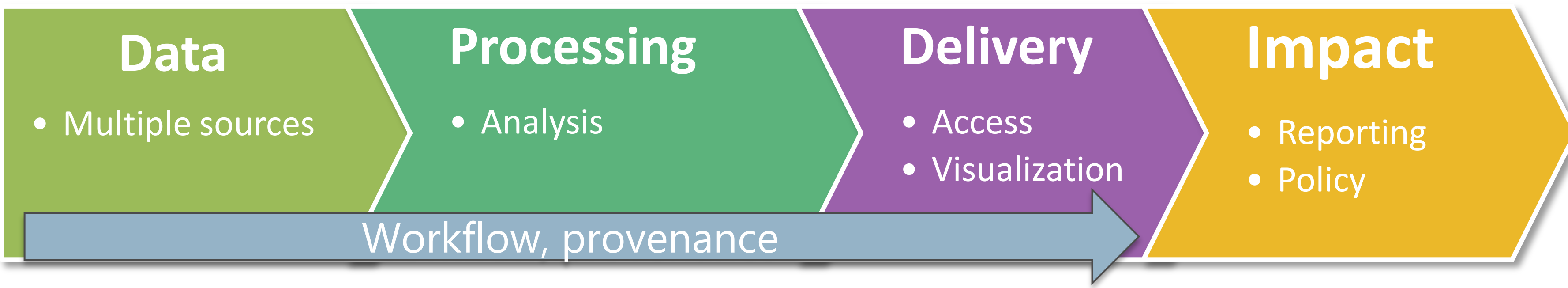
Manaaki Whenua
Landcare Research

Innovative data analysis: Getting the most out of environmental Data

Anne-Gaelle Ausseil, David Medyckyj-Scott, Alistair Ritchie, Jerry Cooper, Andrew Manderson

Introduction

Research question:
"What is the most effective approach to data analysis that would allow most new knowledge and value to be created from existing environmental data sets?"

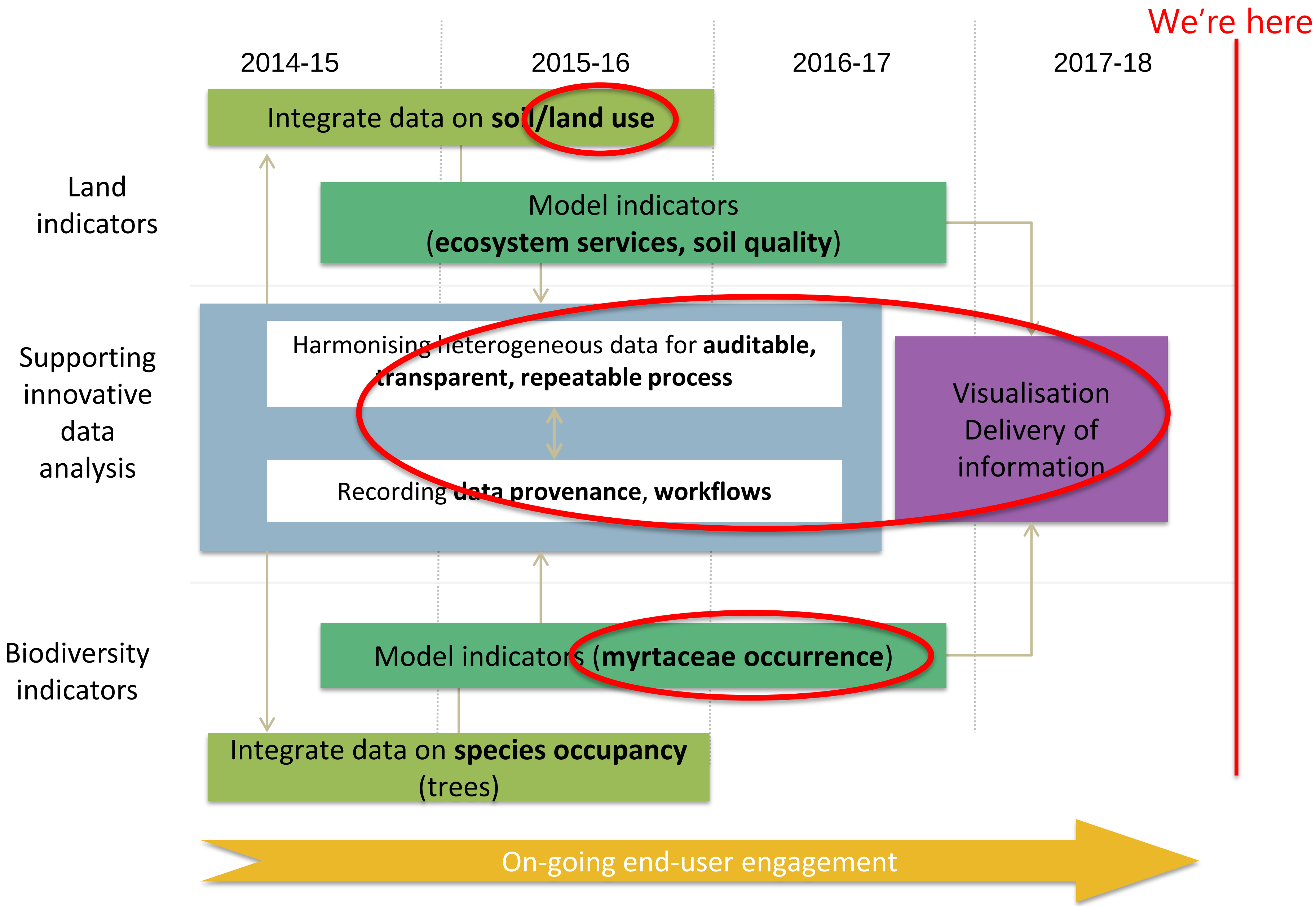


The programme aimed to:

- Bring together heterogeneous spatial data
- Analyse data and model indicators
- Characterise provenance, quality, uncertainties, workflow
- Visualise and deliver data

- 3 domains: land use, soil health, species occupancy

Project plan



Web Resources and more information

- www.landcareresearch.co.nz/science/e-science/ida

Land use

Andrew Manderson

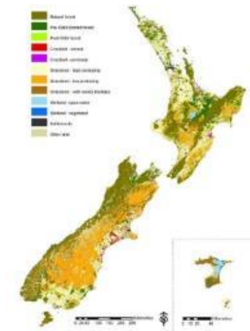
Introduction

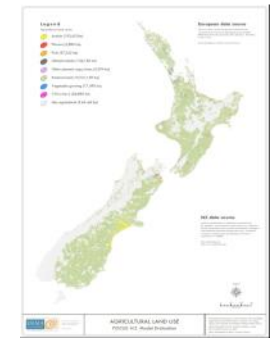
- The NZ Land Use Classifier (IDA development)
- Grassland improvement mapping using Innovative Data Analysis (IDA) techniques (post IDA)
- (Mapping the extent of artificial drainage in New Zealand)

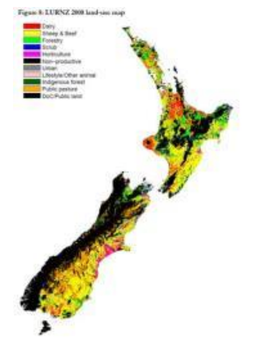
1. The NZ Land Use Classifier (IDA)

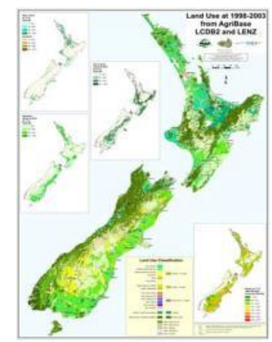
- Problem: NZ LU classifications lack transparency, robustness, temporal relevance, reproducibility (method), and differences in land use class definitions

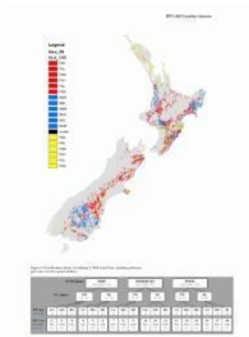
- IDA Method*: Reconstructed x3 (example) classifications, then rebuilt the workflow as software:
    - pyluc (software framework)
    - Data harvested from LRIS portal
    - Desktop or HPC
    - Automated generation of dataset provenance & documentation

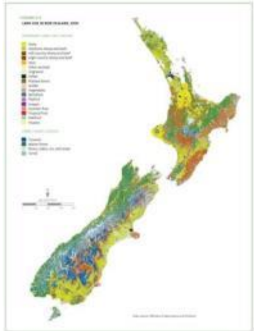










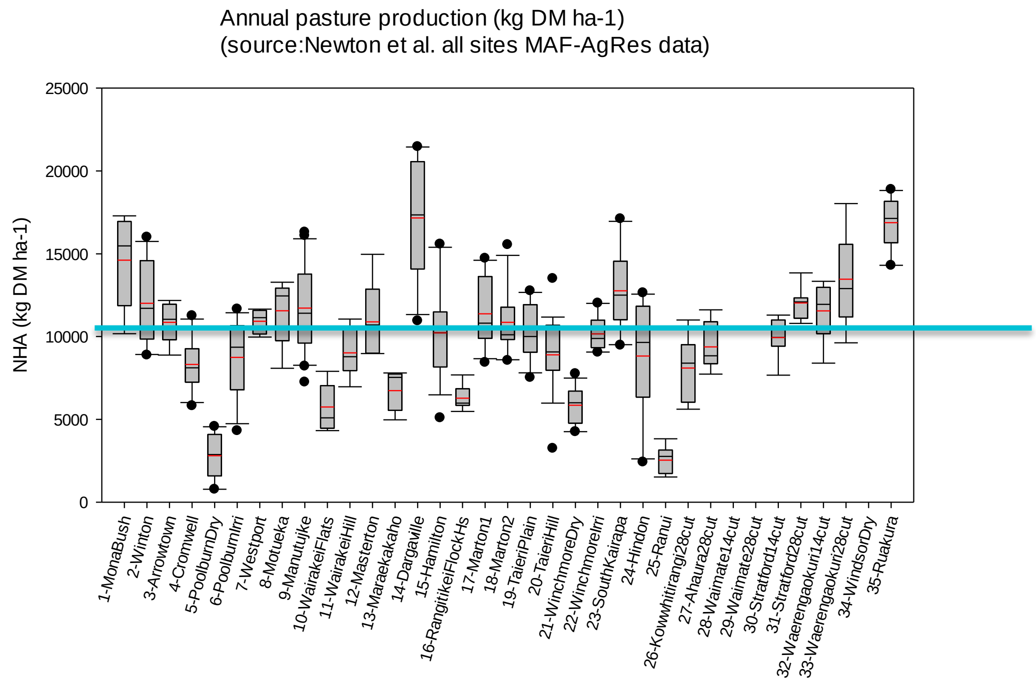


2. LUCAS LUM managed grassland classification (MfE)

- Aim: investigate improvement of LUM's high- and low-producing grassland classification (LUCAS MfE)
- Problem: remote sensing does not reliably differentiate HP from LP grassland. NZLRI used as workaround (now very dated)
- Methods
  - Reviewed HP & LP definitions
  - Reviewed spatial pasture modelling as an option
  - Fuzzy logic classification (likelihood of being high producing)

2. LUCAS LUM managed grassland classification

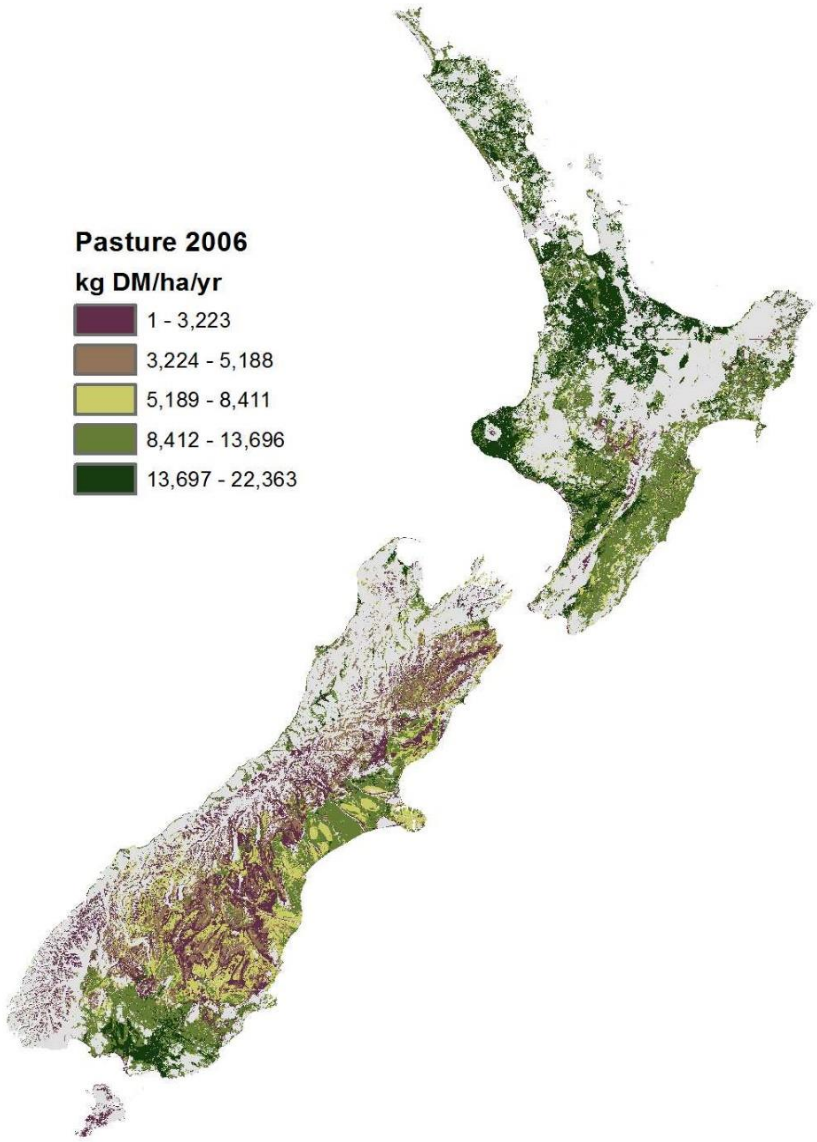
- Definitions are generally vague and qualitative
- Many pasture production models exist but…
  - Annual pasture yield is quite variable (one year HP; another year LP)



  - Land development and farm management have a major effect on pasture yield. We have no (spatial) data.

