Innovative Data Analysis programme:
an update on progress

Anne-Gaelle Ausseil, David Medyckyj-Scott, Ben Jolly, Andrew Manderson

(Jerry Cooper, Andrew McMillan, Bryan Stevenson, Jim Payne, Alistair Ritchie, Daniel Rutledge)

Landcare Research
Manaaki Whenua

Introduction

Data
- State
- Trend

Processing
- Analysis
- Interpretation
- Uncertainty

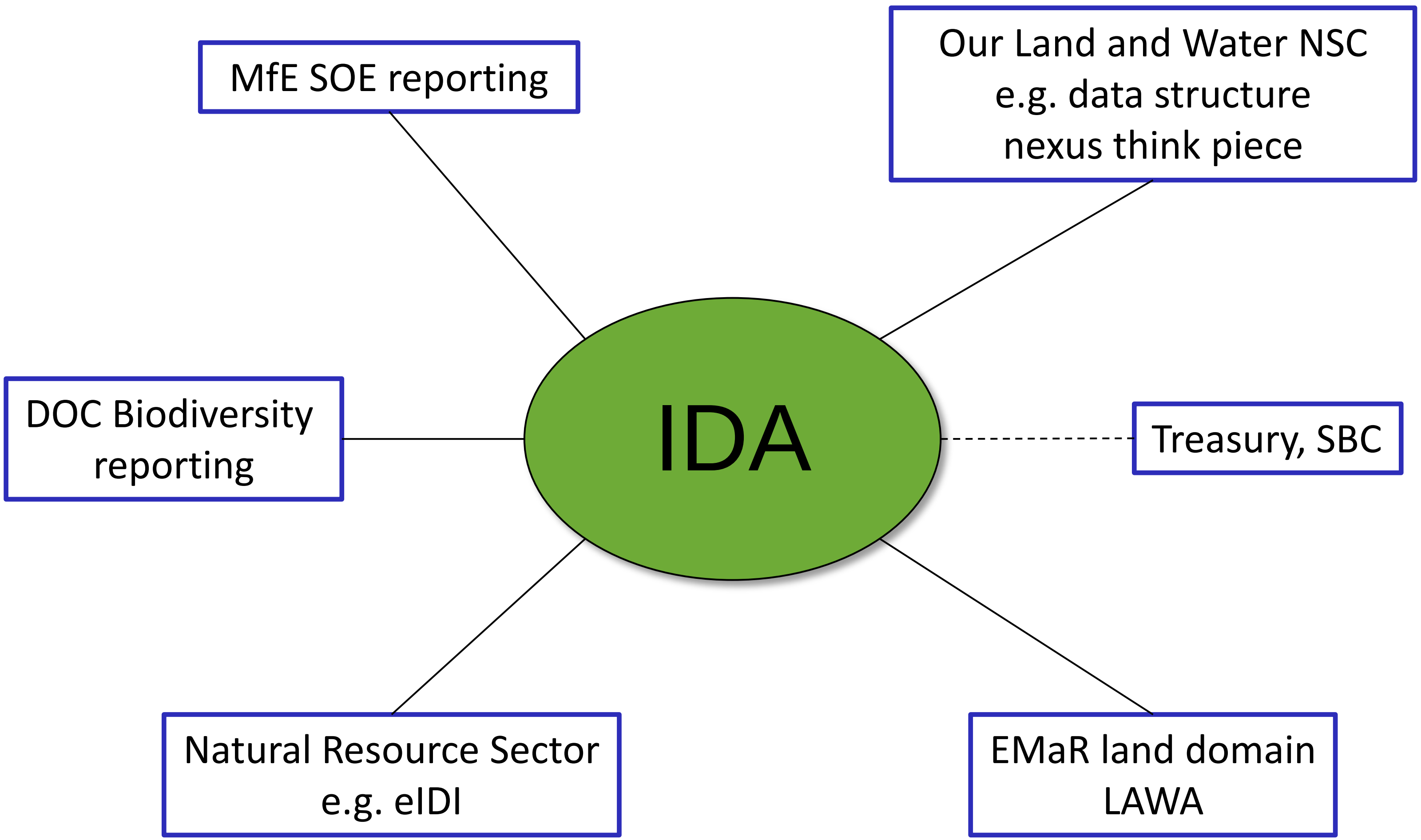
Delivery
- Access
- Visualization
- Inter-operability

Impact
- Reporting
- Policy formation
- Practice

- Bringing together heterogeneous spatial data
- Analysing data and modelling indicators
- Characterising provenance, quality, uncertainties, workflow
- Visualising and delivering data

3 domains: land use, soil health, species occupancy

IDA in relation to other initiatives



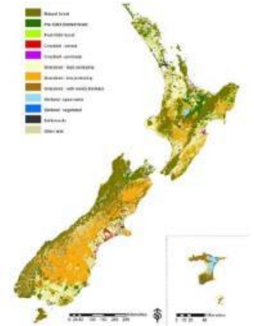
IDA work programme

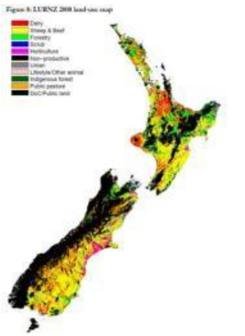
	Data	Processing	Delivery
Land use	Combining spatial data Formalising definitions	Indicator pipeline Validation	Visualisation land cover trends
Soil health	500 soils data into repository Investigate other soil DB	Soil health pipeline Exploring LU impacts on soil resources	Sharing soils data Visualisation of soil quality data
Species occupancy (trees, birds)	Presence/absence and Presence only data	Data pipeline Taxon concept	Visualisation of actual and potential species ranges

