SOiL CRC

Performance through collaboration

DISCUSSION PAPER

Project 2.3.001

VISUALISING AUSTRALASIA'S SOILS:
A Soil CRC Interoperable Spatial Knowledge System

Milestone 4
Governance and Data Stewardship Guidelines

Authors	Peter Dahlhaus, Federation University Australia
Title	Project 2.3.001 Visualising Australasia's Soils: A Soil CRC interoperable spatial knowledge system. Milestone 4 Governance and Data Stewardship Guidelines. Discussion Paper.
Date	December 31st 2019
Keywords	Soil data, FAIR data, data systems social architecture
Publisher	Cooperative Research Centre for High Performance Soils
Preferred citation	Dahlhaus P.G. (2019) Visualising Australasia's Soils: A Soil CRC interoperable spatial knowledge system. Governance and Data Stewardship Guidelines. Discussion Paper. Soil CRC Project 2.3.001. Cooperative Research Centre for High Performance Soils, December 31 2019.

CONFIDENTIAL

Not to be distributed beyond the Soil CRC Participants and Affiliates without the consent of the CEO.

DISCLAIMER

Any opinions expressed in this document is that of the author. They do not purport to reflect the opinions or views of the CRC for High Performance Soils or its partners, agents or employees.

The CRC for High Performance Soils gives no warranty or assurance, and makes no representation as to the accuracy or reliability of any information or advice contained in this document, or that it is suitable for any intended use. The CRC for High Performance Soils, its partners, agents and employees, disclaim any and all liability for any errors or omissions or in respect of anything or the consequences of anything done or omitted to be done in reliance upon the whole or any part of this document.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Visualising Australasia's Soils (VAS) project is a foundational project that commenced in 2018 that aims to establish a soil research data federation, based on agreed data stewardship and governance frameworks, that allows Australasian soils data from all sources (private and public), to be discoverable to all Soil CRC participants through an intuitive-to-use internet portal. This discussion paper meets Milestone 4 in the project.

The inaugural project meeting was attended by 34 participants from 17 organisations. The consensus was an agreement that the project should deliver a clear case (a value proposition) for rolling forward the project as a fundamental foundation activity in the Soil CRC.

The project participants include 14 farmer groups and two catchment management authorities. All except two were visited in the first year of the project and from these meetings three general value cases were determined: 1) access to a free, web-based, supported soil data management system; 2) opportunities to access to locally relevant soil data for enhancing industry productivity and as an evidential basis grants and reporting; and 3) providing the opportunity to provide new data services to support group members.

From the partner discussions two generalised use-cases have emerged. The first is a web-based self-serve data input system that allows groups or individuals to set the controls on who can access their data. The second is a spatial data viewing portal that allows data filtering, basic data analysis, visualisation and printing.

Each project partner has committed to contribute an initial soil data set for experimentation and functionality testing. The VAS portal – data.soilcrc.com.au - was published in December 2018 and initially shows some of the soil data publicly available in repositories of various organisations. At this stage farmer group data can be accessed via login only.

A clear preferred model for the VAS portal is one based on federated architecture in which the data remains in the control and management of the data custodians, who have the control over who can access the data and the licensing conditions. This overcomes many of the barriers to sharing soil data.

Although the aspiration is for a fully federated model, it is recognised that not all the participants (i.e. federation members) have the technical capability from the outset to interoperably serve their data according to the required standards. Therefore, as an interim solution, a brokered model will be offered in which participants can use a self-serve system to provision their data through a secure cloud-based data host, according to the international data standards.

Besides the central access point for VAS, most participants desire access to the portal direct from their organisation's website, since that is the logical trusted place for their members to find soil data and information. Embedding the VAS portal in their websites also fulfils a value case in supporting group membership, as it offers value and services to members.

While it is recommended that the Agricultural Research Federation (AgReFed) is the baseline exemplar for the VAS soil data federation, the VAS Steering Committee have deferred the decision on governance until a greater first-hand knowledge, understanding and experience with the VAS portal and its operation is acquired.

The VAS project is strongly committed to good data stewardship and will work with data custodians in implementing the guidelines to make their data findable, accessible, interoperable and reusable (FAIR), including standard assessments for FAIRness. Also encouraged is the certification of data repositories with the Core Trust Seal, hence complying with International Standards Organisation (ISO) standard ISO 16363 as a trusted repository.

1 INTRODUCTION

The Cooperative Research Centre for High Performance Soils (Soils CRC) was established in 2017 and brings together eight universities, four state agencies, eight industry partners and twenty farmer groups. The overall aim of the Soil CRC is to find practical solutions for Australia's underperforming soils. The Soil CRC will run until 2027 and was established with $59 million in cash contributions and $108 million in-kind.

The Visualising Australasia's Soils (VAS) project is a foundational project that commenced in 2018 and aims to provide access to soil data for Soil CRC participants. This discussion paper meets Milestone 4 in the project.

1.1 Brief overview of VAS project

Project 2.3.001 in the Soil CRC is titled *"Visualising Australasia's Soils: A Soil CRC interoperable spatial knowledge system"*. The VAS project aims to establish a soil research data federation, based on agreed data stewardship and governance frameworks, that allows Australasian soils data from all sources (private and public), to be discoverable to all Soil CRC participants through an intuitive-to-use internet portal.

The project has 20 partners:

Federation University Australia

Manaaki Whenua Landcare Research

University of Southern Queensland

University of Tasmania

Birchip Cropping Group

Burdekin Productivity Services

Central West Farming Systems Inc.

Gillamii Centre

Herbert Cane Productivity Services Ltd

Holbrook Landcare Network

Nutrien (formerly Landmark)

Liebe Group

MacKillop Farm Management Group

Mallee Sustainable Farming

North Central Catchment Management Authority

Riverine Plains Inc.