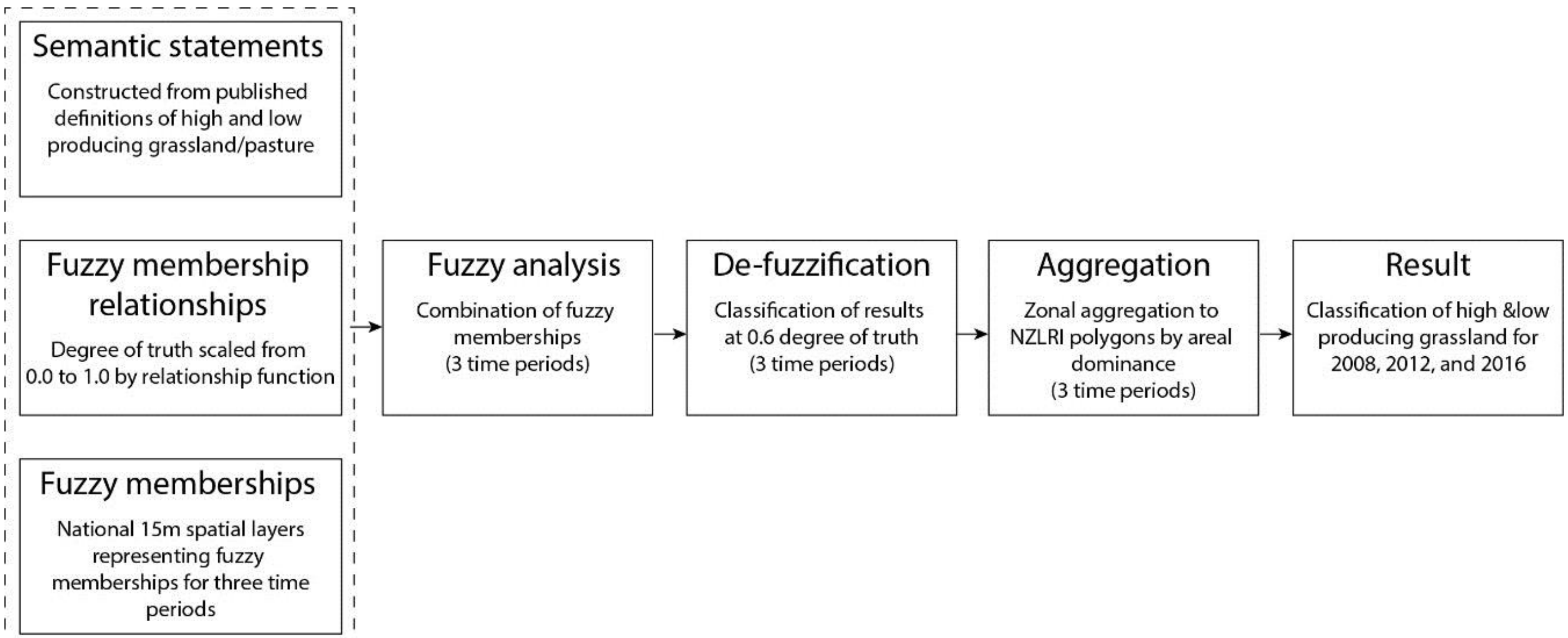
2. LUCAS LUM managed grassland classification

- Spatial-temporal (daily) pasture yield modelling for NZ. (Moir et al. method)



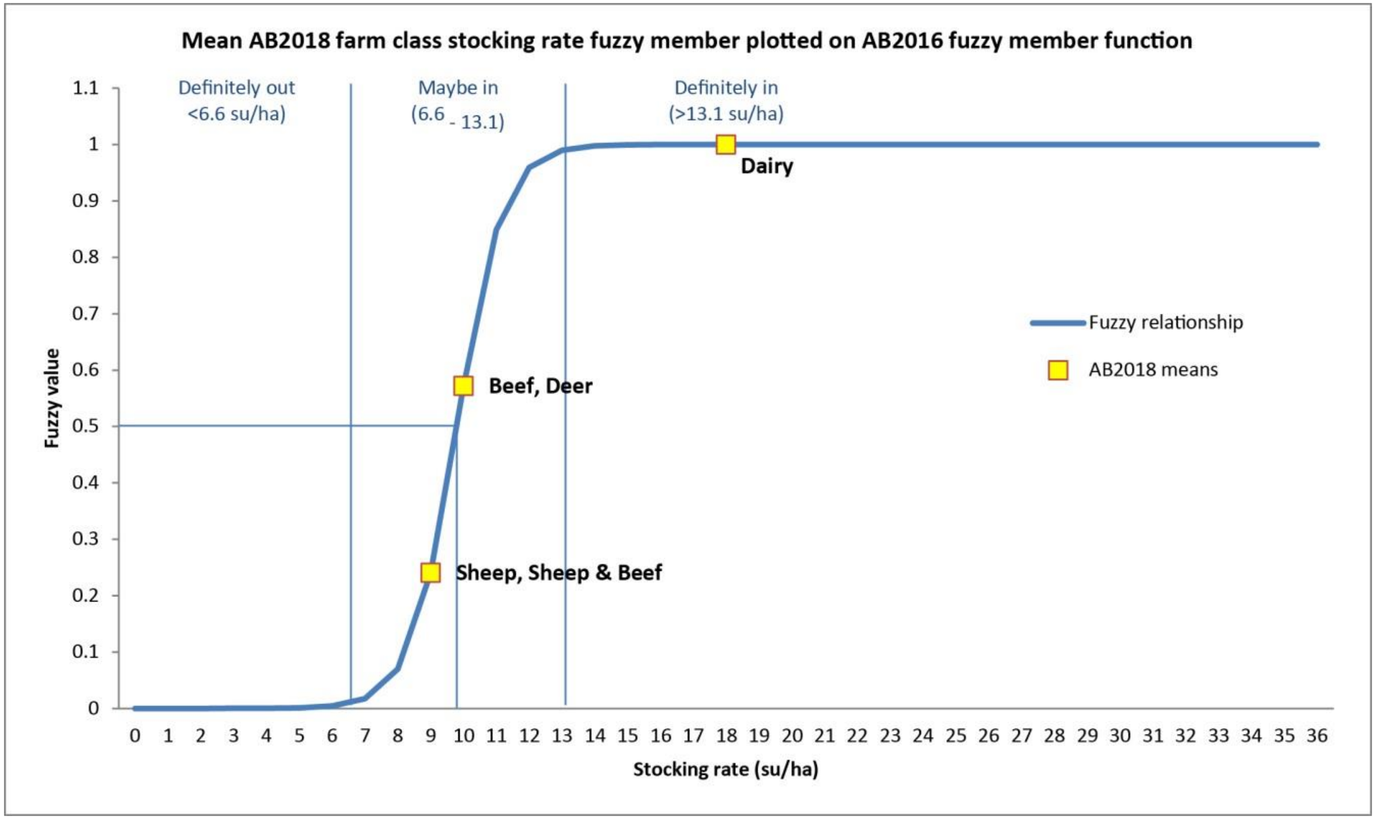
2. LUCAS LUM managed grassland classification

- "Fuzzy logic is an expert-guided weights of evidence method useful in applications that have vague specification and/or imprecise data."
- Degrees of truth



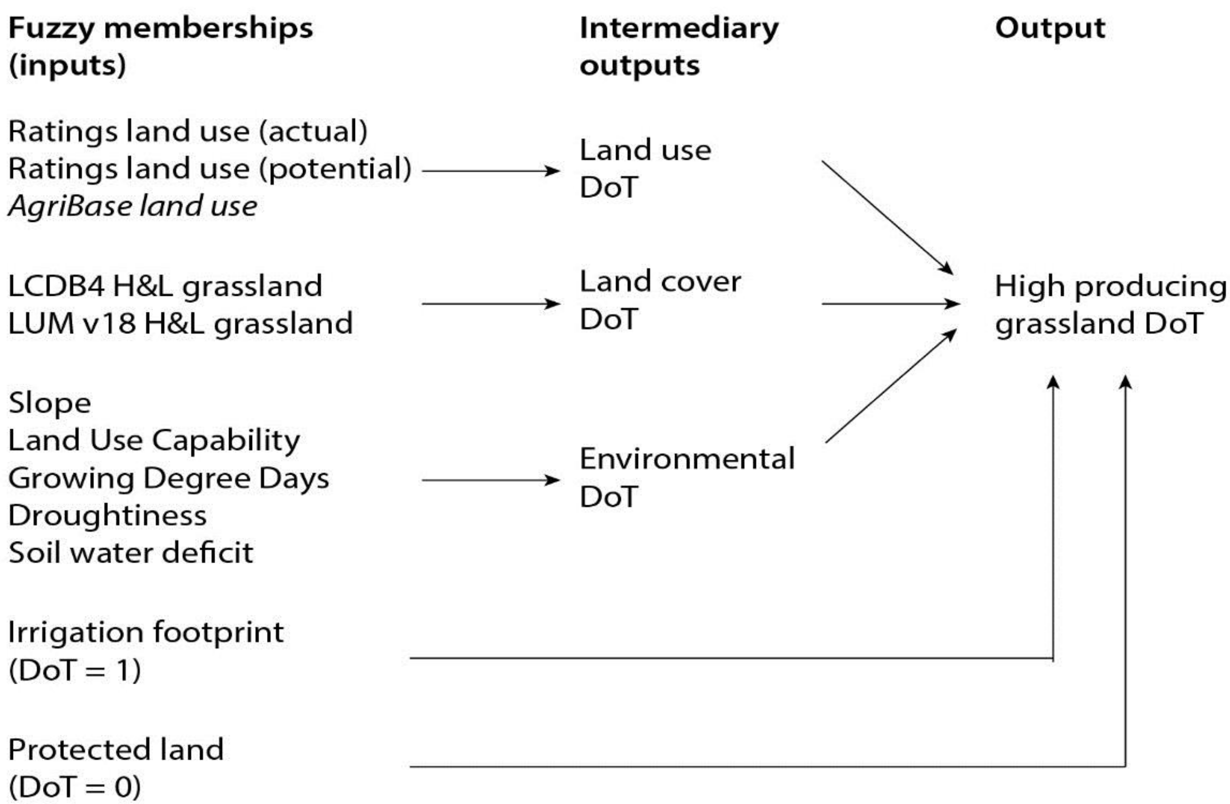
2. LUCAS LUM managed grassland classification

- One example of a fuzzy membership
- (high producing pastures are more common on farms with high stocking rates)



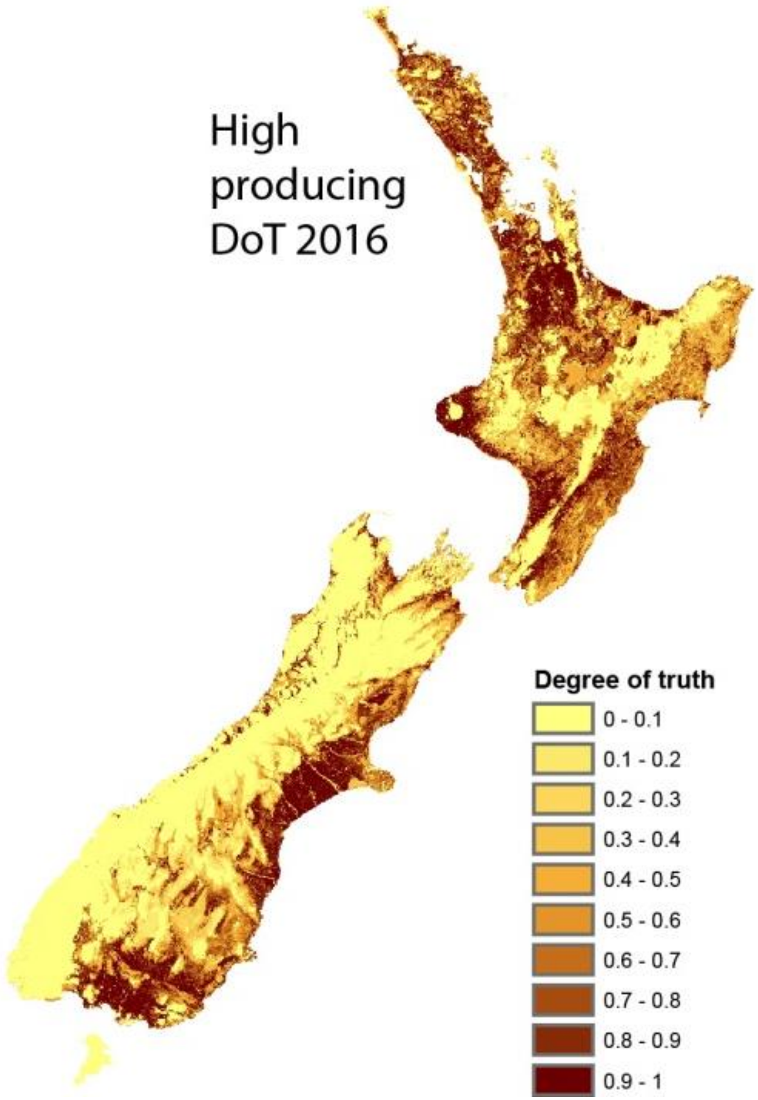
2. LUCAS LUM managed grassland classification