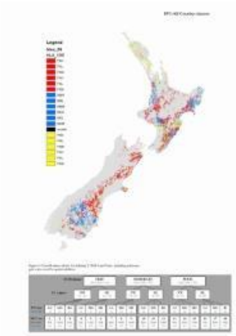
NEXT GENERATION LAND USE MODELLING

Land use classification

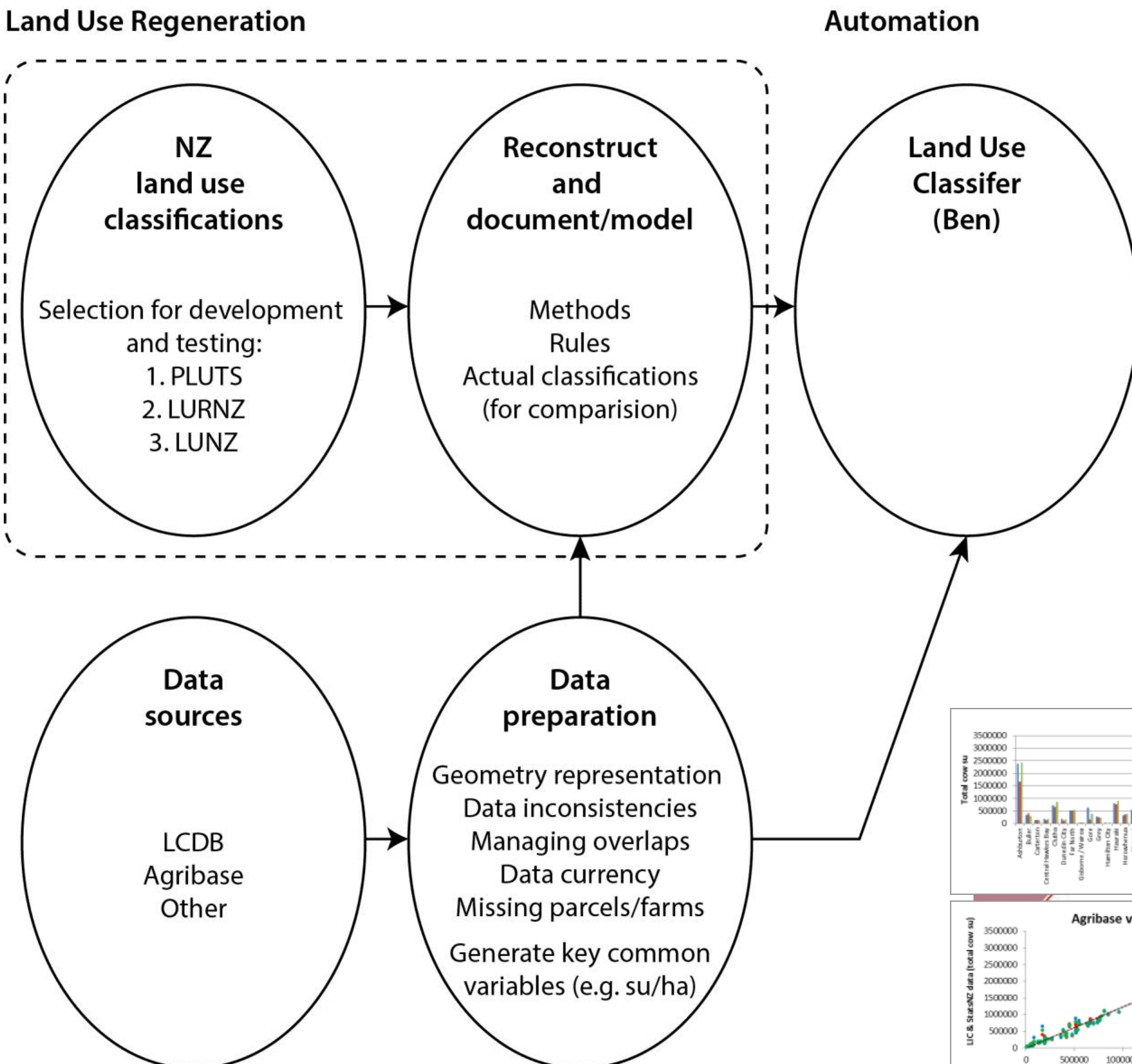
- Many LU classifications past and present
- Different scales of application/development
- Drivers = modelling, policy development, reporting (indicators), economic development
- Tend to draw on the same datasets but…
  - Different but same
  - Limited transparency
  - Assumed trust in the underlying data
  - Reproducible?
  - Rapidly dated

