pH (acidity), Olsen P (fertility), total carbon and nitrogen, mineralisable nitrogen (all related to organic reserves), macroporosity, and bulk density (physical status). Depending on soil type and land use, different target ranges were previously specified by experienced soil scientists in order to define value ranges characterising healthy soil conditions.

The Soil Quality Assessment Tool was implemented using R Shiny. It sources data from different web services. First, generalised boundaries for regional councils are retrieved from the Stats NZ Geographic Data Service using the OGC web feature service (WFS) standard. Second, various map images are loaded into the map window of the Soil Quality Assessment Tool based on the OGC web map service (WMS) standard. By default, an image of the IDA-generated 2017 Land Use of New Zealand (LUNZ) classification is shown, delivered from the PoC WMS. Third, soil attribute data are gathered from PoC WFS services using GeoJSON as data exchange format.

While site information is loaded just once when starting the tool, related soil profile data are retrieved from the WFS every time the user changes the region, soil order and land-use selections. In doing so, the Soil Quality Assessment Tool provides a useful example of how to utilise geospatial data, which is supplied by external web services following OGC standards, and eventually to turn these retrieved data into knowledge that matters to the user.

In addition to the Soil Quality Assessment Tool, general access to infrastructure components was tested using the open-source QGIS desktop GIS tool. Figure 7 shows soil quality site data (as GeoJSON) overlaid on the PoC WMS LUNZ land-use map. The map used HTTP URIs to simultaneously identify land-use classes and provide links to a controlled vocabulary, the description of which is shown to the user as a SKOS RDF response from the vocabulary service**.**



Soil Quality Assessment Tool

By regional council: Bay of Plenty Region

By soil order: Recent

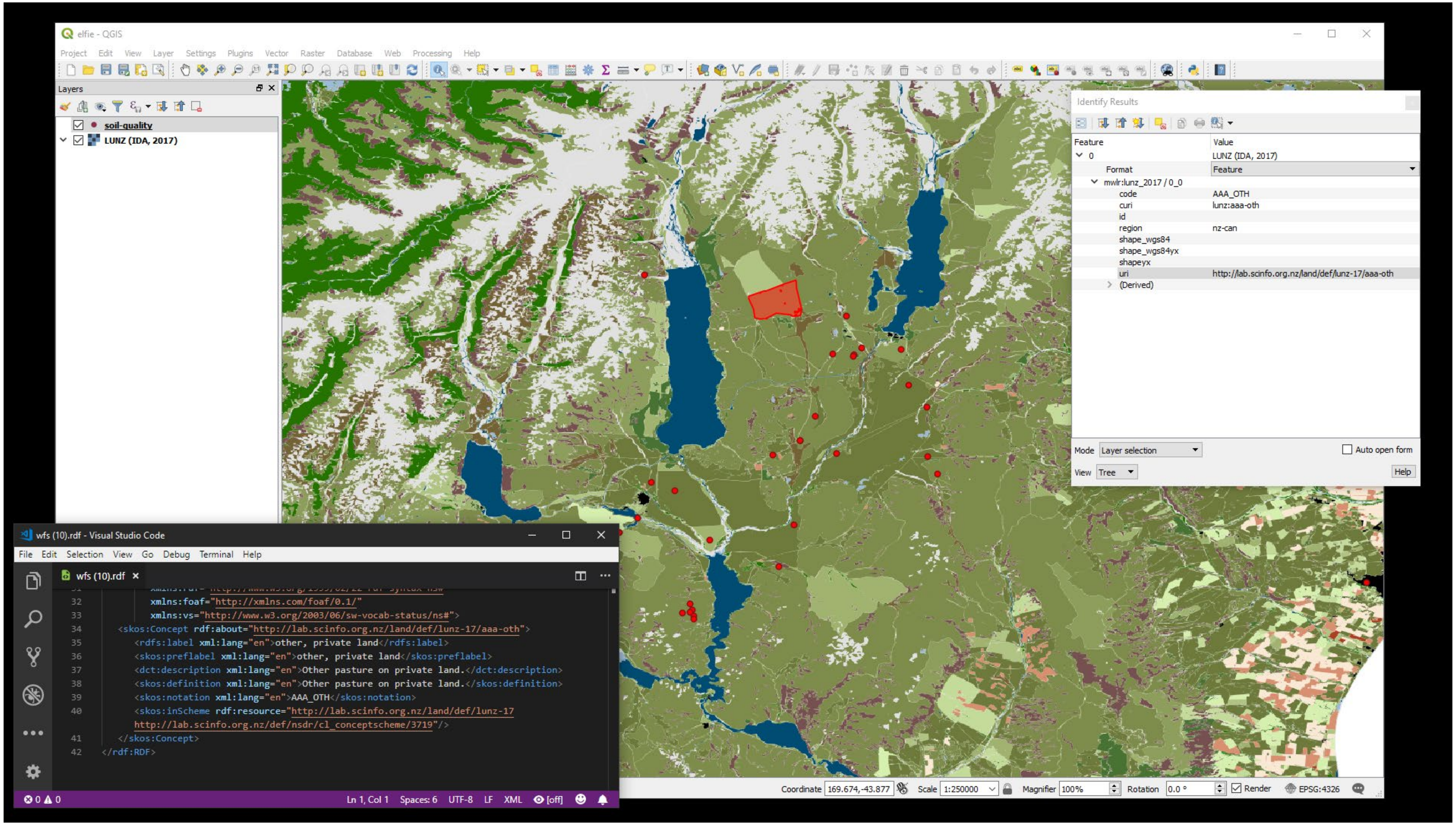
By land use: Dairy

NUMBER OF SITES SELECTED:

14

Note that soil attribute values may have been modified to illustrate technical issues and, therefore, cannot be considered valid

Figure 6. Screen shot of the IDA demonstration client developed in R shiny. (Map shows land use)



elfie - QGIS

Layers

☑ soil-quality
☑ LUNZ (IDA, 2017)

Identify Results

Feature	Value
0	LUNZ (IDA, 2017)
Format	Feature
mwlr:lunz_2017 / 0_0	
code	AAA_OTH
curi	lunz:aaa-oth
id	
region	nz-can
shape_wgs84	
shape_wgs84yx	
shapeyx	
uri	http://lab.scinfo.org.nz/land/def/lunz-17/aaa-oth
(Derived)	