- 12 memberships refined to 3 intermediary memberships



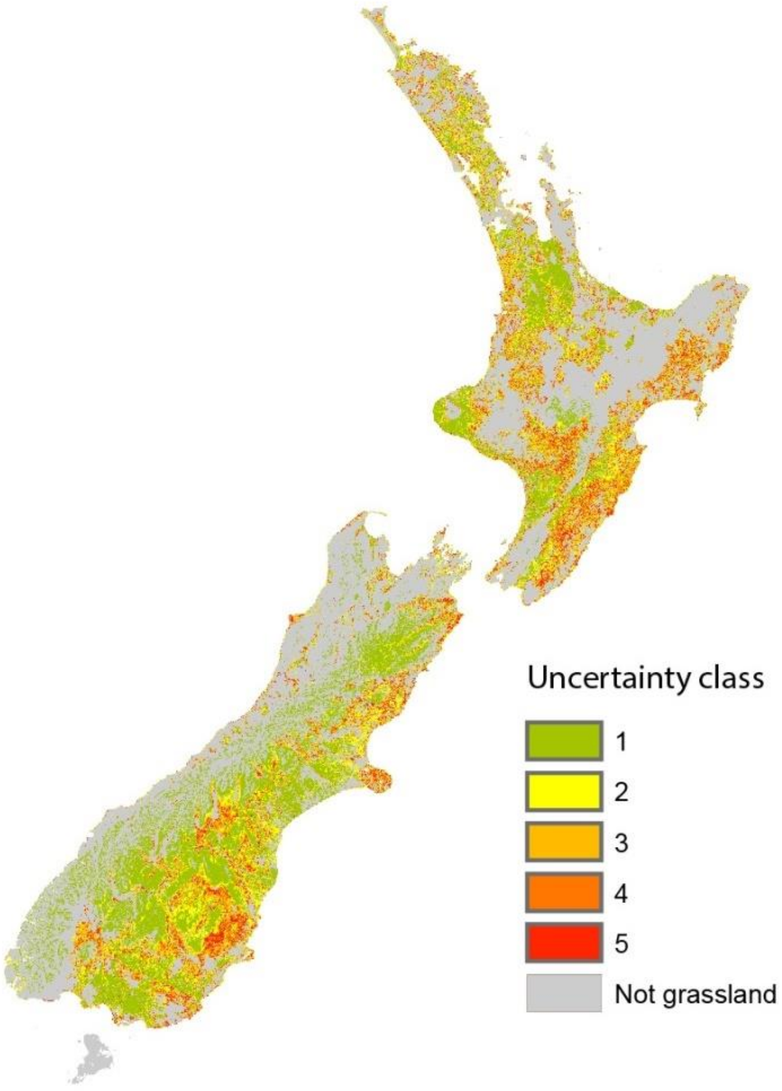
2. LUCAS LUM managed grassland classification

- Fuzzy logic result
- The degree of truth of being high producing pasture
- (likelihood or probability of being high producing pasture)



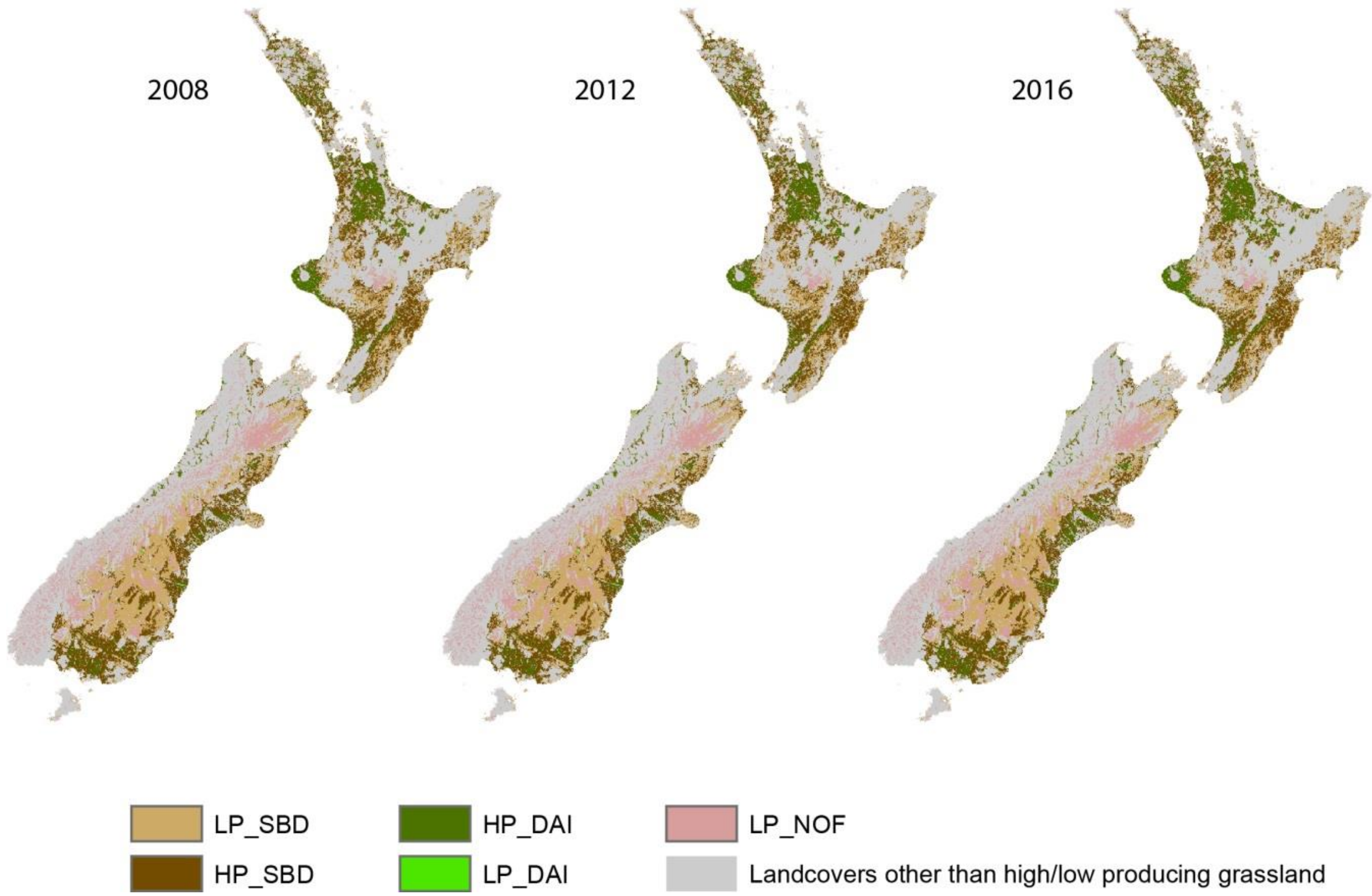
2. LUCAS LUM managed grassland classification

- Uncertainty classes
- (class 5 = highest uncertainty)



2. LUCAS LUM managed grassland classification

- Results classified and combined with simple land use



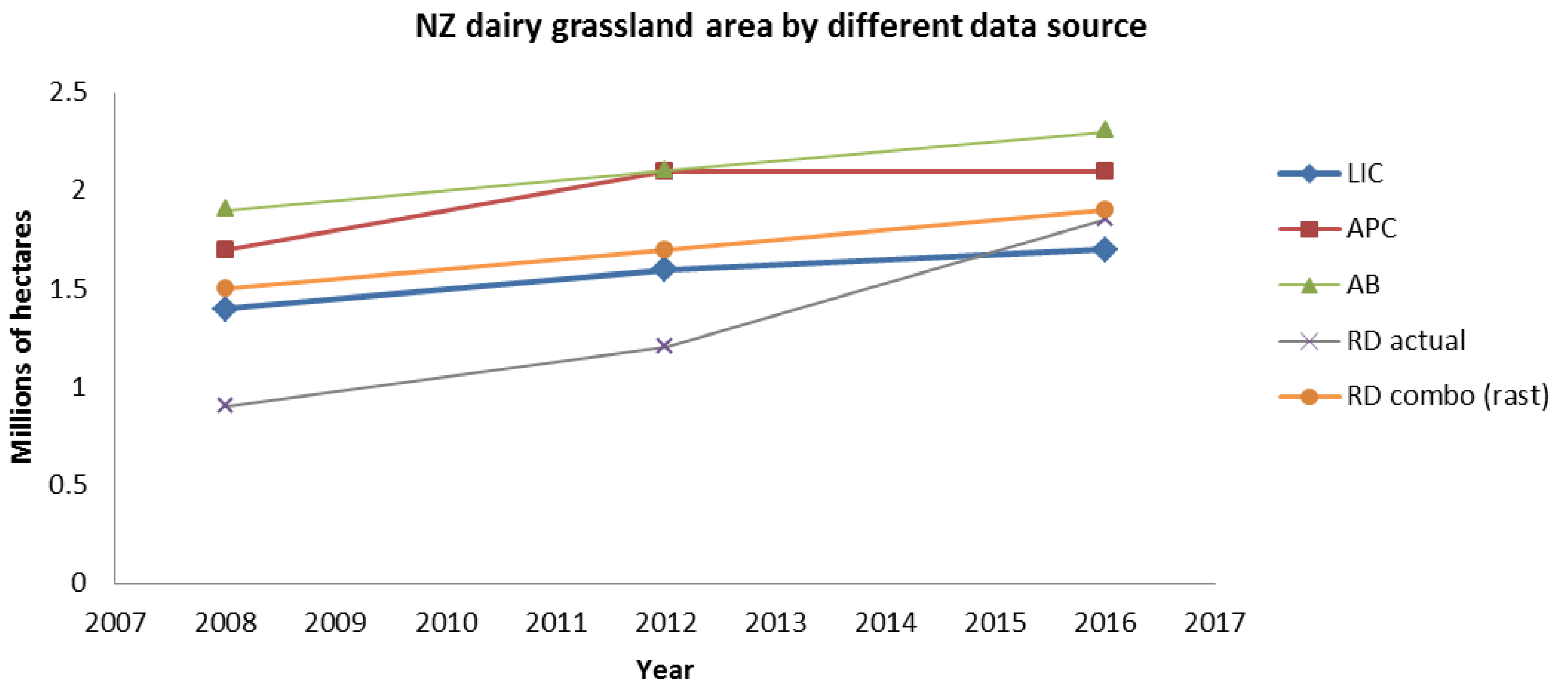
2. LUCAS LUM managed grassland classification

Results and conclusions

- FL method estimates high producing pasture for 2012 at 50% of total grassland area
  - LCDB4 estimated 67% HP
  - Previous LUM estimated 44% HP
- Differences between years was small but HP increasing*
- Improved quality and accessibility to national land use data would improve the fuzzy logic classification of high and low producing grasslands.

* Based on 2012 grassland footprint only.

2. LUCAS LUM managed grassland classification



LIC = Livestock Improvement Corporation. Total effective dairy farm hectares reported by LIC & DairyNZ (2018) for the preceding season (e.g. 2007/08 is used for 2008).
APC = Agricultural Production Census. Total 'Grass land' for Dairy Cattle Farming (ANZSIC06) for the Agricultural Production Censuses 2007, 2012, 2017
AB = AgriBase dairy cattle farming (DAI) intersected with LUM managed grasslands. LUM 2012 extent is used for 2016.
RD = Ratings data. Actual dairy land use intersected with LUM managed grasslands. LUM 2012 extent is used for 2016.
RD combo (rast) = Actual dairy land use from ratings data, plus validated dairy land use from ratings data land use category. 'RD combo (rast)' is used in the final classification. Areas are summarised from the final raster outputs. LUM 2012 extent is used for all years.