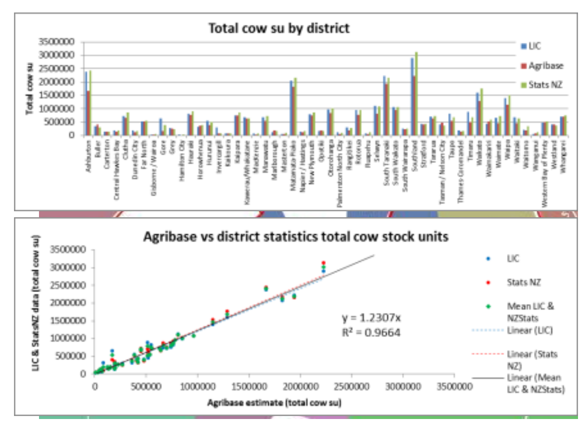




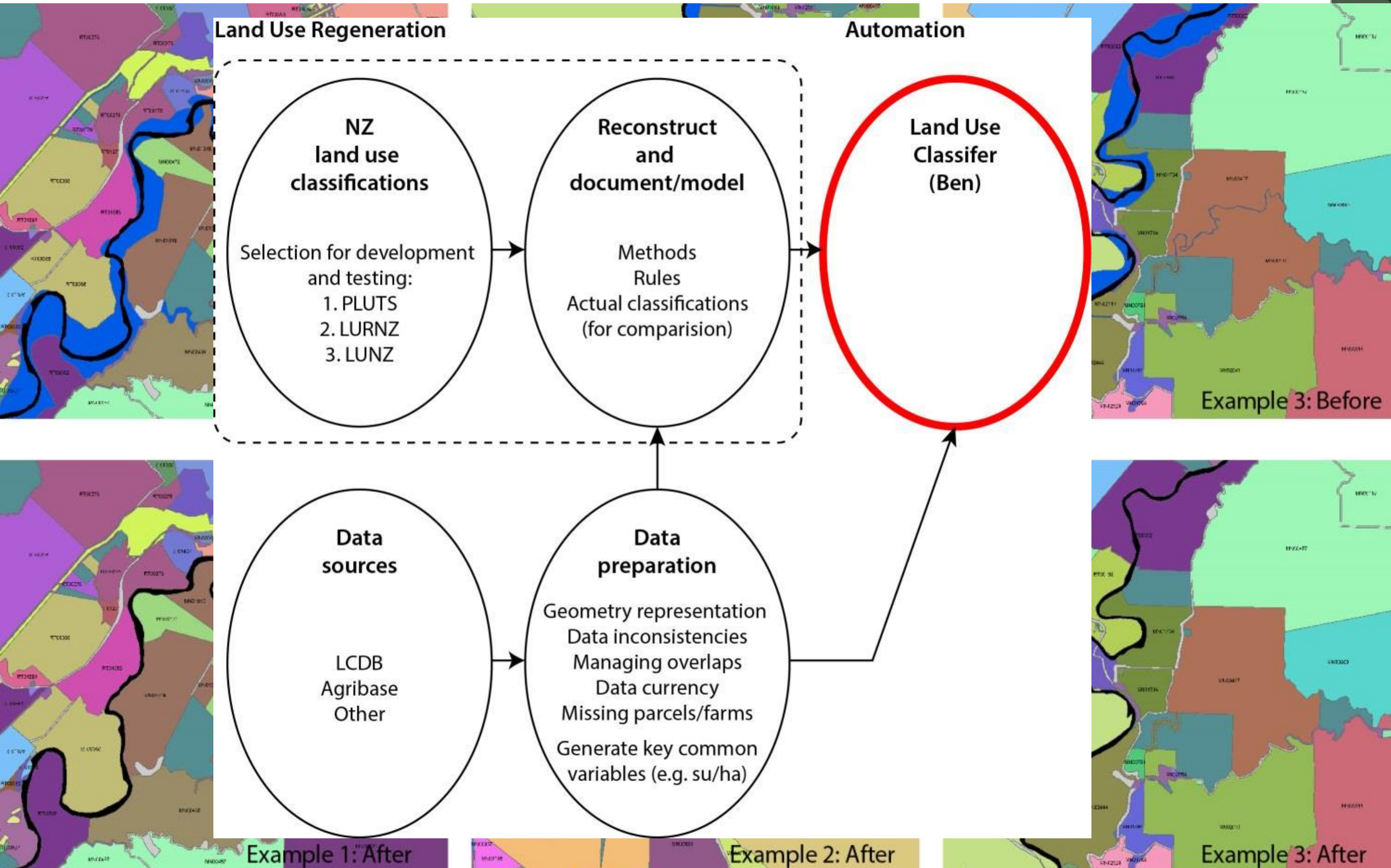


Land Use Classifier (part 1)





Land Use Classifier (part 1)



LUC Engine

A Land Use Classification (LUC) should be:

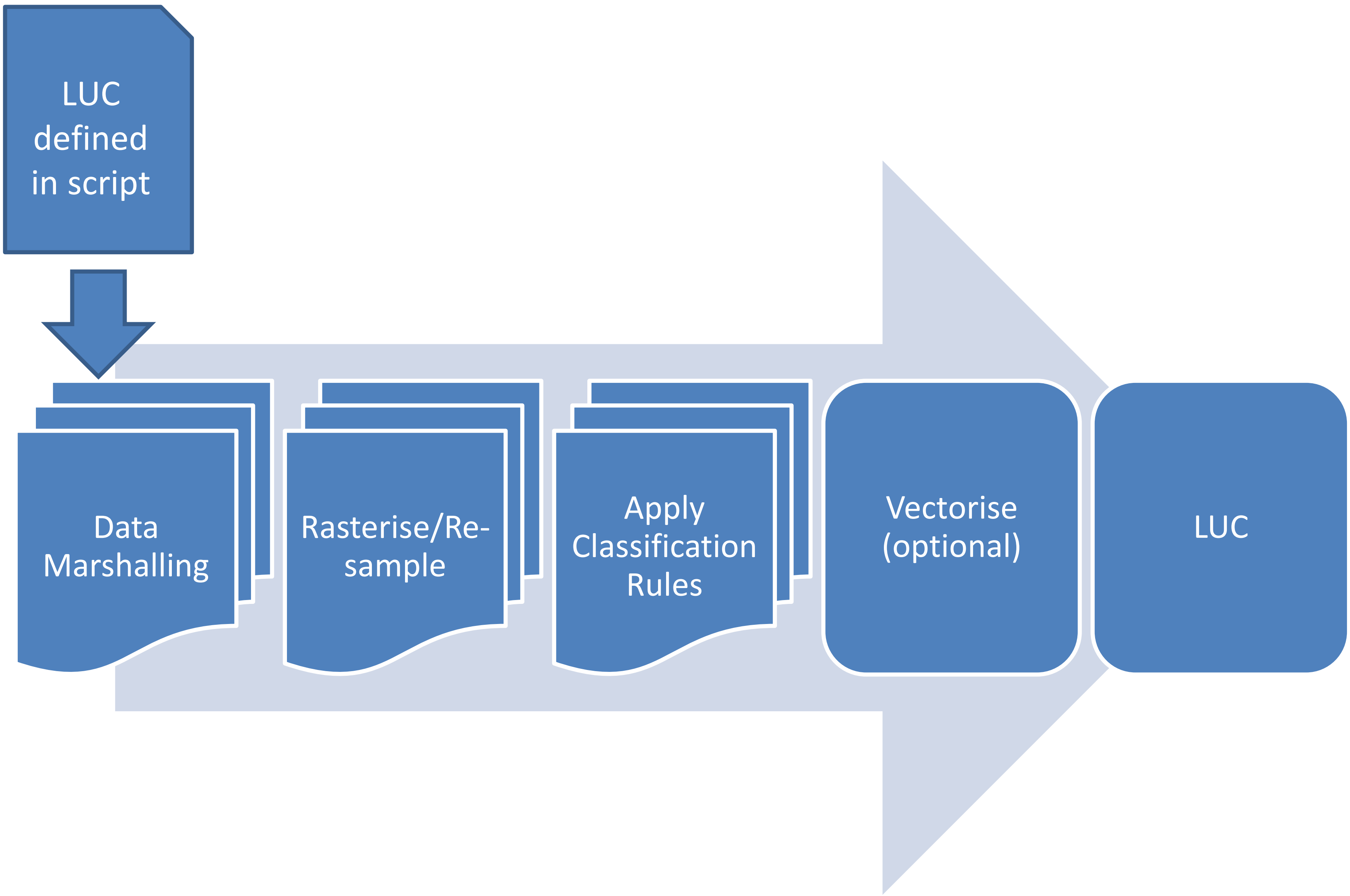
- Well documented
- Transparent
- Repeatable