Mode: Layer selection
View: Tree
□ Auto open form

wfs (10).rdf - Visual Studio Code

```
            xmlns:foaf="http://xmlns.com/foaf/0.1/"
            xmlns:vs="http://www.w3.org/2003/06/sw-vocab-status/ns#">
    <skos:Concept rdf:about="http://lab.scinfo.org.nz/land/def/lunz-17/aaa-oth">
      <rdfs:label xml:lang="en">other, private land</rdfs:label>
      <skos:preflabel xml:lang="en">other, private land</skos:preflabel>
      <dct:description xml:lang="en">Other pasture on private land.</dct:description>
      <skos:definition xml:lang="en">Other pasture on private land.</skos:definition>
      <skos:notation xml:lang="en">AAA_OTH</skos:notation>
      <skos:inScheme rdf:resource="http://lab.scinfo.org.nz/land/def/lunz-17
      http://lab.scinfo.org.nz/def/nsdr/cl_conceptscheme/3719"/>
    </skos:Concept>
</rdf:RDF>
```

Ln 1, Col 1 Spaces: 6 UTF-8 LF XML

Coordinate: 169.674,-43.877 Scale: 1:250000 Magnifier: 100% Rotation: 0.0 ° ☑ Render EPSG:4326

Figure 7. Screen shot of soil quality monitoring site (red dots) and land-use data (thematic map), and land-use vocabulary definitions (XML in bottom left) in QGIS.

4.3.4 Data types and views

Defining the right web service communication protocols is a relatively straightforward exercise: the OGC and Linked Data service interfaces are mature and between them address the needs of most clients. The more important aspect of a useful and robust implementation that performs acceptably is the provision of appropriate data formats and structures. Not every response to a request should have the same information content: there is no 'one-size-fits-all' solution to data delivery, so a pragmatic set of options for content is required.

During the design phase we considered three important data-use cases based on the environment in which the responses from the services were accessed and the reasons why they were being accessed. The boundaries between each are fuzzy, as environments and motivation may overlap.

1. ***Web presentation:*** web applications are probably the dominant user of this infrastructure. The services must therefore support access and responses that align with one or more of the widely used web data paradigms (e.g. ReST; Fielding 2000) and data formats (JSON[30]/GeoJSON[31]). Clients may have limited computing resources, particularly internet bandwidth, so responses must be compact and have minimal complexity.
2. ***Data analysis:*** this is perhaps the most important motivation for using the infrastructure. Analysis may be the data science requirement in order to calculate an indicator or for diagnostic reasons, requiring sampling of measurement metadata to help establish why a value is missing or anomalous. Data packages are therefore more complex because they carry more information. Clients may be web applications, or desktop tools, or development environments (e.g. statistical computing environments such as R). While desktop tools may have access to more computational and network resources, the likelihood of web usage means efficiency is still important, so large and complex data packages should be used only when necessary.
3. ***Data management:*** this is the primary working environment for data providers. These must support the exchange and validation of the complete set of data describing a resource. This includes the values themselves and all metadata, especially those related to provenance and uncertainty. They will also be the raw data source for the smaller and more specific data objects provided for data use cases 1 and 2. Of necessity, data management services must support large, complex and rigorously structured data. Speed and performance are secondary to accuracy and information retention.

The PoC was designed and implemented so that it supported a variety of protocols and formats to meet these use cases, allowing applications to pick combinations of each as

[30] http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

[31] https://tools.ietf.org/html/rfc7946

appropriate to their needs. Figure 8 shows a simple matrix of combinations and their suitability for a use case.

	GML	JSON-LD	GeoJSON
Data management	🟩	🟨	🟥
Data analysis	🟨	🟩	🟧
Web presentation	🟧	🟦	🟩

Data validation:

• structure	XSD/OWL/SHACL
• content	Schematron/SCHACL

Figure 8. Proof of concept data type / use case combinations. Cell colours denote suitability red: unsuitable; orange: limited and/or discouraged; yellow: limited but acceptable; blue: good option in the absence of more suitable options.

Overall, the three main data formats used by the PoC (GeoJSON, JSON-LD and GML) lent themselves well to a specific use case.

1 GeoJSON was very good for the provision of spatial data to web applications and R. Its ubiquity and simple data structure mean it is well supported by most web frameworks. It is, however, limited by the specification's restriction of spatial data to the WGS84 geographic coordinate system[32].

2 JSON-LD was the most appropriate for the structured data needed for more complex data analysis. It was also suitable as an alternative to GeoJSON for simple aspatial data objects. JSON-LD can be provided as a JSON object readily parsed by JSON-aware applications, but with additional robustness: well-identified objects and robust linking between data objects, and semantic rigour afforded by links to ontologies through the JSON-LD contexts (at the very least these can be treated as a data dictionary for property definitions).

3 GML provides a rich format for data exchange and validation use cases. Data can be constrained and checked by XML Schema Documents (structure) and Schematron[33] (content). In addition, the XML stylesheet transformation language (XSLT) is a powerful tool that allows the GML to be restructured into new forms for specific use cases, or indeed new formats (e.g. JSON/GeoJSON). The richness

[32] https://tools.ietf.org/html/rfc7946#section-4

[33] http://schematron.com/

comes at a cost, though, as XML documents are larger than the equivalent JSON documents and GML's very explicit typing of classes (also known as Features) and properties makes documents larger still. This means that while appropriately structured GML can be used for all use cases, it is not suitable for web data delivery (high band width). In future, GML may be replaced or augmented using semantic web technology with OWL/RDF ontologies and the Shape Constraint Language[34] (SHACL) playing a similar role to XSD and Schematron.

4.3.5 Performance considerations

As a PoC the IDA programme focused on the capabilities of the services and the appropriate provision of data. Some consideration of the performance of the services was necessary to ensure stability and reasonable response times, especially when applications requesting data may consider a slow response time to be a time out. In a production environment, good performance will be essential, particularly on the Web: there is little tolerance by users of delays lasting several seconds or longer. Slow data services can be a significant limiting factor in website performance.

A combination of tools was used to speed up the infrastructure's performance:

- appropriately resourced database and application servers – the extraction of data from a database and subsequent translation to GML or JSON can be resource (CPU and memory) intensive in both the database and on the web service software's host
- appropriate database indexing and optimisation, including pre-packaging some data for the web service using materialised views
- appropriate caching (storage of precompiled copies for quick retrieval) of web resources, especially where the construction of a resource can be time consuming, but once it is created is unlikely to change frequently:
  - Geowebcache[35] was used on the WMS, and WMTS (Web Map Tile Service) endpoints were enabled
  - HTTP caching[36] was enabled on the web server to cache all resources identified with an HTTP URI.