Species occupancy

Jerry Cooper

Introduction

- A proposed indicator for assessing one aspect of biodiversity status and change
  - What proportion of the potential range of a species is actually occupied and how is that changing?
- IDA: Improving the processes required to allow species occupancy to be modelled
  - The ‘pipeline’ from initial assembly of available data on species occurrence (and absence) as inputs to modelling and visualisation

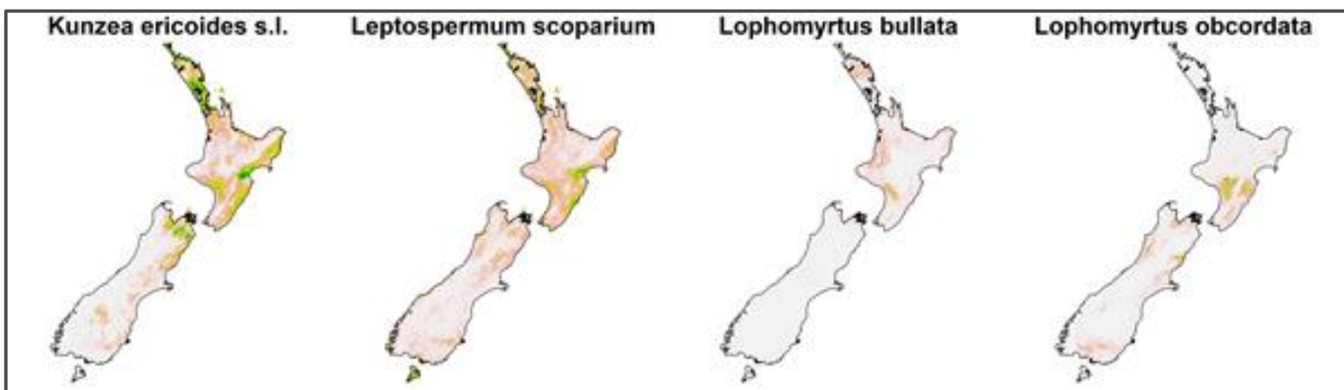
IDA contribution to Species Occupancy

1. Improving bio-data access, integration & quality

2. Online species modelling tools – are they fit for purpose?

3. Visualizing data/model outputs

4. Some test modelling of New Zealand species distributions



1: Improving bio-data access, integration & quality

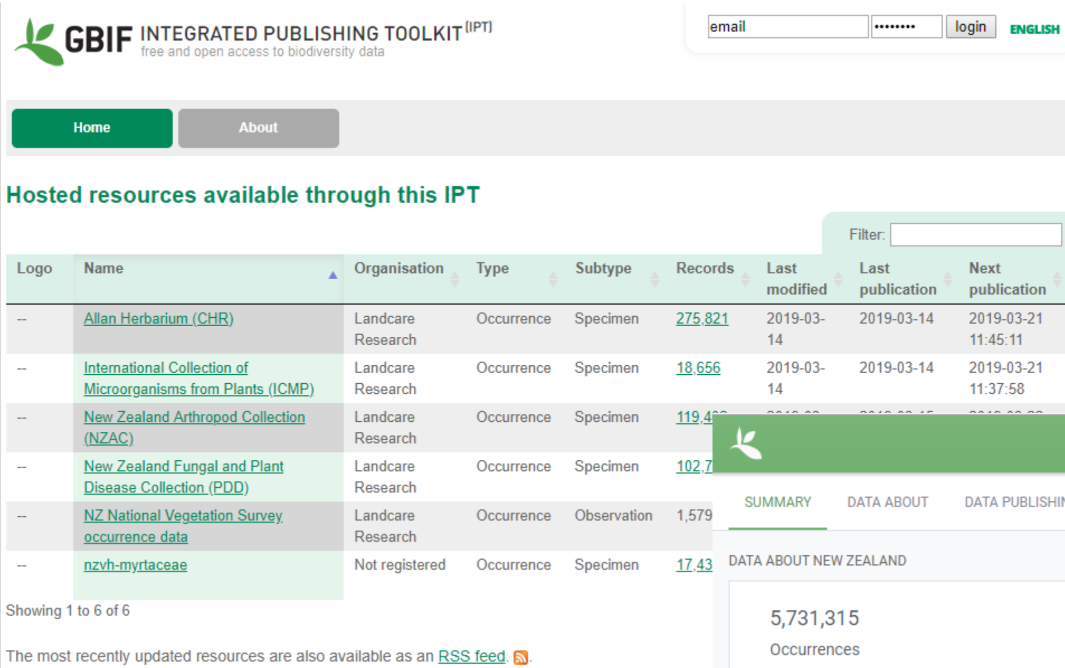
Problems:

- limited species occurrence data
- scattered across agencies/institutes
- in different formats using different collection protocols
- with varying data standards and data quality

Solutions: support & enhance existing initiatives

- The Global Biodiversity Information Facility (GBIF)
- The Atlas of Living Australia (ALA)
- The New Zealand Organisms Register (NZOR)
- Survey/Monitoring programs - the National Vegetatior (NVS) & the Nationally Significant Databases
- The rise of Citizen Science platforms (e.g. iNaturalist)
- The international biodata standards bodies (e.g. TDWG)

GBIF/ALA – the global/regional data aggregators


GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

email | login | ENGLISH

Home | About

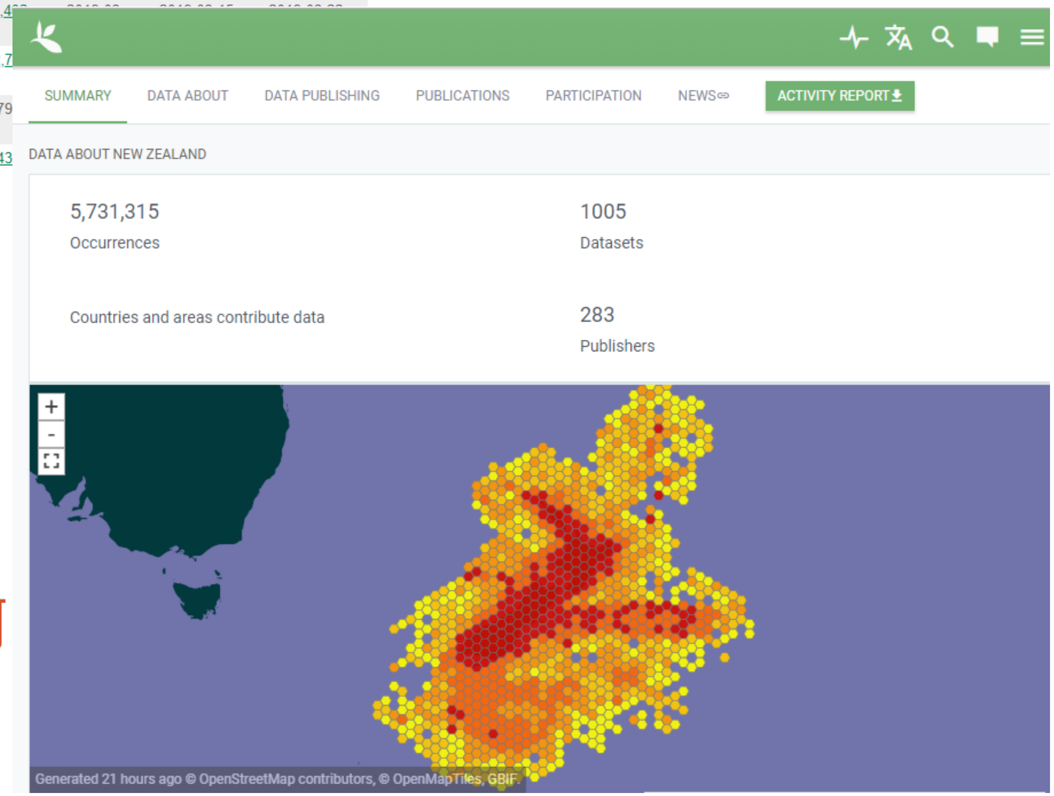
Hosted resources available through this IPT

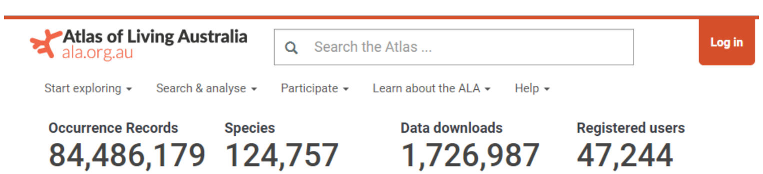
Filter:

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
--	Allan Herbarium (CHR)	Landcare Research	Occurrence	Specimen	275,821	2019-03-14	2019-03-14	2019-03-21 11:45:11
--	International Collection of Microorganisms from Plants (ICMP)	Landcare Research	Occurrence	Specimen	18,656	2019-03-14	2019-03-14	2019-03-21 11:37:58
--	New Zealand Arthropod Collection (NZAC)	Landcare Research	Occurrence	Specimen	119,4	2019-03-	2019-03-	2019-03-
--	New Zealand Fungal and Plant Disease Collection (PDD)	Landcare Research	Occurrence	Specimen	102,7			
--	NZ National Vegetation Survey occurrence data	Landcare Research	Occurrence	Observation	1,579			
--	nzvh-myrtaceae	Not registered	Occurrence	Specimen	17,43			

Showing 1 to 6 of 6

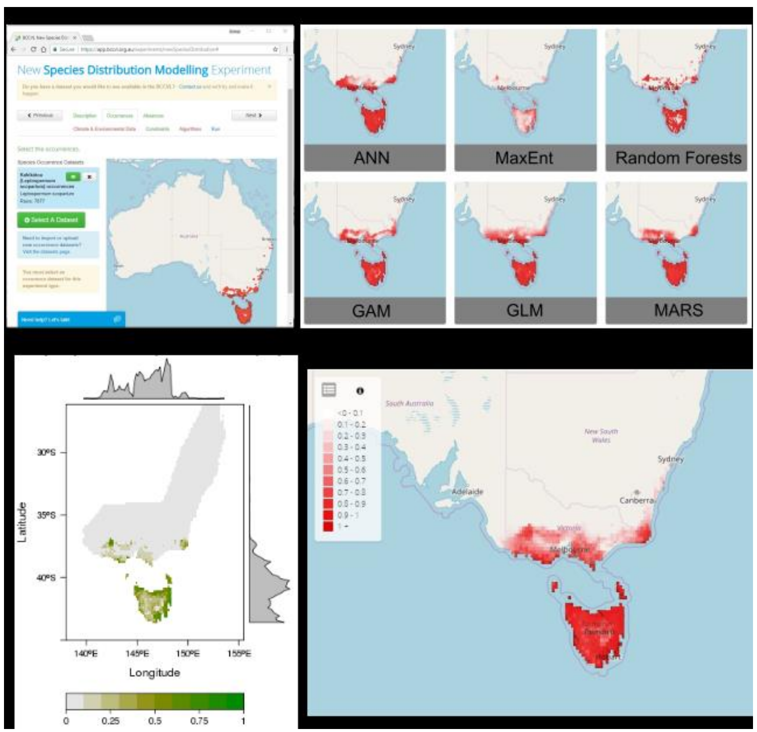
The most recently updated resources are also available as an RSS feed.





2: Online species modelling platforms

- Species occurrence data can be used to generate **species distribution models** by combining with environmental data – rainfall, altitude, soil chemistry …
- We reviewed some ‘point & click’ online toolboxes
- **Conclusions**: Easy to use – but easily abused by the inexperienced



Australia:
The Biodiversity and Climate Change Virtual Laboratory (BCCVL)

Links:

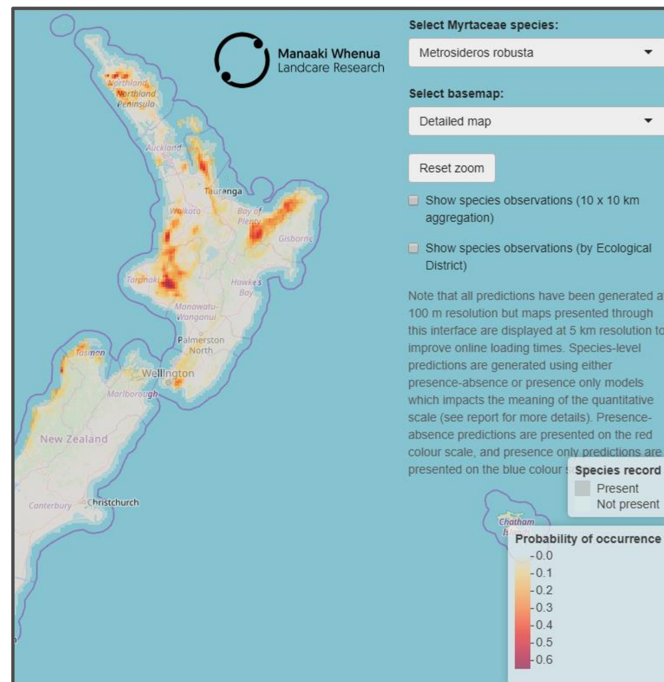
- Occurrence data held by the ALA
- To numerous modelling tools
- Running on the Au high performance computing resources

3 & 4: Species distributions, modelling, visualisation

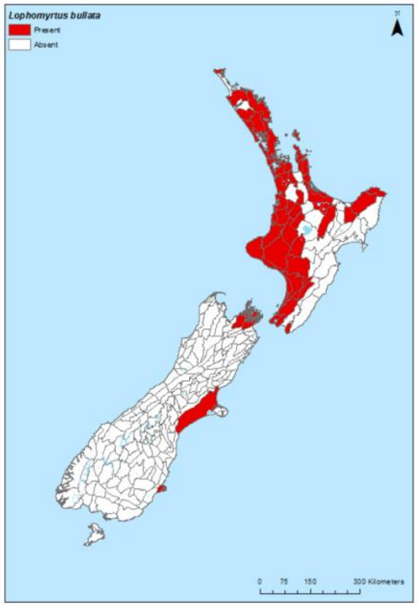
- Myrtle rust reported in NZ in April 2017
- We re-purposed the IDA pipeline
- Within a week initial maps of the distribution/abundance of native myrtaceae species for DOC
  - To support targeted seed-banking
  - To inform disease-spread models

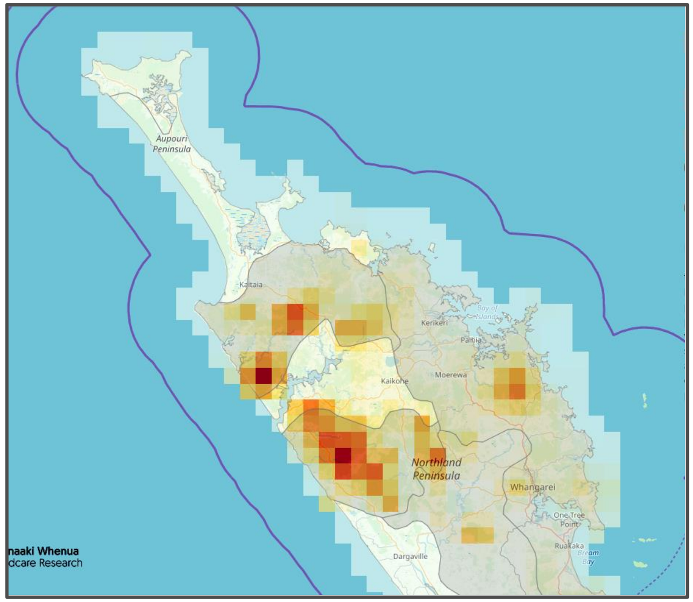
- We went on to produce species distribution models, including the first potential range maps for recently described species in *Kunzea* (funded by MPI/DOC)

- We have R-Shiny apps to visualise the models and underlying data.