for both data sources and classification rules

To achieve this, we want to use a single file to define the LUC that is:

- Human-readable
- Machine-readable

LUC Engine - Workflow



LUC Engine - pyluc

LUC rules defined within simplified, human readable, Python scripts

Framework takes care of everything else

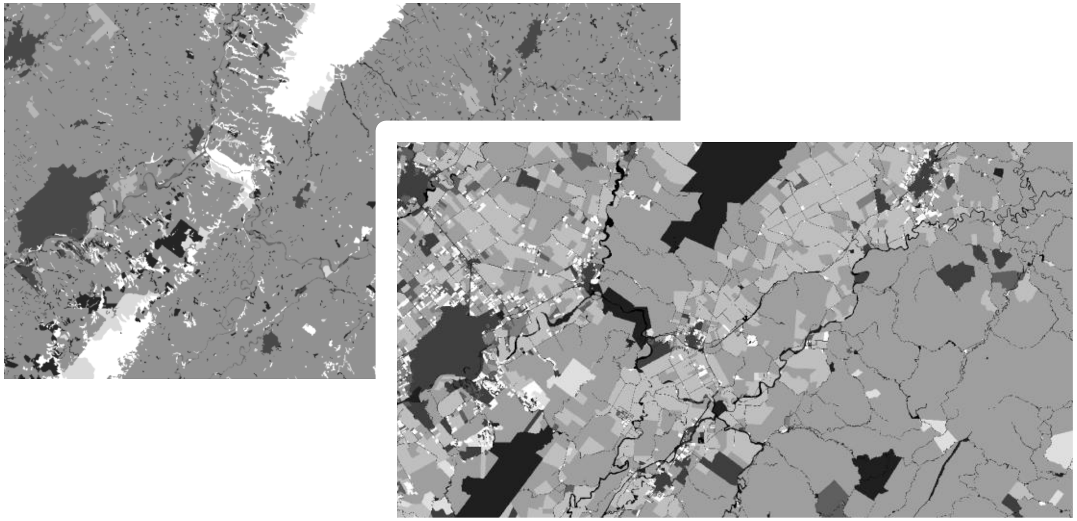
Pros

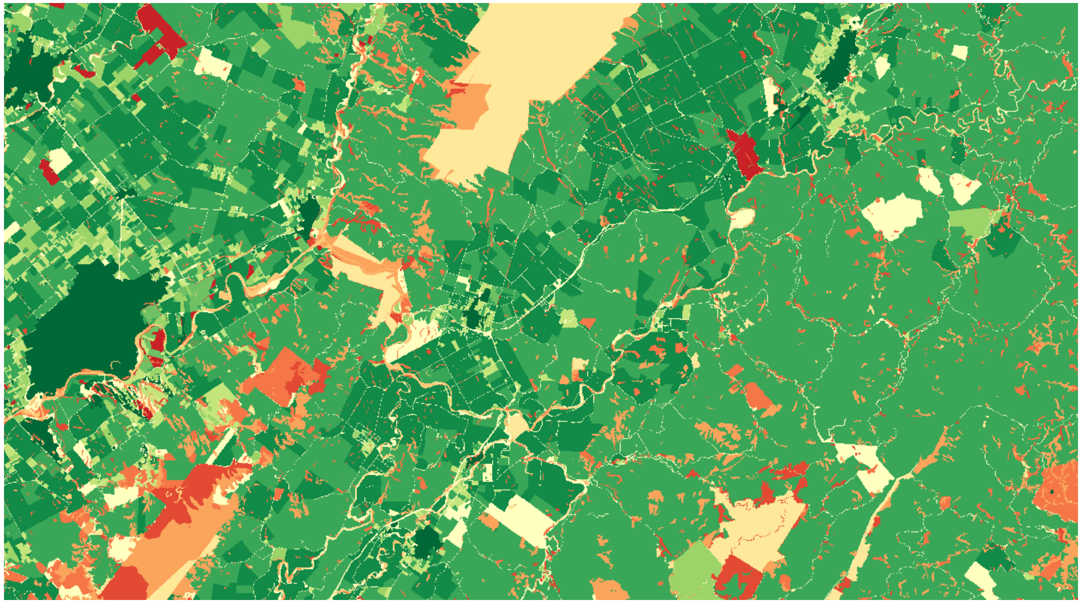
- Flexible, powerful LUC rules possible
- Very basic Python knowledge required
- Runs on HPC (performance, stability/consistency) or desktop
- Accepts any SHP, KEA, or IMG file
- Will take data straight from LRIS portal or other Koordinate sites