4.3.6 Publication and analysis pipeline

A multi-indicator data infrastructure is a crucial nexus in a data pipeline. It provides access to the data repositories that hold the data from which indicators are generated and provides access to the results of the analysis. For the purposes of the discussion below, the soil quality data pipeline can be considered as a proxy for publication of most solid earth environmental data.

[34] https://www.w3.org/TR/shacl/

[35] http://geowebcache.org/

[36] https://www.w3.org/Protocols/rfc2616/rfc2616-sec13.html

The current data pipeline is fragile (Figure 9). As discussed above, the data harmonisation of the source data is severely compromised, and it is impossible for data to flow through a service-oriented infrastructure into an analytical tool, the portal used to visualise the source data, or the indicators generated through analysis without significant human intervention and manual data processing. The PoC has shown that once the data are well organised, the necessary services can be deployed and used by analytical/visualisation tools.

Figure 10 shows two options for a functioning environment. In both cases they introduce a data cube (alternatively a local data cache) to support analysis in a high-performance computing environment (these were evaluated by the IDA programme and are summarised by Jolly (2018). The data cube would pull its source data into a local cache to ensure fast and stable access and therefore faster processing.

Each environment assumes a well-managed set of distributed data sources[37] published using web services that conform to the standards of the multi-indicator infrastructure. These may be the feeder services for a central repository that feeds a central service (Figure 10a), or they may be published through a broker service that acts as a proxy or a redirection agent. As far as the analytical and visualisation tools are concerned, the central service/broker will behave in the exact same way. This allows the source services to change from one option to another, or a hybrid, without affecting the client. In this model we assume that analytical results are stored in the data cube, but other options are permissible.

A recommendation on the best configuration of the infrastructure is beyond the scope of this report and requires further research.

[37] These data sources will be the raw data (as held by regional councils), but could also include snapshots (either held by the councils or MfE or an agent of MfE) of data used by a particular State of the Environment report. Snapshots allow a given report to be faithfully reproduced, something not necessarily possible in a single, dynamic data set.

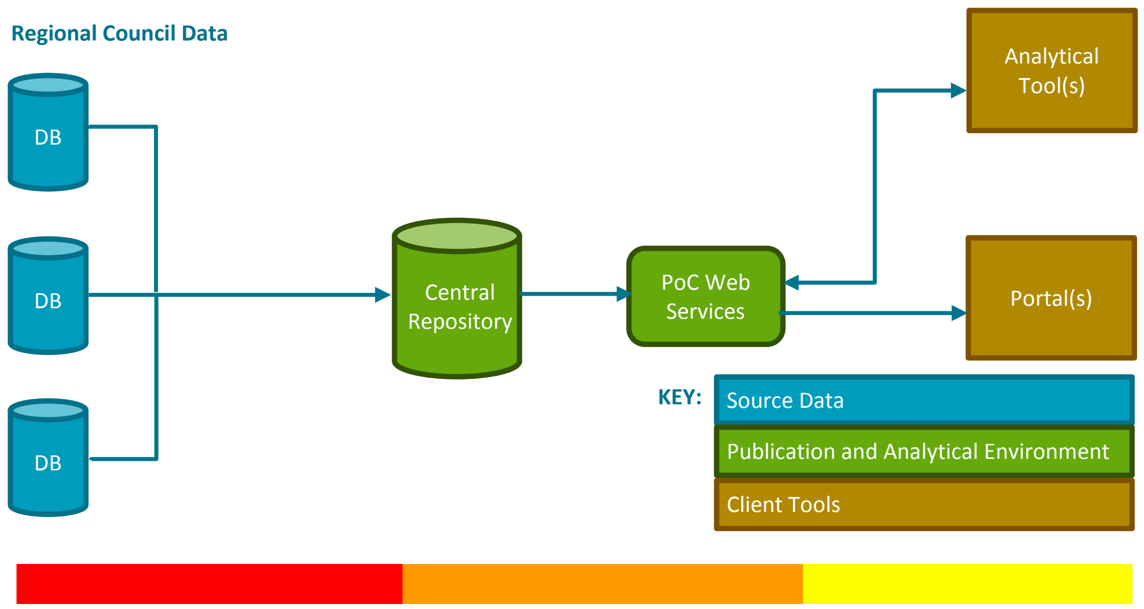


Figure 9. The soil quality data pipeline, as used by the IDA programme
The coloured bar at the bottom shows the impact of compilation, publication and analysis components on the functioning of the whole system. The impact is governed by the state/maturity of each part: red: severely compromised, orange: possible but badly compromised by upstream effect; yellow: compromised by upstream effects, with potential for upstream components to reduce the impact of other upstream components.