Example: From species occurrences to models

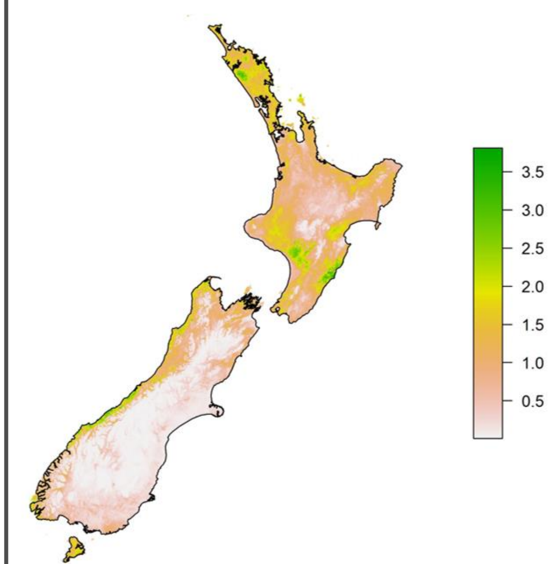




AUC – 0.92

Important predictors:
1. Growing season heat index (27%)
2. Mean annual humidity (15%)
3. Winter/summer precip. ratio (14%)

4. Remaining variables (44%)



Lophomyrtus bullata

- Modelled using boosted regression trees
- Many environmental layers
- Model at 100m but degraded for visualisation
- Online R-Shiny App

James McCarthy – Manaaki Whenua

IDA biodiversity – where next?

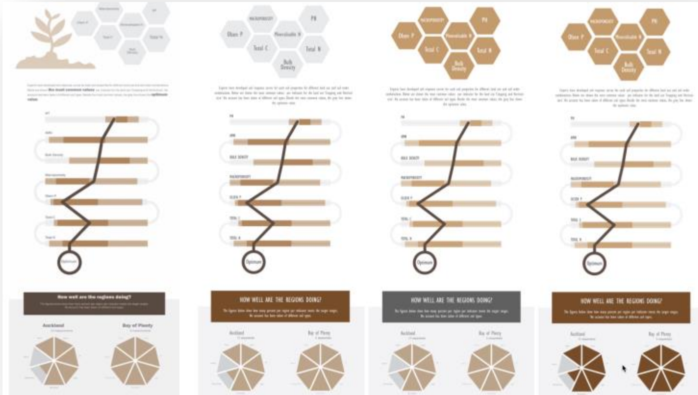
- Manaaki Whenua now has increased capacity for species modelling - supporting both conservation and biosecurity needs
- But … their work is contingent on having adequate and accessible baseline biodiversity data
- **NZ needs a financially supported national bio-data infrastructure**
    - Nationally Significant Collections & Databases – under review. Meanwhile capability eroded due to flat funding and rising costs
    - National coordination between data-holders does not exist
    - Technical expertise exists in NZ but is capacity-limited, and ageing!
    - NZOR supported by MPI & DOC – currently at least
    - Some key technical components could be adopted – e.g. ALA
    - GBIF NZ is not financially supported – but we signed the agreement
    - iNaturalistNZ survives on occasional project funding
    - Short-term project funding for data-science is not a solution

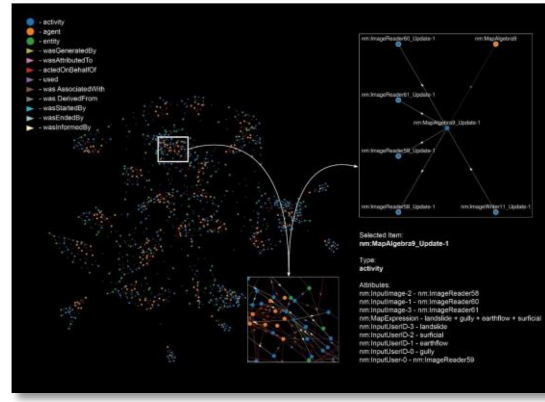
Supporting IDA

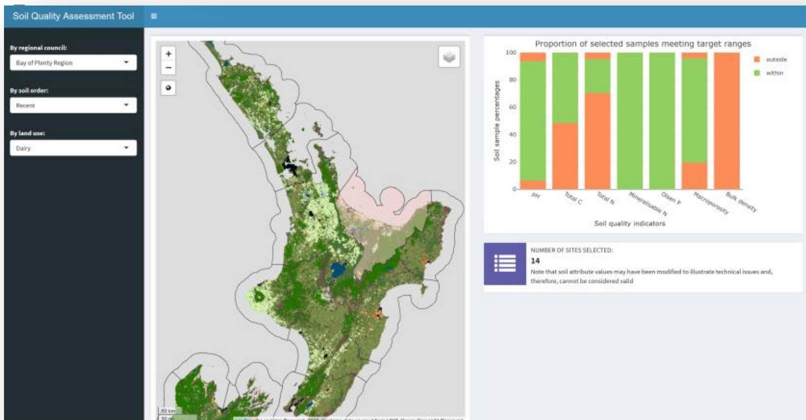
David Medyckyj-Scott

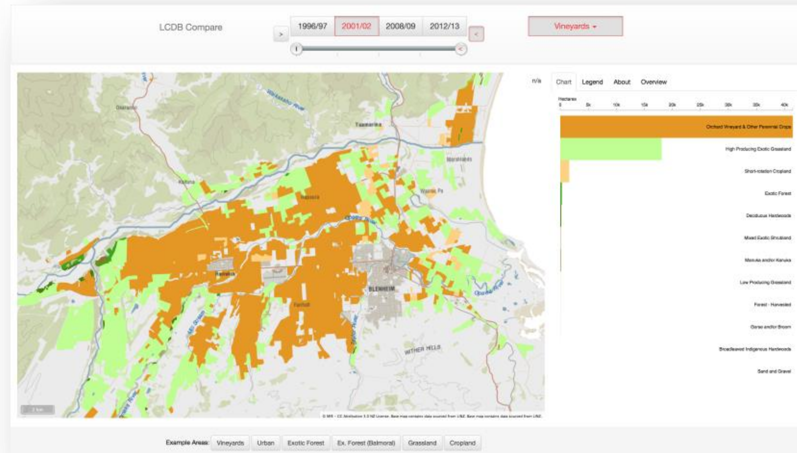
Activities and outputs

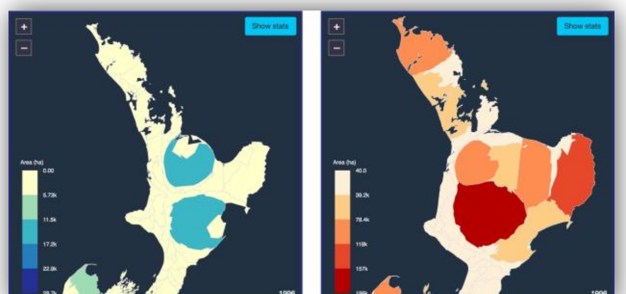
- Technology, processes, pipelines and tools e.g. validation to integrate, harmonise and standardise heterogeneous land resource and biodiversity datasets *(e.g. pyLUC, taxon scrubber, geovalidation tools)*
- Multidimensional database (*review of data cubes*, ➲*Discrete Global Grid System*)
- New ways to present and share data on state and trend *(visualizations, tools, standards, APIs)*

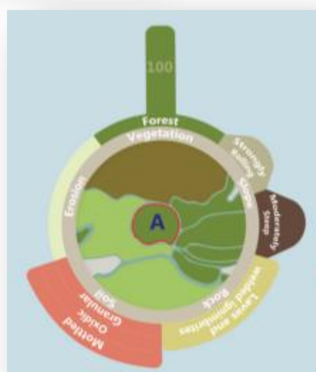












https://vizdemo.landcareresearch.co.nz/#about

Activities and outputs

- Technology, processes, pipelines and tools e.g. validation to integrate, harmonise and standardise heterogeneous land resource and biodiversity datasets *(e.g. pyLUC, taxon scrubber, geovalidation tools)*
- Multidimensional database (*review of data cubes*, ➲ *Discrete Global Grid System*)
- New ways to present and share data on state and trend *(visualizations, tools, standards, APIs)*
- Improvements in environmental data management practice/data science *(DOIs for data, provenance in modelling systems, ➲ best practice documents)*
- Use of best practices and standards to integrate environmental data *(standards, vocabulary services, ontologies, Linked Data, OGC ELFIE interoperability experiment)*
- (Multi-indicator) environmental data infrastructures *(POC, OGC SoilIE interoperability experiment, social architectures)*
- Outreach and capability building (*Environmental Data Summit, LINK Seminars*)

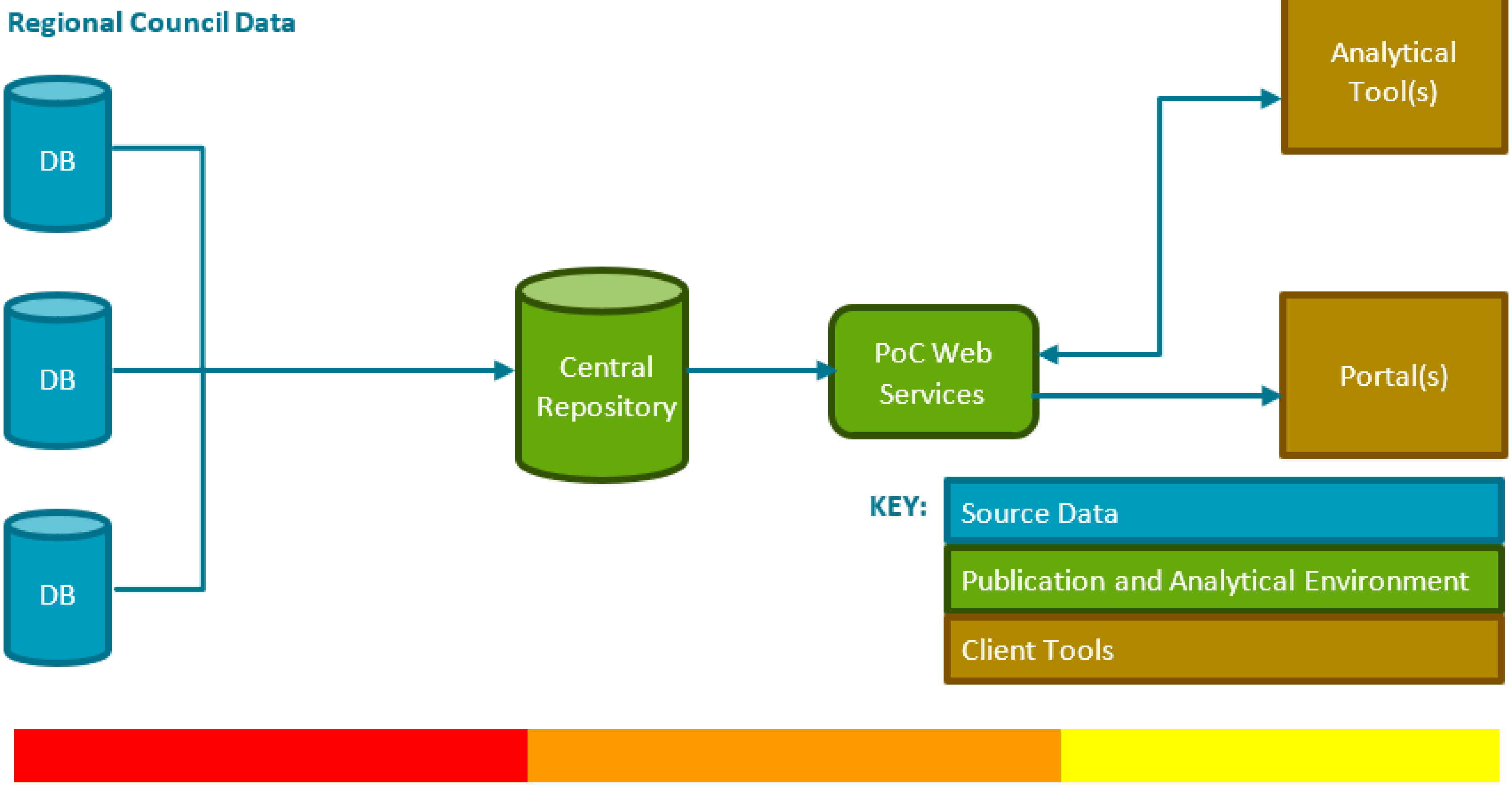
Soil Quality Data Case Study

Data harmonisation case study using a data set that exposes the range of needs, conditions and issues faced when aggregating data for analysis and reporting