Cons

- Some basic Python knowledge required to create scripts
- Rasters involved in processing chain

LUC Engine - pyluc





SMARTER WAYS TO DELIVER DATA

A Soil Data Interoperability Experiment

- Soil data needs to be freely available for a wide range of purposes
  - *global/national/local initiatives require interoperable solutions*
- An accepted, global soil data standard is required
  - *for the transfer of soil feature data, including data about soil bodies, profiles, and horizons, and related entities*
- Problem: existing standards are diverse and don't cover all our data exchange needs
  - *need to reconcile them in a single, comprehensive, global standard*
- A OGC Interoperability Experiment was seen as the first step in achieving this
- Initiated by the International Union of Soil Sciences Working Group on Soil Information Standards
  - *Led by Landcare Research, funded in part by IDA*

Open Geospatial Consortium (OGC)

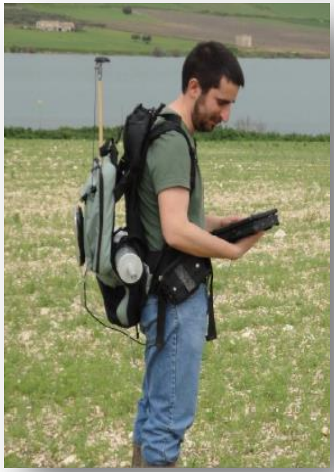
- An international industry consortium of 500+ companies, government agencies and universities participating in a consensus process to develop publicly available data exchange standards
- Different types of initiatives, e.g. Testbeds and Interoperability Experiments, used to develop standards
- Interoperability Experiments are brief, low-overhead, formally structured and approved initiatives led and executed by OGC members to achieve specific technical objectives
- An IE must focus on a single interoperability issue

OGC®
Making location count.

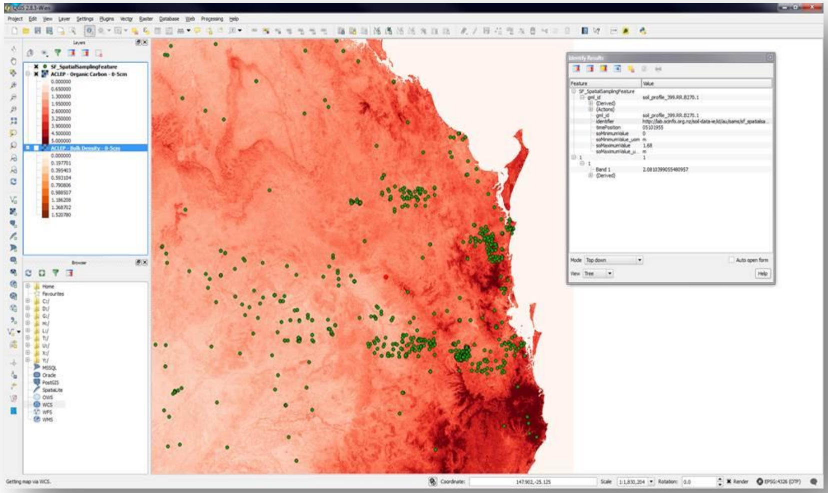
The IE

- Goal - the development and testing of a Soil Markup Language, a data encoding for soil features
- Participants: IE Initiators (CSIR0, Landcare Research, IRSIC) plus a number of other agencies e.g. USDA, USGS, Fed Uni (Aus), CRA-ABP (Italy) + Horizons Regional Council
- 6 months duration
- 4 main use cases: soil data integration & publication, soil sensor data, soil property modelling and predictions, pedo-transfer functions
- Develop a conceptual model and common exchange language (a GML-XML application schema)
- Deploy a set of demonstrators using web services (e.g. WFS, SOS) that used the schema

Collecting Soil Observations



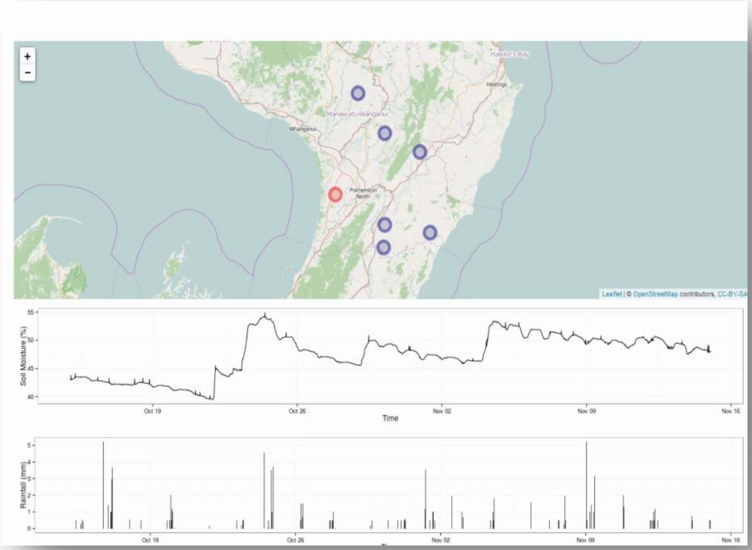
Demonstrators



Provision of soil sensor data as time series

Provision of soil sampling and observation data

Provision of soil property models and predictions



We’ve used open standards to…

- create a simple *information model* of soils data
- *harmonise* the structure and some content of soils data between agencies
- bring data from different international soil agencies together in applications for users (*interoperability*)
- provide a way to describe and organise soil concepts, features, methods, etc (*semantics*)
- use *linked data* ideas to provide supporting information about the data inside the responses to queries
- to chain services for *processing* data