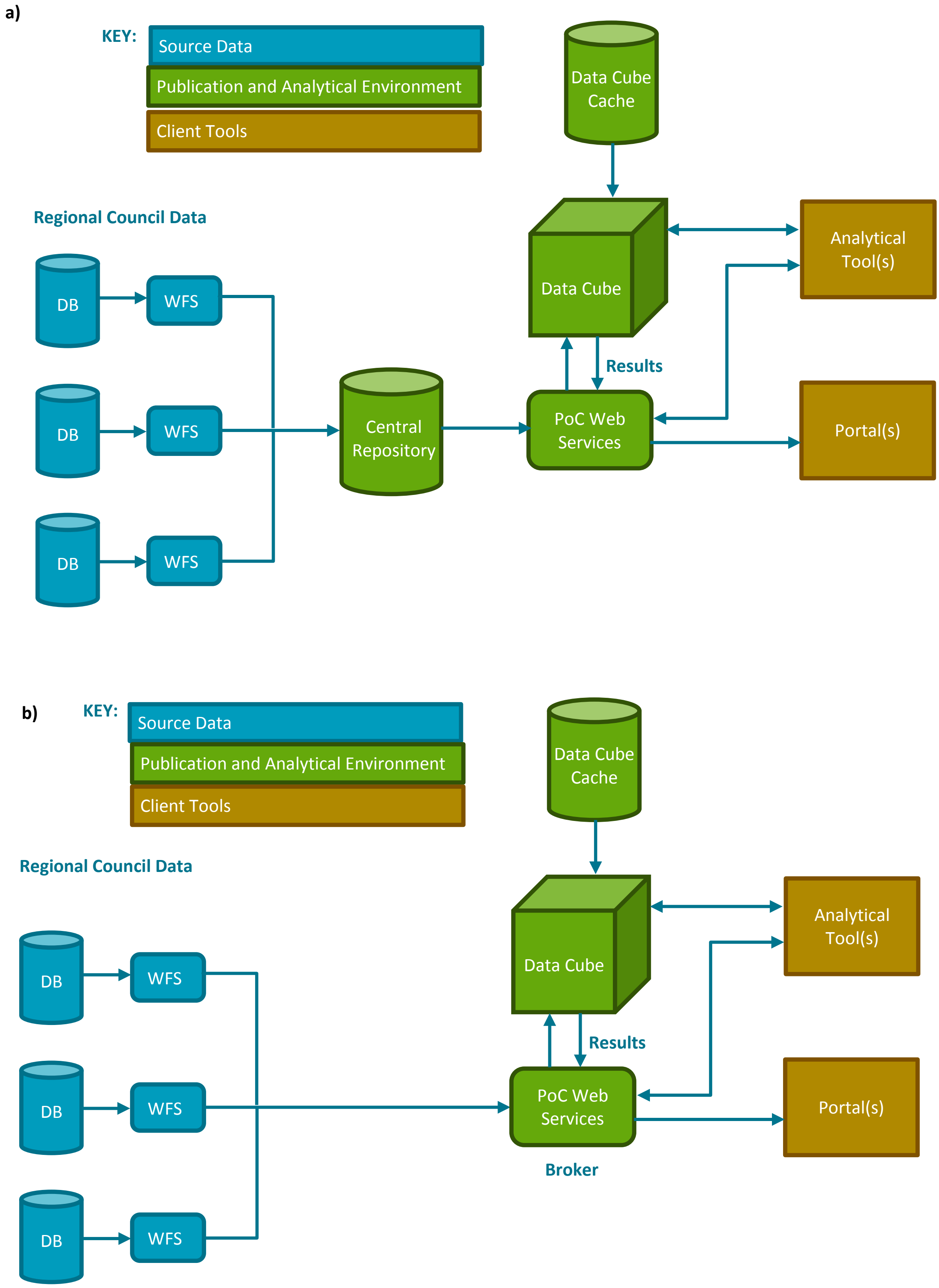


Figure 10. Potential configurations of a mature multi-indicator infrastructure and analytical data cube (with a performance-boosting cache providing fast access to locally stored data): (a) well-managed data sources publish web services that feed standardised data into a central repository for redelivery; (b) well-managed data sources publish web services that are accessed via a central service broker. All web services in the system conform to the same standards.

4.3.7 Role of data infrastructures and data management

The PoC has shown that poor management of data can severely compromise an entire infrastructure (Medyckyj-Scott et al. 2016). In the case of the soil quality data, providers can be forgiven because there is no clear guidance for the publication of soil quality data. This is, however, solved by the definition of the data and content standards for the multi-indicator infrastructure, as they can be used as the template for data delivery. In situations where no internal practices exist, they can be used to design the local systems for robust storage (this is MWLR's experience, having used the Observations and Measurements and SKOS standards to build the National Soil Data Repository).

4.3.8 Discussion and future work – technical improvements

The PoC was a qualified success. It proved that a set of web data services could be deployed to provide raw data for analysis (the soil quality data set) and the results of an analysis (the IDA LUNZ data set). The provision of soil and land-use data (considered alongside the work of the hydrology and geology working groups in the OGC) shows that multi-domain / multi-indicator infrastructure, at least for the solid earth, is achievable.

To further enhance the infrastructure, we recommend the following supporting work.

- The tight integration of provenance information into every part of the system – the PROV evaluation conducted by the IDA programme (Spiekermann et al., submitted) uses technology and standards that could be integrated into those used in the PoC.
- The tight integration of uncertainty data into every part of the system, represented using both quantitative and statistical methods.
- Development of authentication and authorisation technology: by their nature indicators, and the data that support them, can be contentious and may not be in the public domain. Ownership and usage constraints on the data must be respected, and access to different parts of the system must be managed. Any technology used must be able to support different or changing models of access.
- The development of policies and tools to ensure the creation of the persistent, nationally unique identifiers needed to identify and link the data in the distributed infrastructure: we recommend continuing the work on establishing national monitoring site identification started by regional councils and MWLR in the context of Envirolink (Ritchie & Osorio-Jaramillo 2017).

Ultimately, the success of the PoC is not surprising. Standardised infrastructures simply work with existing technology, with a defined set of constraints on data structure and content, and well-established communication protocols. Once agreed and honoured, these constraints make for a stable and consistent system that users can connect to with confidence. Essentially, participants enter into a contract to provide and use a very clearly defined system.

The challenge when deploying an infrastructure is establishing a willing and empowered community that will create, maintain and use the infrastructure. This requires a clearly defined need for the system, a mandate to operate part or all of it, and the human and financial resources to do so. Ultimately the infrastructure will succeed or fail due to its social architecture.

5 Social architecture

The term *social architecture* is used in this document to describe the social aspects that would underpin the development, implementation, and operation of a common data infrastructure for environmental data in New Zealand. The term complements the more common focus of the technical architecture.

This section starts by describing the technical–social challenges faced in each data domain considered by the IDA programme, how these challenges relate to the multi-domain indicator infrastructure (multi-indicator infrastructure), and how the consideration, design and adoption of a social architecture may reduce the complexities and increase the likelihood of a successful environmental data infrastructure.

5.1 From single-domain to a multi-indicator infrastructure

Sound and fit-for-purpose infrastructure and technologies for leveraging data and information are key components of the IDA programme. However, as noted earlier, the challenges in building a multi-indicator infrastructure are in some sense different and more complex than the challenges related to single-domain data infrastructures.

Some of the common challenges within the single domain data infrastructures were related to data quality, the complexity of data harmonisation, the rights to access and use data, and the selection of appropriate technologies. Tackling these challenges requires understanding and acknowledging human behaviours and motivations to share data and comply with agreements. Although the IDA programme managed to weather these challenges to create appropriate implementations per domain, it was very much a 'hand-crafted' solution and more work needs to be done to avoid these issues continuing to emerge again in the future. The work that needs to be done is an interweaving of data, technology and social aspects.