Soil quality data are

- Fundamental data set for State of the Environment reporting
- Collected, stored and maintained by a disparate set of agencies for both data management and analytical reasons
- Stored and maintained separately but are functionally a single, logical data set – clear need for consistent management
- No history of coordinated, nationally consistent, capture and management of data, but widespread recognition of the need
- Technology and processes required to implement the case study should be appropriate to other environmental domains

Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept

Regional Council Data

DB

DB

DB

Data Aggregation

- 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)
- undeclared changes of units of measure for analyses, often in the same column of a single spreadsheet
- missing, or ambiguous, laboratory method metadata
- missing site locations (not usable without human intervention)

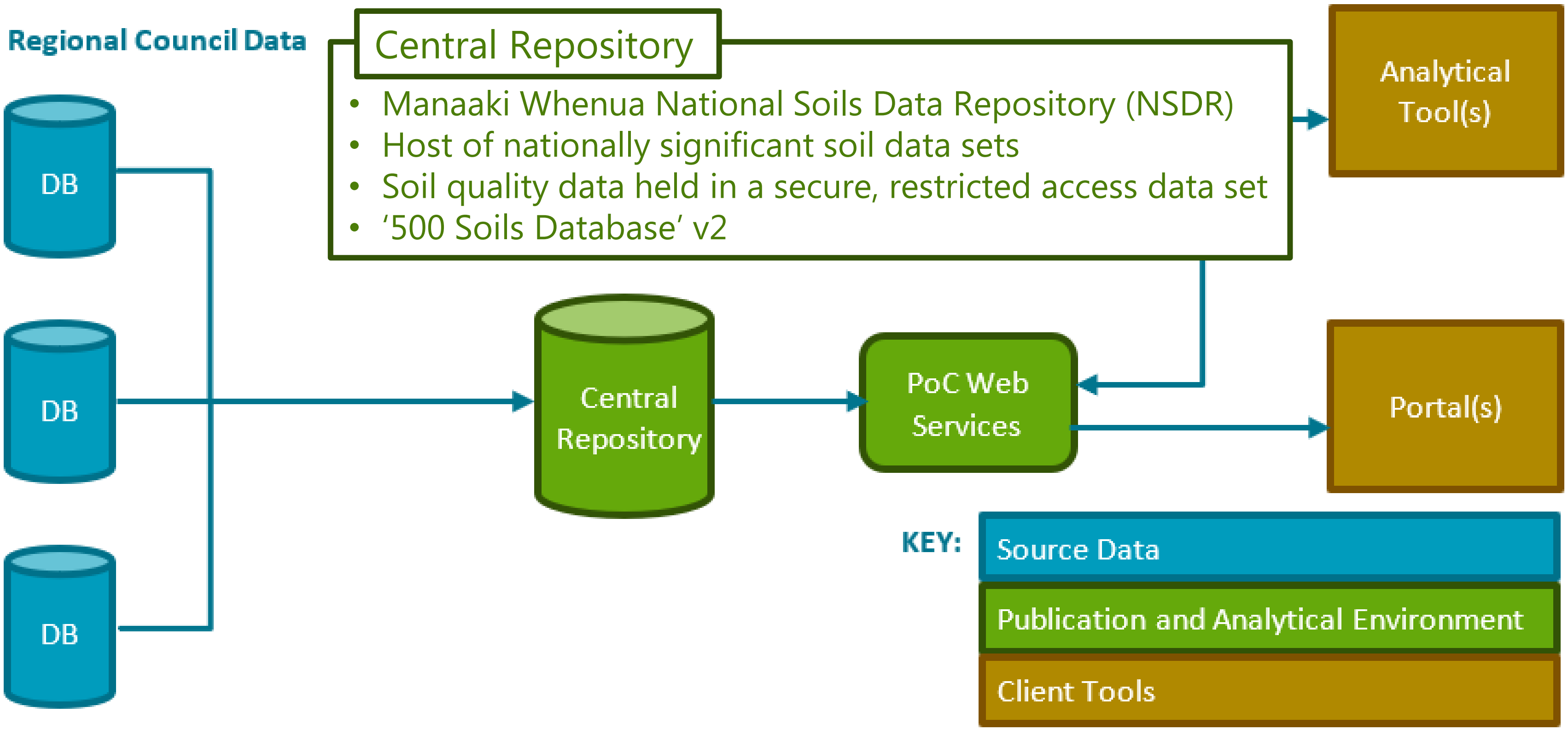
datatamer
HILLTOP SOFTWARE

Analytical Tool(s)

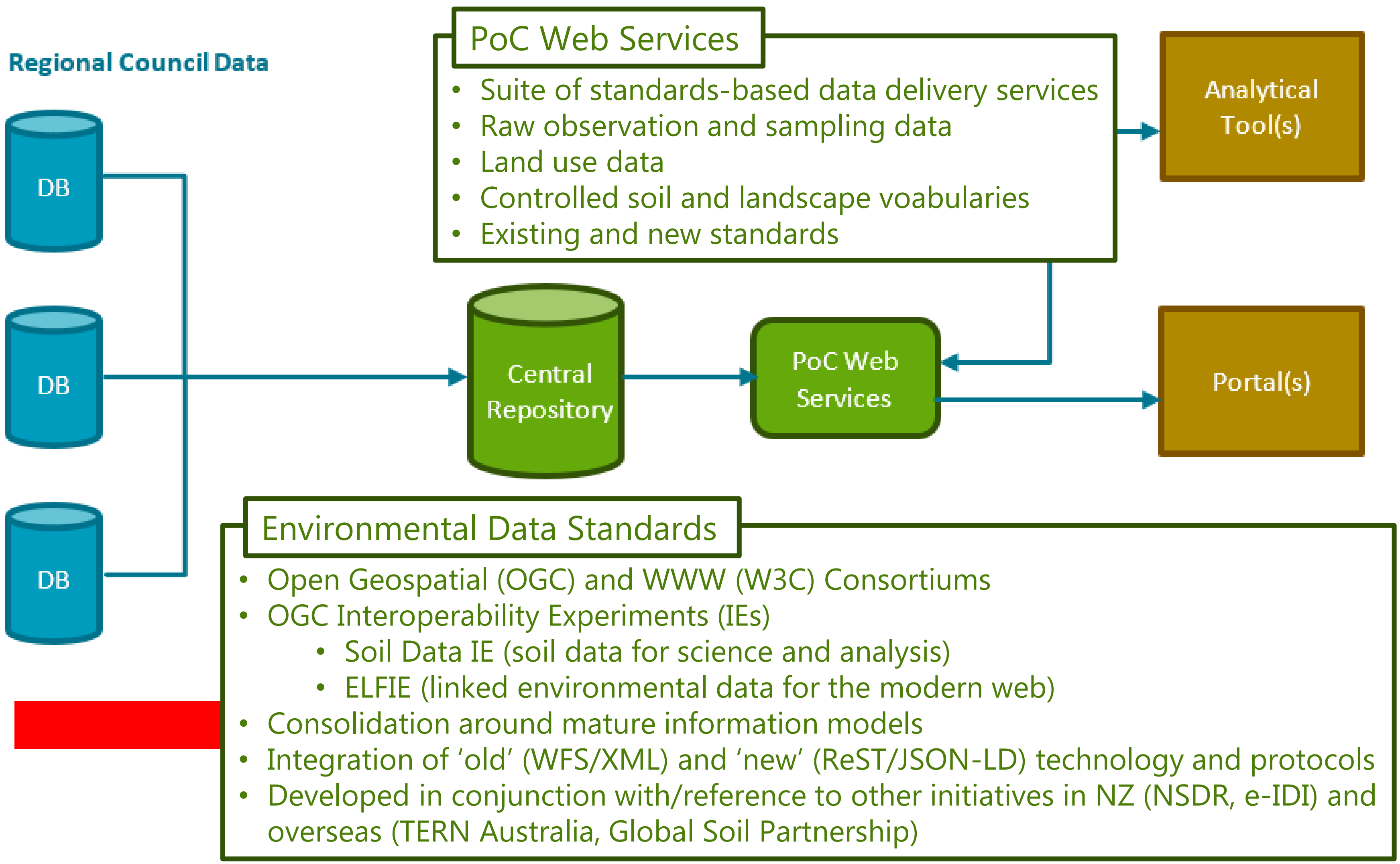
Portal(s)

cal Environment

Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept

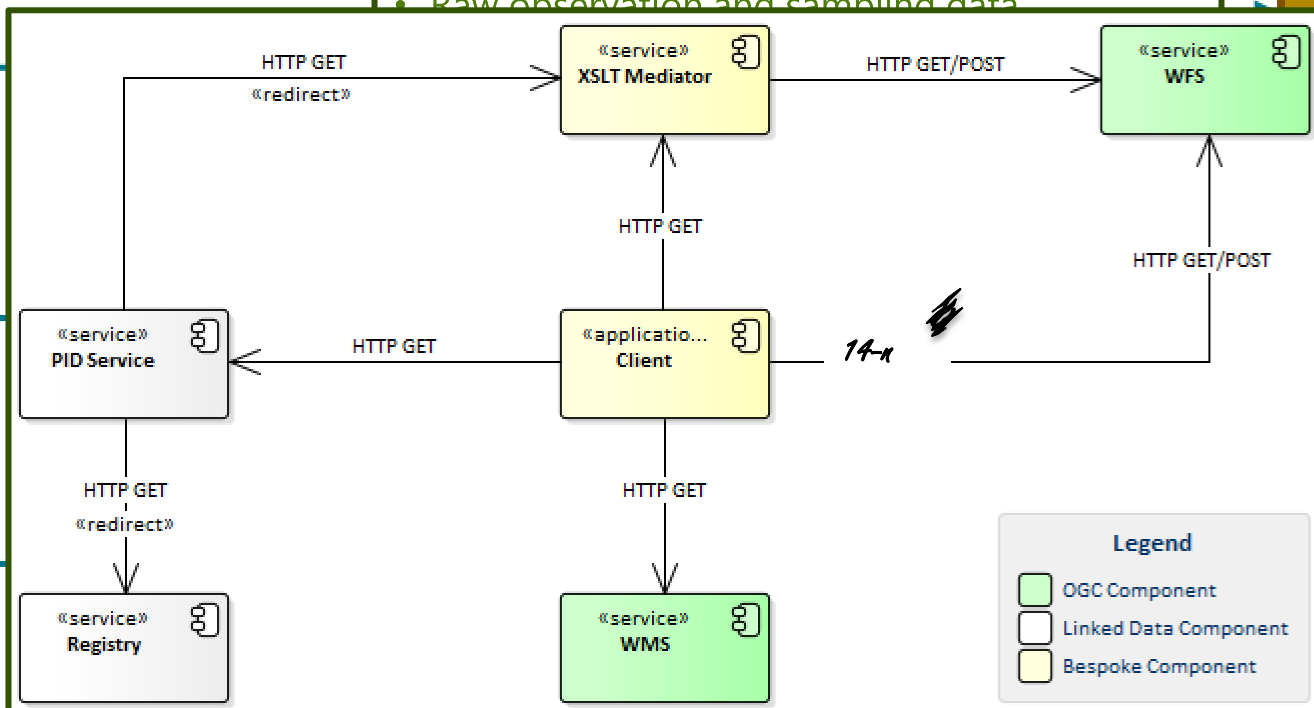
Regional Council Data

PoC Web Services

- Suite of standards-based data delivery services
- Raw observation and sampling data