➢ And demonstrated cross agency *data sharing*

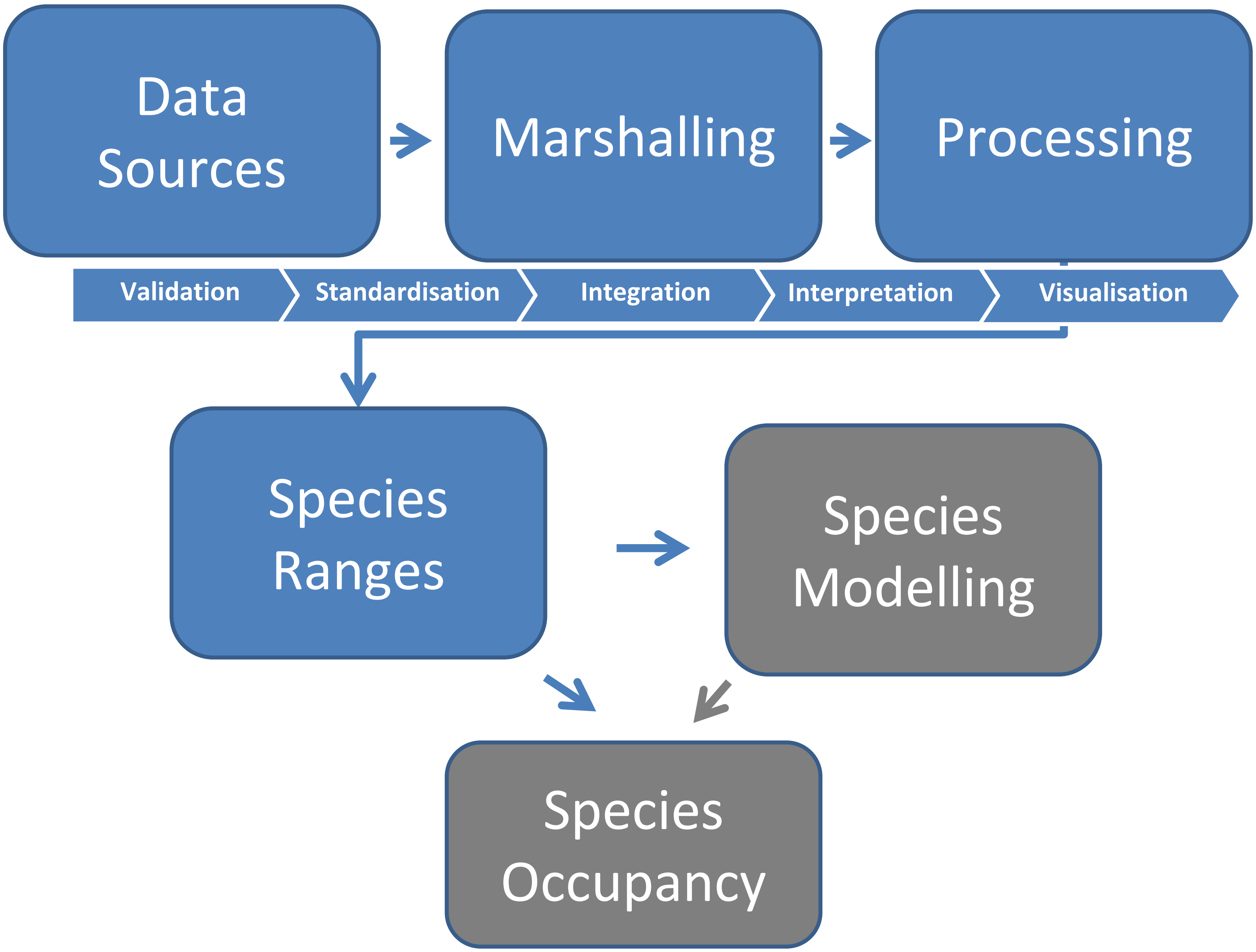
SPECIES OCCUPANCY

IDA: Supporting Species Occupancy Metrics

Species Occupancy:

Where do species live now VERSUS where could they live, and how is that changing over time?

1. **Where do (selected) species live now?**
   Requires marshalling of historical/current data from multiple distributed sources, of varying quality/standards
   IDA work will provide working solutions/tools
2. **Where could species live?**
   Requires robust modelling of potential range using actual species range data as input
   IDA work will quantify the issues



Data Sources
Marshalling
Processing
Validation
Standardisation
Integration
Interpretation
Visualisation
Species Ranges
Species Modelling
Species Occupancy

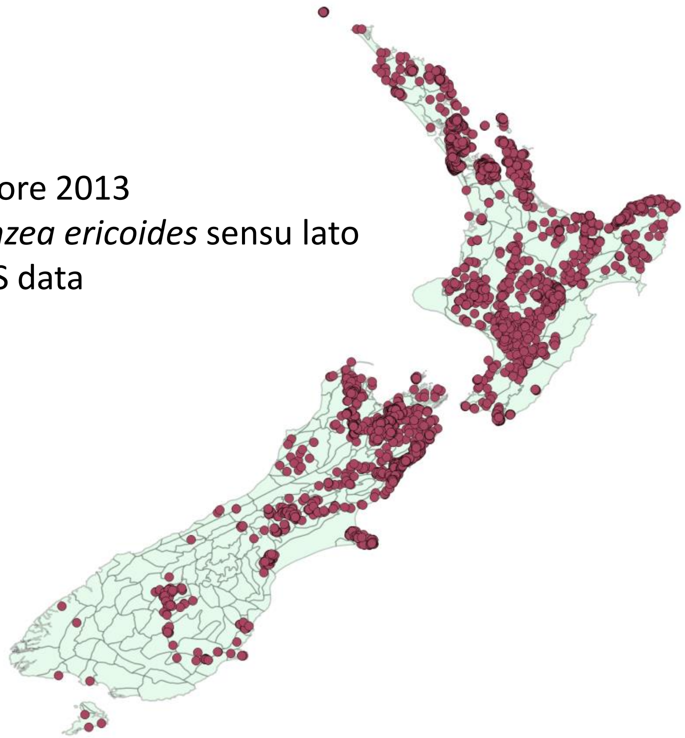
One example - The impact of changing taxon concepts