5.1.1 Data quality, format and harmonisation

There is a general deficiency in the adoption of data management practices across the different domains (this has been recognised by others; e.g. Medyckyj-Scott et al. 2016 with respect to the Our Land and Water National Science Challenge). In some cases, there are defined data management maturity frameworks, standards, and even technology/ applications that should be used to collect, harmonise and manage the data. However, the use of such best practices, standards and technologies is not properly enforced or incentivised. As a result, there are data quality problems and differences in data formats within the domains that make harmonisation difficult and time consuming.

Ideally, collecting and harmonising data should be an automated process. Such automation is possible when the input data is standardised with (or without) the use of applications. The standardisation may be achieved at different levels by compliance to agreements, use of centralised naming systems (e.g. for species or sites), or the use of software applications for collecting information that validate the data as it is entered or captured against agreed business rules.

In the same way, data quality can be guaranteed by enforcing the use of standards (methods, rankings, etc.) accompanied by the implementation and use of data collection applications.

In summary, standards and technologies are enablers, but if there are no agreements and incentives for using them, data collection will continue to be a problem, with resulting issues with data harmonisation, sharing and data quality.

5.1.2 Access to data

Another common challenge among the different domains was the right to access, expose and reuse data. In many cases the data that MWLR manages on behalf of the country is open data and accessible to the public, researchers, etc. However, MWLR is also contracted by regional councils and other parties to collect and analyse date for them. Some of these contracts may have restrictions on how the data can be subsequently used. Likewise, increasingly land owners want a say in how and who uses the data collected from their land; this is a particular issue for Māori where past knowledge 'giveaway' has not brought them benefits, with the result that they are now much warier about sharing data[38].

In some cases, as has happened with some of the soil-quality data, MWLR is authorised to collect the data and put them into the system (e.g. NSDR), but they cannot be used for other analysis or shared with other parties. From the point of view of the infrastructure, this is not a problem: MWLR has in place the authentication and authorisation components and servers that secure access to the data. But this restricts the gain in knowledge that MWLR and New Zealand as a whole can get from using the data in future analysis.

In the future, negotiation with the different entities that own the data is a key aspect to include in agreements in the design of a social architecture that underpins the operation of the infrastructure. What data are to be collected? Who can access them? Under what circumstances, at what spatial and temporal scale? Can the data be released if anonymised and/or aggregated, etc.? In this way the value of the data can be maximised through reuse rather than having their use restricted to the scope of the project the data was collected for.

[38] See Te Mana Raraunga – the Māori Data Sovereignty Network: https://www.temanararaunga.maori.nz/ng-mahi/

5.1.3 Choosing the technology

The process to choose the technologies for the infrastructures in each domain was done in isolation, with little co-design or co-development. As a result, a variety of protocols/standards, databases and services were used and built (see section 3, 'Domain-specific infrastructure'). This is understandable to some extent: each domain was trying to meet the demands of different users, funding levels and cycles, and complying with international developments and standards. However, a common architecture design would had been valuable to facilitate the maintenance and evolution of the systems.

The use of different programming languages, services, standards and technologies to store and expose data adds complexity to infrastructure maintenance. Consequently, not much depth of understanding of each different piece of technology can be pursued due to the need to spread time and resources on maintaining a varied set of technologies.

Going forward, the use of standardised designs, data architectures, protocols and services should guarantee robustness and scalability of the infrastructure. Nevertheless, the implementation of agreements on standards and common approaches needs to somehow be incentivised or enforced, as well as recognised by stakeholders and funders.

The challenges discussed above are transferable to a multi-indicator infrastructure context. The major challenge for a multi-indicator infrastructure is to define the standards for integrating and delivering data from different domains. As discussed in section 3, each domain uses its own standards, and so exposing the different data in a way that can be easily consumed and combined by the end-user is not straightforward. In part this is a failing at an international level, where cross-domain standards development is rare.

The OGC Geoscience Domain Working Group is perhaps unusual, in that it encompasses hydrology, groundwater, geology, and agriculture (soils), but there is no biodiversity focus[39]. The GEOSS Standards and Interoperability Forum[40] and the Research Data Alliance are other forums for cross-domain development of standards, but the former seems to be inactive and the latter is more of a community of practice.

Within New Zealand those bodies responsible for environmental data services across the different domains rarely come together to discuss best practice and data interoperability standards. Agreements on which standards to use, which standards need to be developed, and how to integrate and harmonise data across domains (data transformations) need to be worked on.

As expected, social challenges get more complex when interaction within different domains, people and interests converge. Therefore, a social architecture design that encourages multiple parties and domains to work together in the evolution, sustainability and usability of an environmental data infrastructure is key.

[39] There was an MOU between OGC and TDWG Biodiversity Information Standards group to collaborate on joint standards, but little came from it.

[40] http://geoss.omstech.com/index.php?option=com_content&task=view&id=41&Itemid=170

5.2 Definitions

Other initiatives around the world (INSPIRE[41], CGDI[42], FSDF[43], ALA[44], and NEII[45], among others) similar to the IDA programme have faced the same issues that IDA staff encountered doing the work undertaken in this programme. These initiatives have recognised the necessity of building not only the information infrastructure but the social foundation, agreements and governance groups that are required to create, keep alive and ensure fitness for purpose of the technical architectures in these types of initiatives (Box & Lemon 2015).

The fundamental aspects of a social architecture can be divided in three main components: governance, participation, and agreement framework. These three aspects should be understood in the context of legislation, policy and standards that constrain the operation of the data infrastructure (Figure 11).