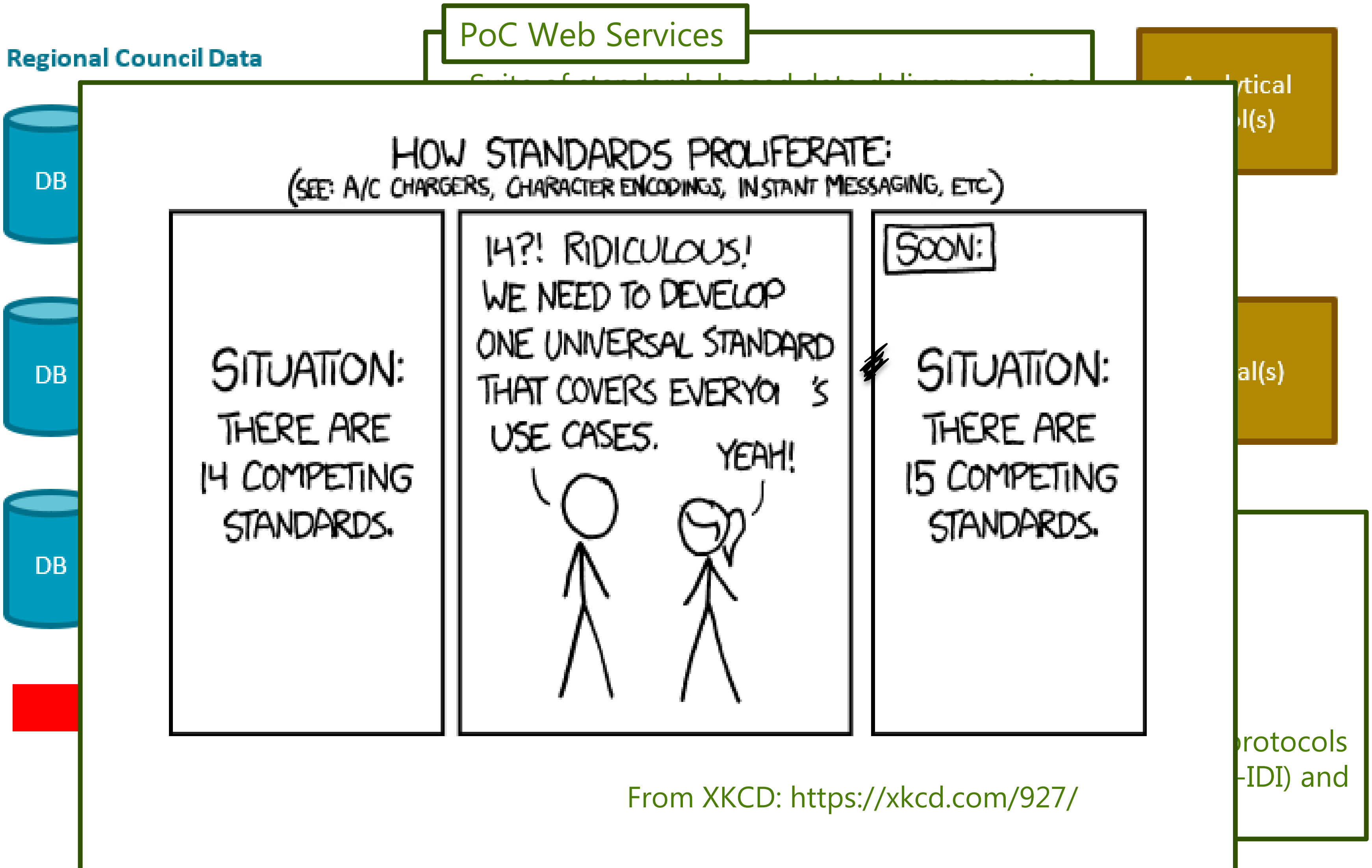
Analytical Tool(s)

Portal(s)



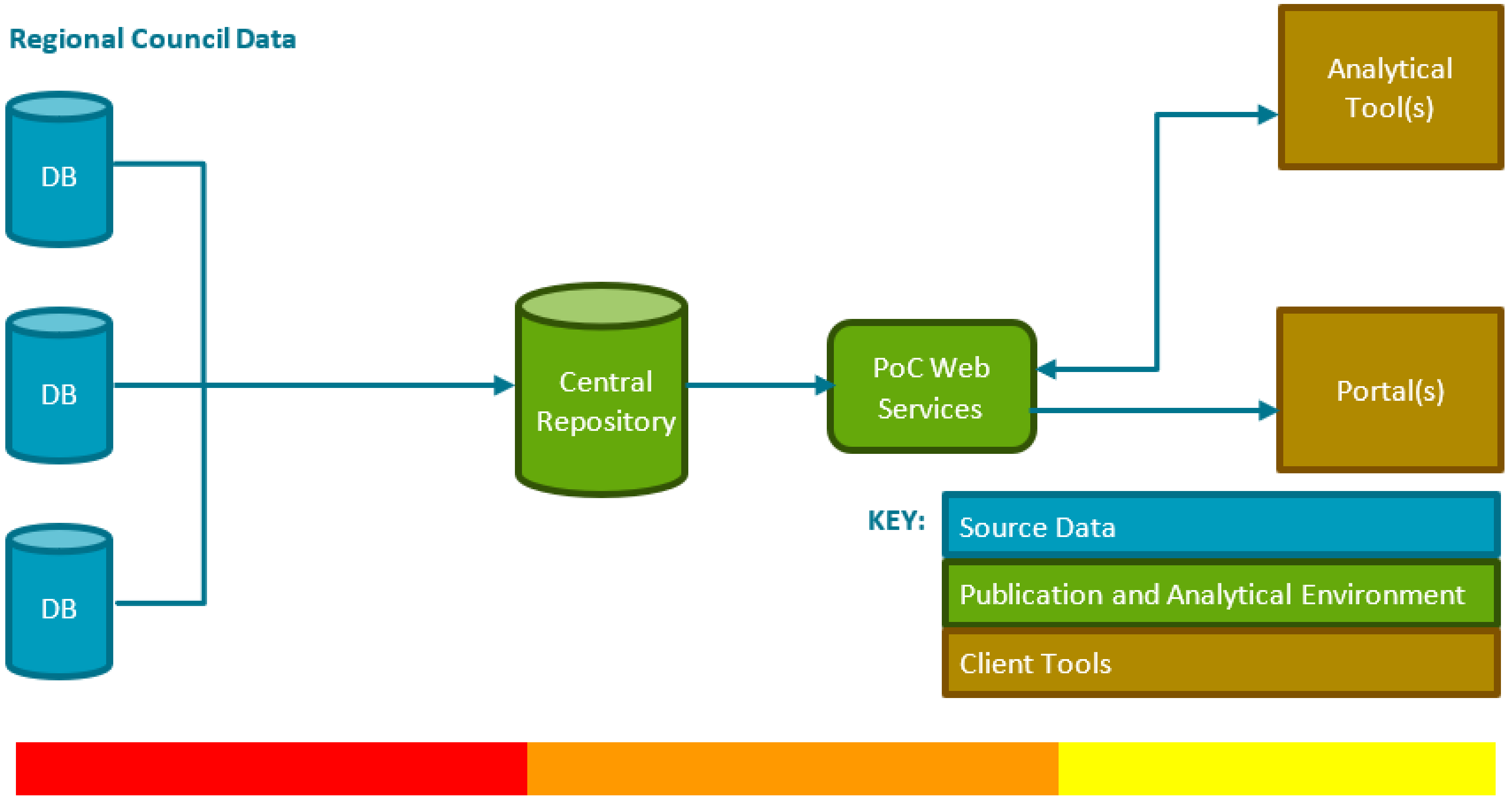
- Consolidation around mature information models
- Integration of 'old' (WFS/XML) and 'new' (ReST/JSON-LD) technology and protocols
- Developed in conjunction with/reference to other initiatives in NZ (NSDR, e-IDI) and overseas (TERN Australia, Global Soil Partnership)