- The names of species can change over time
  - Species X changes its name to Species Y (taxonomic synonyms)
    - easy to address (NZOR)
    - e.g *Kunzea ericoides* (Kanuka) used to be known as *Leptospermum ericoides*
  - Species X and Species Y and Species Z ... are now considered to be Species X (taxonomic lumping)
    - easy to address (NZOR)
  - Species X is now considered to be Species X and Species Y and Species Z ... (taxonomic splitting)
    - Not easy to address. All historical observation-only data unusable!
    - e.g. In 2013 *Kunzea ericoides* (sensu lato) was split into 10 different species including a redefined *Kunzea ericoides* (now sensu stricto)

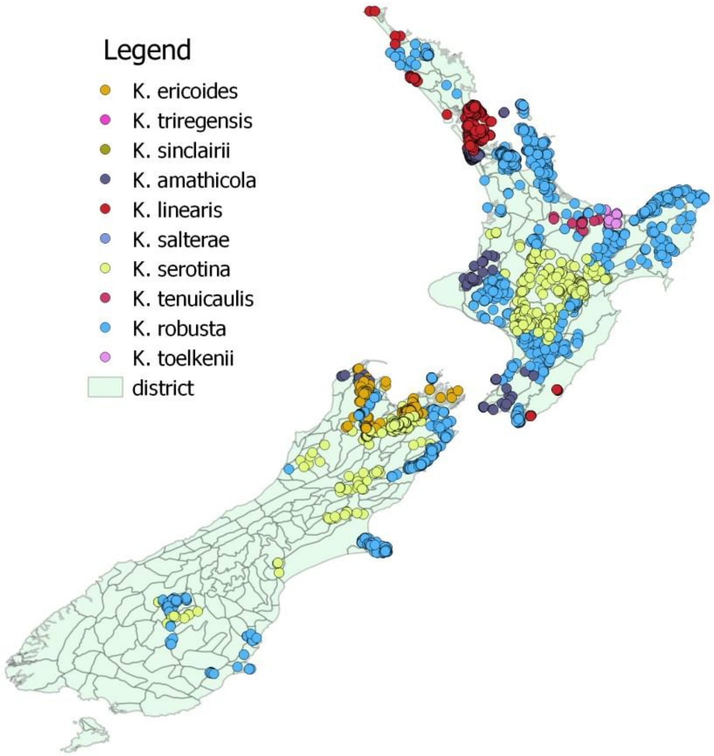
Interpretation

One example - The impact of changing taxon concepts

Before 2013
Kunzea ericoides sensu lato
NVS data



After 2013
Interpreted NVS data:
Kunzea ericoides sensu stricto + 9 other species



Historical data is not unusable.
The data has been re-interpreted.

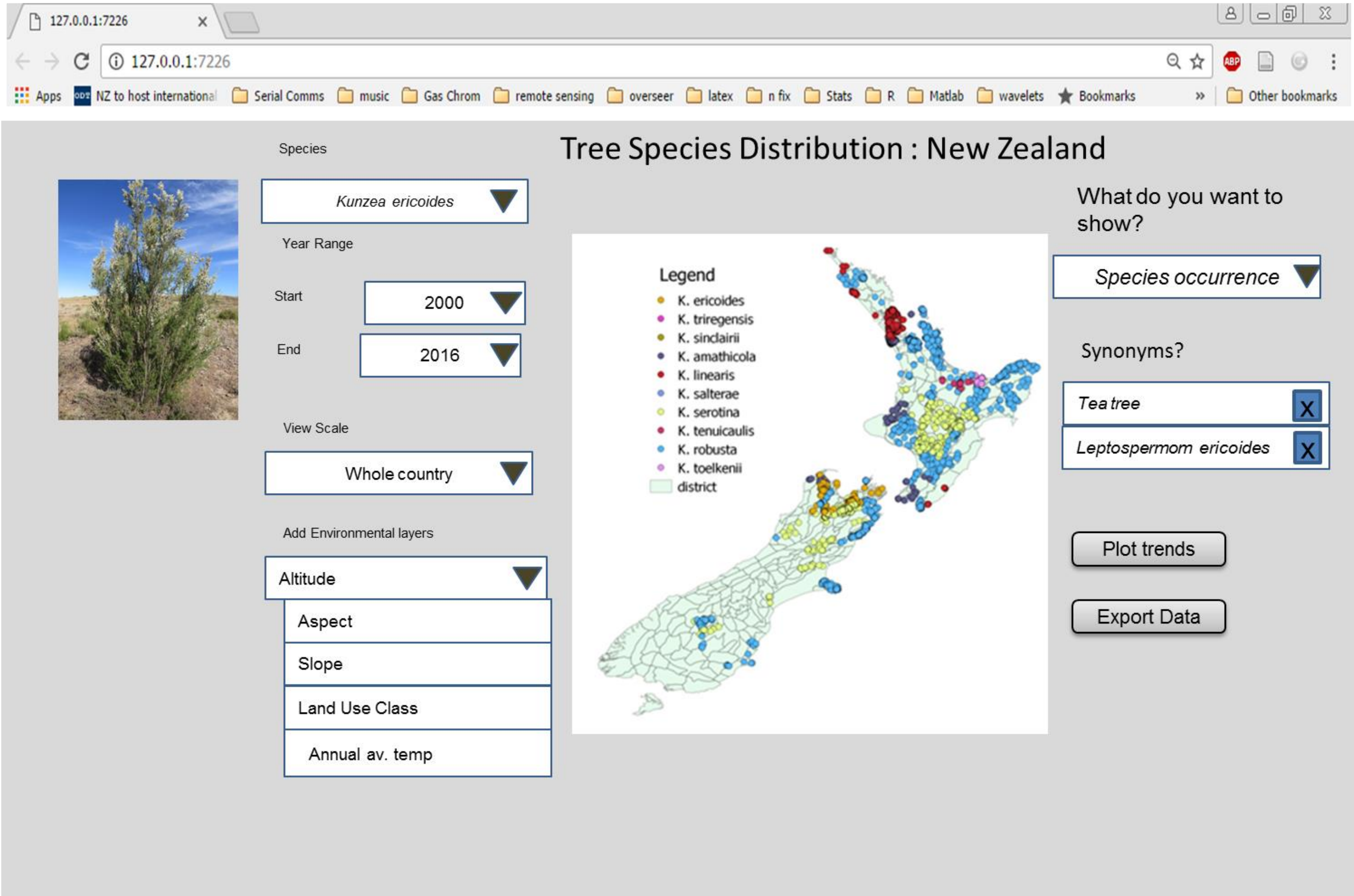
Achieved by developing a taxon concept relationship map employing constraints, such as location, altitude, tree height, DBH, growth form – all attribute data from NVS