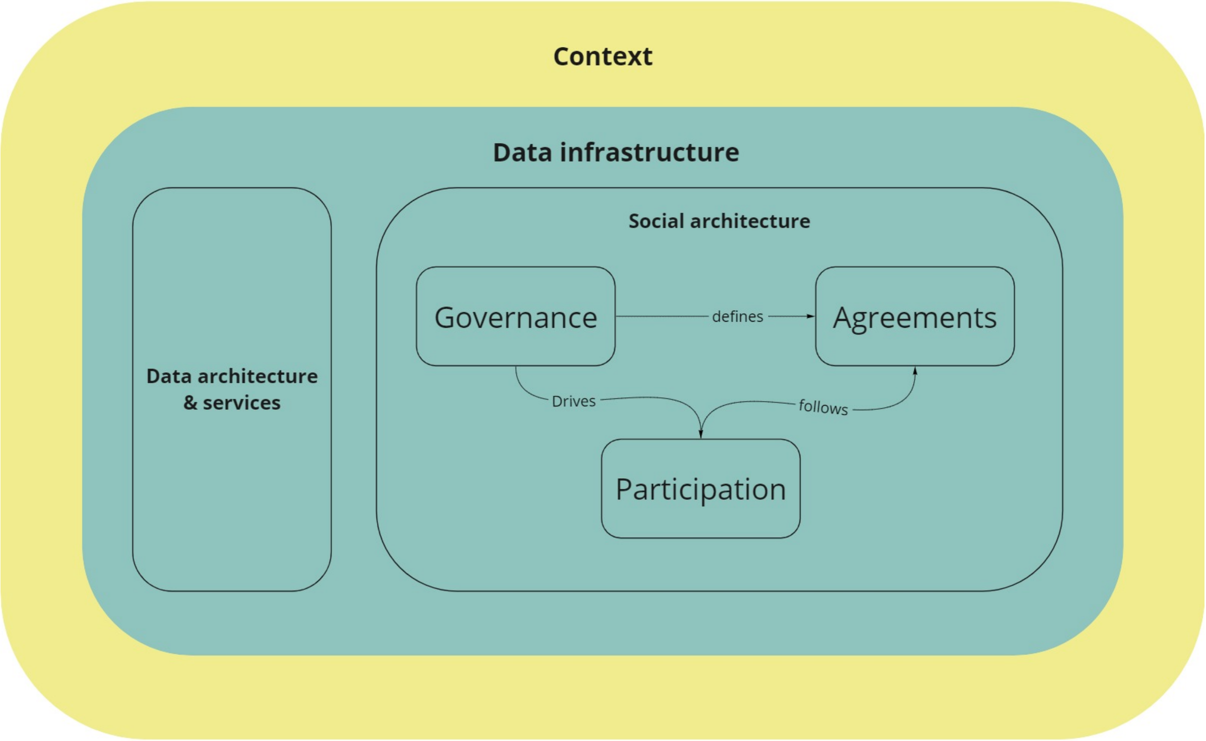


Figure 11. Data infrastructure. Adapted from Box & Lemon 2015.

[41] INSPIRE: (environmental) Infrastructure for Spatial Information in the European Community.

[42] CGDI: Canadian Geospatial Data Initiative.

[43] FSDF: Australian and New Zealand Foundation Spatial Data Framework.

[44] ALA: Atlas of Living Australia, Australia's national biodiversity database.

[45] NEII: the Australian National Environmental Information Infrastructure.

Governance: The governance body in the social architecture oversees compliance with the agreements; makes decisions on cooperation, agreements, interactions and participation; and drives the vision of the data infrastructure consortium.

The governance body has an authority structure and a scope of action, and defines memberships and representation of the different parties. A central authority is recommended, along with authorities for each data and/or application domain.

Agreements: These can be about common goals for participation, data provision, storage, rights, access and sharing. The agreements should be defined formally and stored in a registry. Versioning of the agreements, level of access and the document registry should be undertaken by a control body designated by the governance body.

Participation: This component is about stakeholders' engagement. Stakeholders should be identified, recognised and incentivised to take collective action. Any risk threatening participation should be addressed by the governance body, which may result in new agreements, or modifications to existing ones if necessary. Participation is possible through information sharing, common infrastructure and applications, definition and use of standards, and knowledge exchange, among others.

5.3 Proposed social architecture

The three basic components of social architecture defined for the Australia National Environmental Information Infrastructure (NEII)[46] may be adapted to the IDA context and to the creation of an environmental data infrastructure. Figure 12 is an adaptation of the social architecture defined by NEII. This variation of the NEII model aims to provide more detail on what other considerations are important for a social architecture. It also includes information on which tools may be used to provide a better understanding of the often-complex context and interactions that data infrastructure architectures reveal.

5.3.1 Governance

As mentioned in the previous section, governance provides the steering for the social architecture, helping to define agreements and incentivising participation. Defining a common vision or expected outcomes may facilitate the development of future agreements and ease participation. The outcomes space framework is one of the tools MWLR is currently exploring to help define the vision of projects. This framework helps to identify where the expectations of the different parties are focused. The desired outcomes may vary depending on the party, and may be related to improving a current situation, generating knowledge, or learning processes (Mitchell et al. 2015).

Once the different parties agree on the situation[47] they want to improve and how important it is, the knowledge[48] they want to generate and its importance, and the

[46] http://www.neii.gov.au/

[47] Situation: the situation the project aims to improve

learnings[49] they are expecting from the process and their importance, this informs the direction of the governance and can also direct future agreements (Mitchell et al. 2015).