Soil Quality Data Infrastructure Proof of Concept

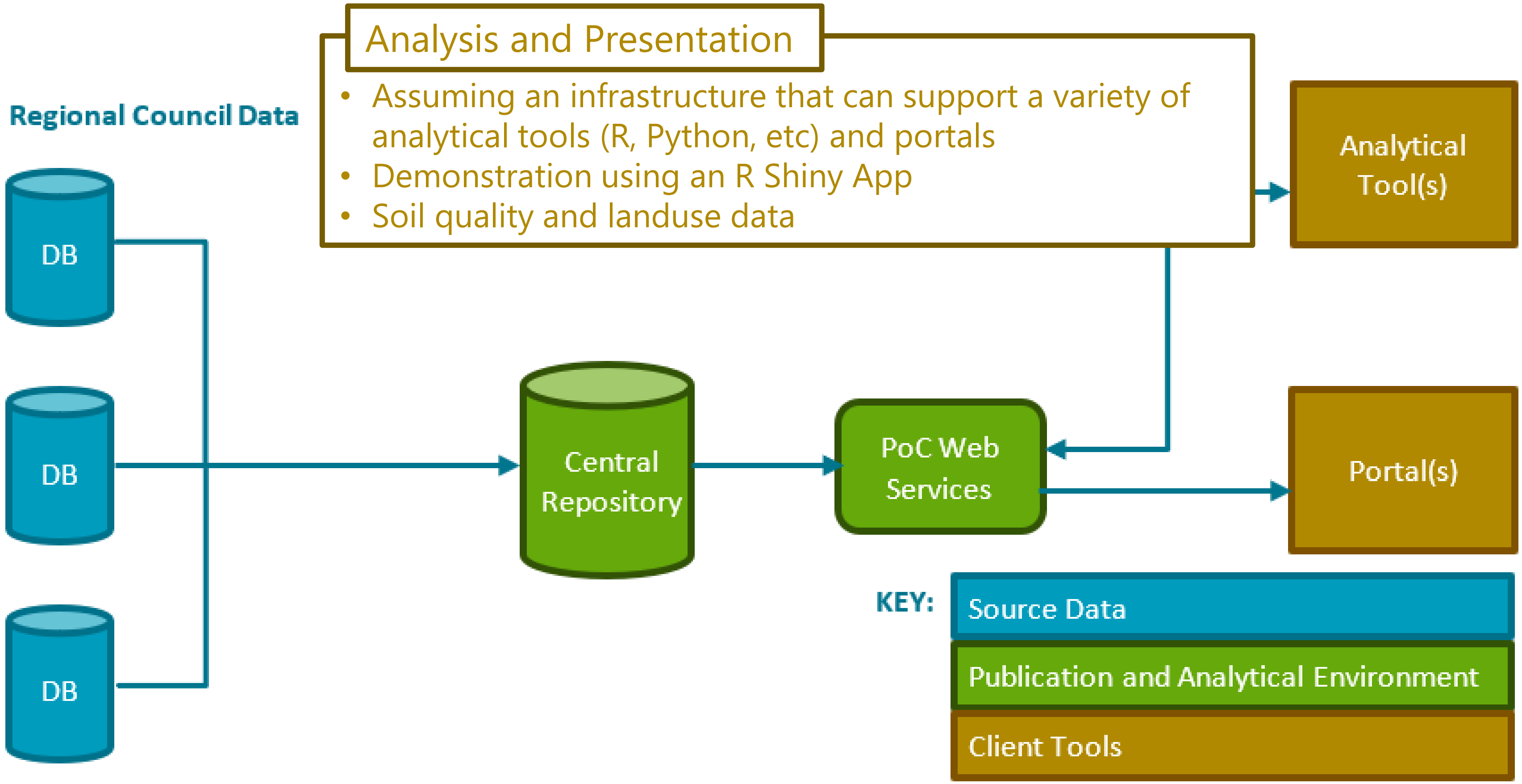


From XKCD: https://xkcd.com/927/

Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept

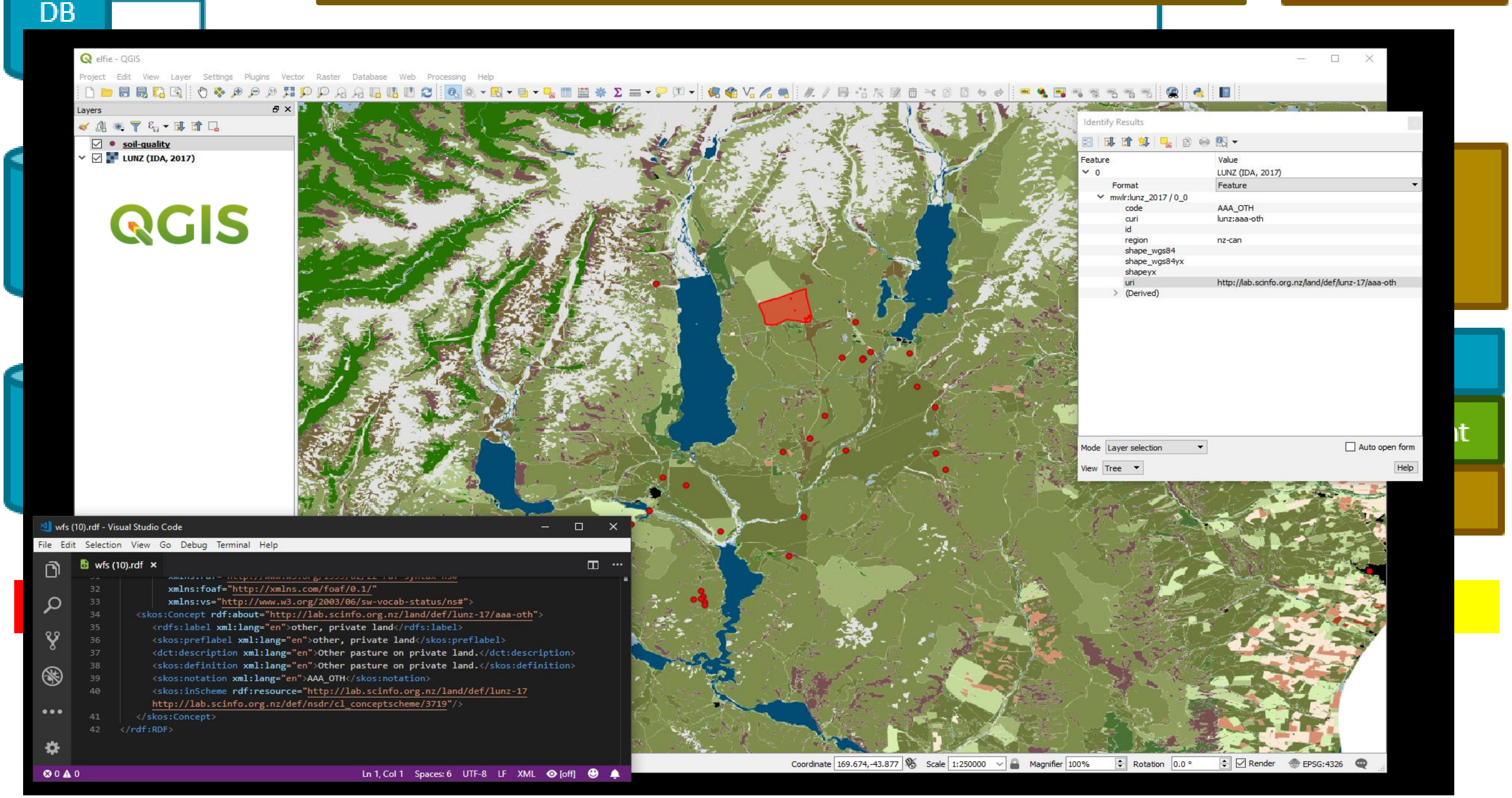
Regional Council Data

DB

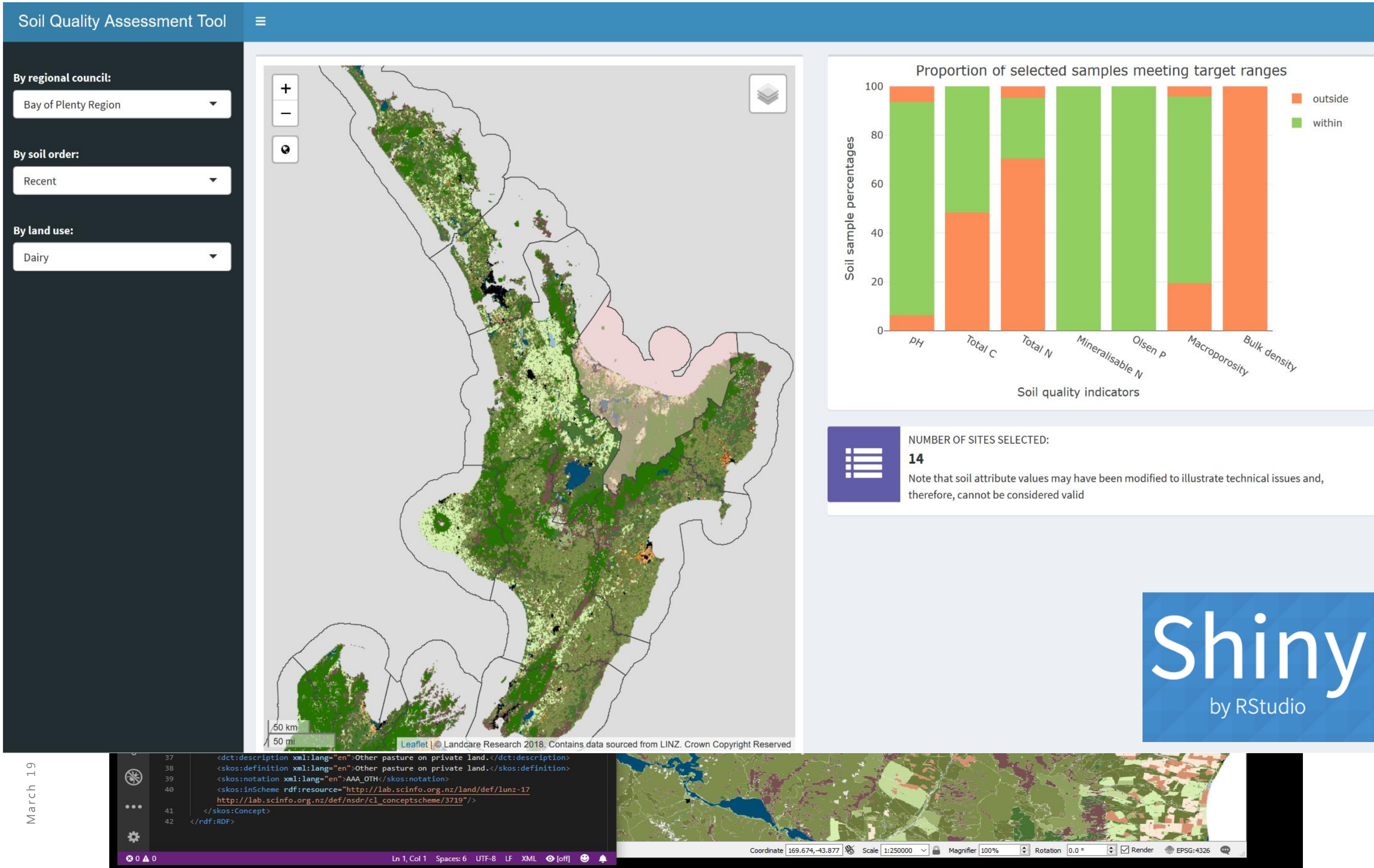
Analysis and Presentation

- Assuming an infrastructure that can support a variety of analytical tools (R, Python, etc) and portals
- Demonstration using an R Shiny App
- Soil quality and landuse data

Analytical Tool(s)



Soil Quality Data Infrastructure Proof of Concept



Soil Quality Data Infrastructure Proof of Concept

Conclusions

Regional Council Data

Analytical Tool(s)

Source Data

- Need significantly improved data storage and management
- Well defined specifications for structure, content and maintenance of data
- Distributed data set but needs to be treated as a single national data set
- Well defined, unchanging identifiers for sampling locations
- RC's need support

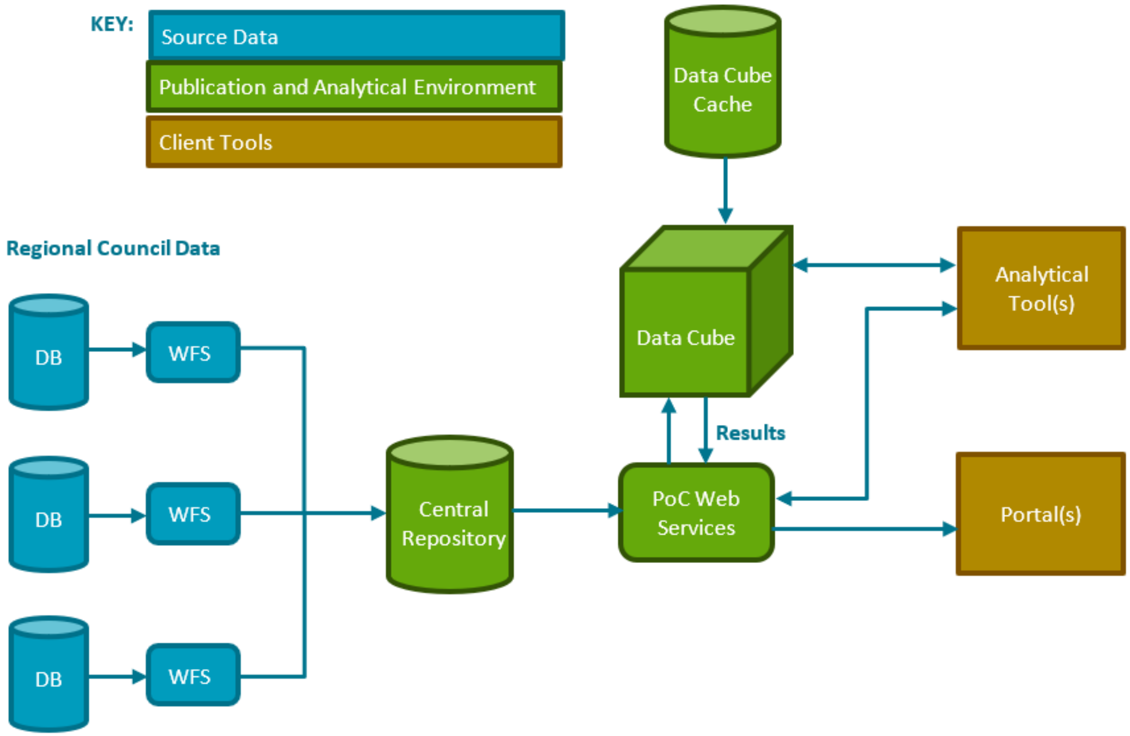
Publication/Analysis

- Have the standards, protocols and tools needed
- Need community agreement/will to use them
- Some protocols and formats are end of life
- Provenance, uncertainty, repeatability
- Security and access control – ownership and trust
- Need clarity on how data are federated
- HPC analytical environments

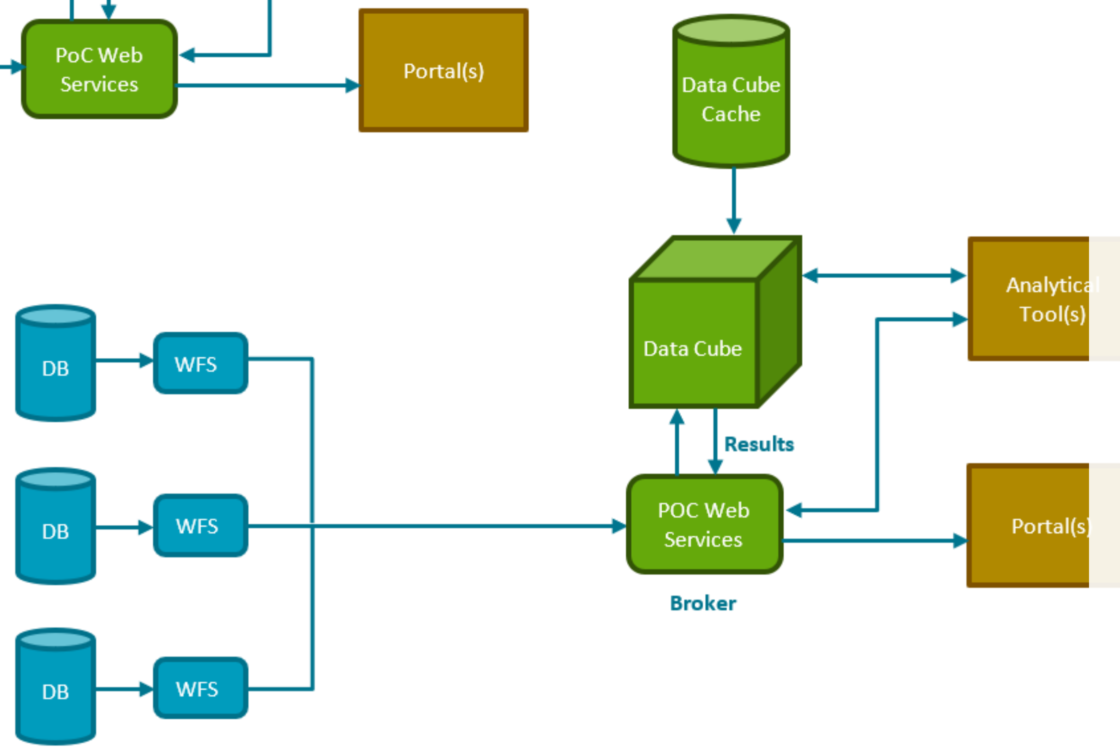
Tools/Portals

- Pretty good shape
- Could be green now
- Need the rest of the infrastructure in place and stable then …
- Potential for a varied, vibrant and robust science and reporting

Soil Quality Data Infrastructure – Where Next?



- Next generation standards - e.g WFS 3.0
- HPC/Datacubes
- Federated data infrastructure



Soil Quality Data Infrastructure – Where Next?

'***The PoC was a qualified success.*** *It proved that a set of web data services could be deployed to provide raw data for analysis […and…] shows that multi-domain / multi-indicator infrastructure, at least for the solid earth, is achievable.*

'***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.***'

From Ritchie et al (2019), Manaaki Whenua Contract Report LC3396

Soil Quality Data Infrastructure – Where Next?

'***The PoC was a qualified success.*** *It proved that a set of web data services could be deployed to provide raw data for analysis […and…] shows that multi-domain / multi-indicator infrastructure, at least for the solid earth, is achievable.*

'***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****.*'

From Ritchie et al (2019), Manaaki Whenua Contract Report LC3396

Enduring value activities

- **Know-how**
  - IDA website *(https://www.landcareresearch.co.nz/science/e-science/ida)*
  - Presentations etc e.g. todays LINK Seminar
  - Todays workshop with key stakeholders
  - Engagement with stakeholders e.g. GBIF secretariat, MfE
  - Publications and reports e.g. ELFIE Technical Engineering report
  - *Workshop – Trends in environmental data management*
- **Data**
  - Available through MW's online services (IP, privacy, etc permitting)
- **Technology**
  - Pipelines, processing, models tools part of MW BAU activities
  - Soil – continued standards work and engagement in FAO GSP and regional soil systems activities
  - Land use – pyluc, LUMASS extensions and visualisation tools available
  - Looking at use of data cubes, provenance, APIs, linked data, DGGS, in future projects and services
  - *Workshop - Next Generation Environmental Data Sharing – Achieving Harmonisation*