Processing

Current and Future work

Visualisation

Web-based Visualisation Tool prototype for actual species range

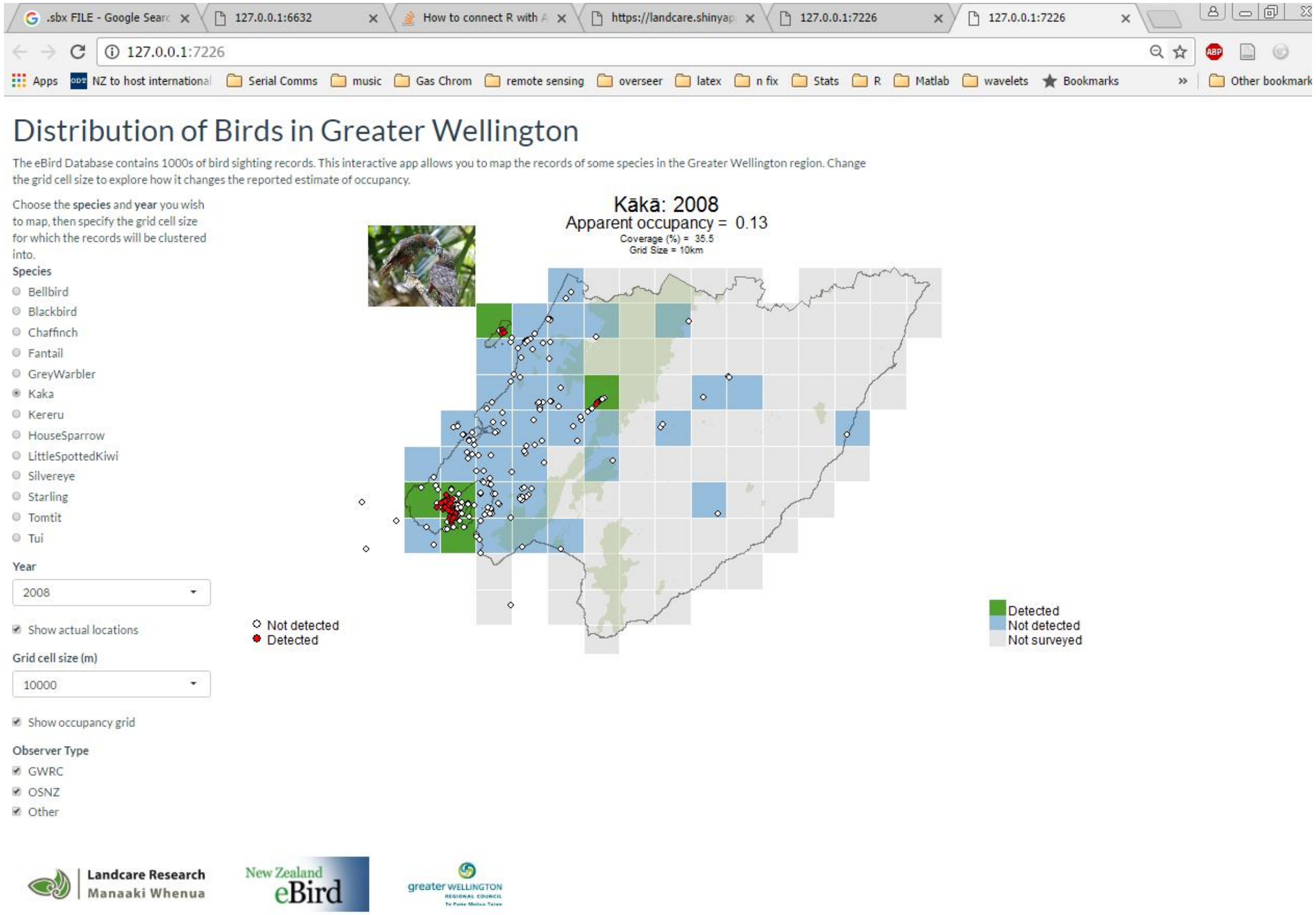


Processing

Current and Future work

Visualisation

- Quantifying issues for modelling of potential species range



Work to come…

- Continue work on pipelines for the 3 domains, focusing on indicators
- Interoperability (web services)
- Characterising provenance, quality, uncertainties, workflow
- Visualisation work

Thank you for your attention!

e: AusseilA@landcareresearch.co.nz
t: 04 382 66 42

Soil IE Engineering Report at http://www.opengeospatial.org/docs/er