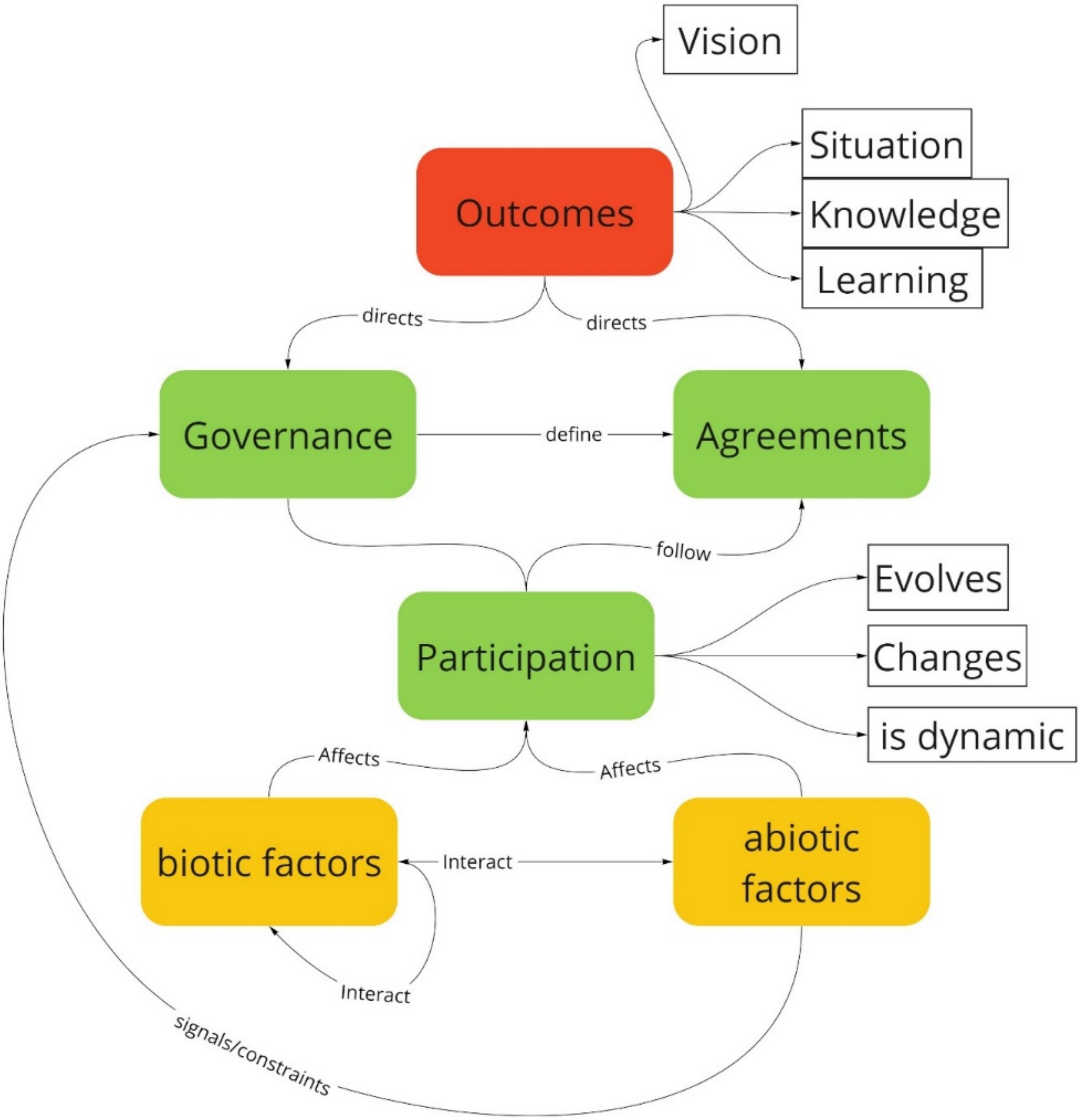


Figure 12. Proposed social architecture developed during the IDA programme.

[48] Knowledge: the types of knowledge they want to generate (data infrastructure, services, standards, etc.)

[49] Learning: the lessons learned from the process, are they important, and where to keep them

5.3.2 Agreements

These are the rules to operate the infrastructure and the relationships between parties. They are defined by the governance group and directed by the vision. The agreements can cover a variety of topics, from information technology and standards, to communications between different domains, data sharing rights, funding for future implementations, maintenance of the infrastructure, and where and how to house it. The implementation framework for the agreements should cover their entire lifecycle, from the requirements identification, to definition, application, and finally deprecation. Such agreements should be kept in a repository with version maintenance and related metadata.

The definition of agreements and their implementation define the participation within domains and across them.

5.3.3 Participation

Participation can be defined as the required collective action that guarantees the successful implementation of the infrastructure. The vision promoted by the governance group, plus compliance with the agreements and incentives, should boost the participation of the different parties.

To understand the means by which to incentivise participation, a knowledge ecology framework may be valuable. This framework helps to identify the biotic and abiotic factors that can underpin or undermine an initiative or project. Biotic factors are defined as the knowers and actors (stakeholders) that hold any type of knowledge that is part of the ecosystem under analysis – in this case the multi-domain data infrastructure. Abiotic factors refer to all the non-living components of the ecosystem, such as the policy context, legislation, standards, funding, centres and networks (Sofoulis et al. 2012).

Defining participation, and the corresponding agreements to encourage it, requires the identification of the biotic and abiotic factors and understanding their relationships and interactions. Using an ecosystem approach enables a bigger picture to be built and the identification of adequate incentives that consider context, risk and potential barriers (Fam & Sofoulis 2017). Demonstrating a deeper understanding of the different stakeholders and their environments may also foster their participation.

6 Conclusions

A multi-indicator environmental data infrastructure is a crucial nexus in the data pipeline, from collection of data to the publication of data for environmental indicators. It provides access to the data repositories that hold the data from which indicators are generated and provides access to the results of the analysis.

No such data infrastructure exists for New Zealand, so the IDA programme undertook to define one and test its viability with a simple proof of concept (PoC). Building a true, whole-of-environment system is an ambitious task, so the scope was tightly constrained to soil quality and land-use data.

The PoC showed that once the data are well organised, web services can be deployed and used by analytical/visualisation tools relatively easily. The Soil Quality Assessment Tool (client) provided a useful example of how to utilise the PoC data infrastructure, with retrieved data turned into knowledge that matters to the user. Further work would expand the scope of the PoC to the other domains, and in doing so could provide a technical basis for a multi-indicator environmental data infrastructure.

The PoC was a qualified success. It proved that a set of web data services could be deployed to provide raw data for analysis (the soil quality data set) and the results of an analysis (the IDA LUNZ data set).

Data quality, format, harmonisation, the existence of relevant domain data exchange standards, and access to data are challenges for both single-domain and multi-domain infrastructures. However, the challenges are more complex for multi-domain infrastructures. These challenges have a strong technology component. Much of the technology is available for appropriation and easy adoption, but more work on the domain standards needs to occur. However, it is the associated social architecture that will determine the successful implementation of an infrastructure, whether single or multi-domain.

MWLR is currently using the single domain data infrastructures to maintain and expose data from its Nationally Significant Databases. These infrastructures adapt and evolve as user needs change and new standards and best practices evolve. Future work on an implementation of the multi-domain infrastructure will occur in the context of the development of the Australian Centre for International Agricultural Research funded Pacific Soils Portal development in 2019. This project will provide an opportunity to progress from the PoC to an implementation encompassing a wider range of environmental data.

The international standards work that MWLR is undertaking with different international agencies such as the OGC and ISRIC (the World Soils Data Centre), help evolve the design and implementation of fit-for-purpose infrastructures. At the same time, the PoCs we have undertaken generated insights that may refine the standards, technologies and social interactions of the international communities.

7 Recommendations

- Several technology frameworks and possible architectures were investigated in the IDA programme, and decisions need to be made on which is the most appropriate for the broader New Zealand context.
- Adopting standards and international best practices across domains and agencies is critical: it reduces the risk we will reinvent the wheel, and we can build on the lessons gained by overseas institutions.
- The provision of data needs to be in data formats and structures appropriate for their use.
    1. GeoJSON is good for providing spatial data to web applications and tools like R.

2. JSON-LD is the most appropriate for the structured data needed for more complex data analysis.
3. GML provides a rich format for data exchange and validation use cases, and best suits scientific use.

- Interoperability of cross-domain data requires the adoption of common data service architectures to allow distributed environmental data to be located and integrated for analysis and visualisation.
- The performance of any infrastructure is an important consideration and needs to be thought about in terms of how the data will be accessed. Good performance will be essential, particularly on the Web.
- To further enhance the infrastructure, the following supporting technical work is required:
  1. the tight integration of uncertainty data and provenance information into every part of the system
  2. The development of suitable authentication and authorisation technology
  3. the development of policies and tools to ensure the creation of the persistent nationally unique identifiers needed to identify and link the data in the distributed infrastructure; data providers then need to publish URIs for their environmental features.
- National standards should be adopted for data collection within a domain (e.g. soil quality) so that harmonisation work can focus on the more difficult challenge of harmonising across data domains, through using semantics, ontologies and linked data.
- The technology aspect of data interoperability is not a high research priority. However, to ensure we can respond to the demands of new techniques in data delivery and analytics, interoperability becomes a responsive task, ensuring that state-of-the-art technology is used in a consistent and disciplined way.
- Being responsive also means infrastructure providers must be prepared, and willing, to update the technology they deploy as the community of users adopts new tools.
- The present document was intended to start the discussion on social architecture and highlight its importance for the successful implementation of complex information systems and infrastructures. Nevertheless, more work needs to be done to describe, analyse and design a complete social architecture for a national environmental data infrastructure in New Zealand. The development of a social architecture that fits the national context will boost data integration, sharing and collaboration between different domains and agencies, and increase the likelihood of success.

8 References

Box P, Lemon D 2015. The role of social architecture in information infrastructure: a report for the National Environmental Information Infrastructure (NEII). neii.gov.au: CSIRO:EP152134. https://doi.org/10.5072/83/5849a28b08365

Cavanagh J, Munir K, McNeill S, Stevenson B 2017. Review of soil quality and trace element State of the Environment monitoring programmes. Envirolink Advice Grant: 1757-HBRC226. Landcare Research Contract Report LC2861. http://www.envirolink.govt.nz/assets/Uploads/1757-HBRC226-Review-of-soil-quality-and-trace-element-State-of-the-Environment-monitoring-programmes.pdf

Fam D, Sofoulis Z 2017. A 'knowledge ecologies' analysis of co-designing water and sanitation services in Alaska. Science and Engineering Ethics 23(4): 1059–1083.

Fielding RT 2000. Architectural styles and the design of network-based software architectures. PhD. University of California, Irvine.

Glavic B, Dittrich K 2007. Data provenance: a categorization of existing approaches. Datenbanksysteme in Business, Technologie und Web, volume 103 of LNI, page 227–241. http://citeseerx.ist.psu.edu/viewdoc/download? doi:10.1.1.94.5718&rep=rep1&type=pdf.

Gong X, Marklund LG, Tsuji S 2009. Land use classification proposed to be used in the System of Integrated Environmental and Economic Accounting (SEEA). 14th meeting of the London Group on Environmental Accounting, Canberra, 27–30 April 2009. https://unstats.un.org/unsd/envaccounting/londongroup/meeting14/LG14_10a.pdf (accessed 28 August 2018).

Hewitt AE 2010. New Zealand soil classification. 3rd edn. Lincoln, Manaaki Whenua Press.

Jolly B 2018. A review of the current state of freely available data cube software. Manaaki Whenua – Landcare Research Contract Report LC3377.

Land Management Forum 2009. Land and soil monitoring: A guide for SoE and regional council reporting. https://www.mfe.govt.nz/publications/land/land-and-soil-monitoring-guide-soe-and-regional-council%C2%A0reporting

Lesslie R 2004. Land use and land management practices: concepts, terms and classification principles. Canberra, ACT, Australian Department of Agriculture, Fisheries, and Forestry, Bureau of Rural Sciences.

Manderson A, Jolly, B, Ausseil A-G 2017. The NZ Land Use Classifier. Manaaki Whenua – Landcare Research Contract Report LC3335.

Medyckyj-Scott D, Stock K, Gibb R, Gehagan M, Dzierzon H, Schmidt J, Collins A 2016. Our Land and Water National Science Challenge: a data ecosystem for land and water data to achieve the challenge mission. Palmerston North, Manaaki Whenua.

Milne JDG, Clayden B, Singleton PL, Wilson AD 1995. Soil description handbook. Lincoln, Manaaki Whenua Press.

Mitchell C, Cordell D, Fam D 2015. Beginning at the end: the outcome spaces framework to guide purposive transdisciplinary research. Futures 65 : 86–96. https://doi.org/10.1016/j.futures.2014.10.007.

Ritchie A et al. 2016. OGC Soil Data Interoperability Experiment technical engineering report. Open Geospatial Consortium. 16-088r1.

Ritchie A, Osorio-Jaramillo J 2017. National Environmental Monitoring Site Identification System. Envirolink Grant: 1729 - HZLC137.

Sparling GP, Schipper LA, Bettjeman W, Hill R 2004. Soil quality monitoring in New Zealand: practical lessons from a 6-year trial. Agriculture, Ecosystems & Environment 104(3): 523–534. https://doi.org/10.1016/j.agee.2004.01.021.

Spiekermann R, Jolly B, Herzig A, Burleigh T, Medyckyj-Scott D 2018. Automated generation of data provenance for transparent environmental research. Submitted to Environmental Modelling and Software Journal.

Sofoulis Z, Hugman S, Collin P, Third A 2012. Coming to terms with knowledge brokering and translation: background paper. Knowledge Ecologies Workshop, 28 November 2012. Background paper. Institute for Culture and Society, UWS, Parramatta.

Young A 1998. Land resources: now and for the future. Cambridge, Cambridge University Press.