Southern Farming Systems

Western Australia No Till Farmers Association

West Midlands Group

Wimmera Catchment Management Authority

The project is overseen by a Steering Committee comprising the following members:

Project Leader – A/Prof Peter Dahlhaus, Federation University

Soil CRC Program Coordinator – Dr Richard Doyle, University of Tasmania

Technical expert – Alistair Ritchie, Manaaki Whenua Landcare Research

Broadacre farming group – Dr David Minkey, WA No Till Farmers Association

Cane grower group – Rob Milla, Burdekin Productivity Services

Mixed farming group – Jane McInnes, Riverine Plains Inc.

Government/agency – Warwick Dougherty, NSW Department of Primary Industries

Global initiatives – Dr Peter Wilson, CSIRO.

The VAS project commenced with a four-day meeting at Federation University, Ballarat, 11-14 February 2019, with a total of 34 participants from 17 organisations attending on various days. One of the outcomes of the workshop was an agreement that the project should deliver a clear case for rolling forward the project as a fundamental foundation activity in the Soil CRC. This case would include:

- A working portal with farmer group / industry group data across the nation
- Groups advocating use of the portal because there is an obvious benefit
- Case studies / examples of the value proposition to farmers / advisors / catchment managers
- Value add to data, for example by showing trends
- Greater emphasis on point data than raster data (although both are important)
- Map the way ahead for sustainability
- Technology neutral to adapt to future advances in applications

Hence it was clear from the outset that the value proposition for the soil data federation was crucial to its success.

Following on from the workshop, the vast majority of the time in the initial year of the project was spent visiting each of the project participants at their office to elicit their views on the value proposition for the VAS project. Two groups – Southern Farming Systems and Riverine Plains Inc. – were chosen as 'pilots' on the basis that they were already involved in other data management projects with the research team. Both groups contributed to the establishment of the initial data governance and stewardship principles that were then taken to all the remaining partners. Two participants were missed in the engagement cycle – Leibe Group, due to a last minute meeting cancellation, and Nutrien (formerly Landmark) due to a change of business ownership.

Table 1. List of engagement with participants

Date	Participant	Place
17 April 2019	Southern Farming Systems	Mt Helen, Vic
13 June 2019	Riverine Plains Inc.	Mulwala, NSW
13 July 2019	Holbrook Landcare Network	Holbrook, NSW
17 July 2019	North Central Catchment Management Authority	Huntly, Vic
3 September 2019	Burdekin Productivity Services	Ayr, Qld
4 September 2019	Herbert Cane Productivity Services Ltd	Ingham, Qld
24 September 2019	Mallee Sustainable Farming	Mildura, Vic
25 September 2019	Birchip Cropping Group	Birchip, Vic
15 October 2019	Wimmera Catchment Management Authority	Horsham, Vic
22 October 2019	WA No Till Farmers Association	Floreat, WA
23 October 2019	West Midlands Group	Dandaragan, WA
25 October 2019	Gillamii Centre	Cranbrook, WA
12 November 2019	Mackillop Farm Management Group	Naracoorte, SA
25 November 2019	Central West Farming Systems	Condobolin, NSW

1.2 Value propositions

In our meetings with each of the project participants, we have generally heard the following value cases for their participation in the project.

1. Free access to a trusted, supported, web-based spatial soil data management system to:
    - store and retrieve soil data for the organisation (data management), including tools to easily enter data, or link to the organisation's database, and extract data in various formats
    - filter data, graph data and present data via the web and/or email (pdf) for members of the farmer groups or CMA stakeholders
    - improve management, quality and value of organisational data, by adding metadata and/or data standards in a supported way
    - be able to monitor trends in soil properties over time
    - be able to benchmark farms and local areas against each other
    - combine organisational soil data with all the other available soil data to examine trends and/or soil properties at a chosen location or area
    - store, catalogue and retrieve non-spatial data such as documents, images, videos, etc.
    - store, catalogue and retrieve trial data and on-farm experimentation data (i.e. non-soil data, but related to soils management)
    - easily find soil data or information to assist in member or stakeholder queries

2. Having access to locally relevant data for enhancing industry productivity and as an evidential basis to:
    - identify research gaps for funding proposals
    - know who is accessing an organisation's data and for what purpose, perhaps leading to new funded collaborations
    - save time and effort in reporting to funding bodies and investors
    - access monitoring, evaluation, reporting and improvement (MERI) metrics on demand

3. Support the organisation's members or communities by offering services via the VAS data portal and tools such as:
    - online educational materials or training courses to offer members (i.e. region specific)
    - pest and disease reporting and alerts
    - early warning of climate and biohazard events
    - view more extensive data such as groundwater levels, seasonal forecasts, terrain/drainage, soil moisture trends, etc.
    - individual farm logins to include the ability to store, retrieve and visualise soil data and other agricultural data, such as paddock management records, yield maps, NDVI maps, FOO, etc.

- independent (i.e. non-commercial) soil additive calculators, variable rate application calculators, response curves, etc. for production analysis and decision support
- farm management decision support tools (unspecified) to assist farmers in practical actions to deal with soil constraints (e.g. sodicity, salinity, non-wetting soils, organic matter, etc.)
- farm-scale carbon-budget calculator and farm-scale water-budget calculator, and evidence of best practice metrics for social licence to operate

1.3 Use-cases

From the value propositions and discussions with the participants to date, the following generalised use-cases have emerged.

1. A website accessed via a login in a standard internet browser, with capability to provision data to the VAS soil data federation. Note that this use-case is required by all groups visited thus far, since they currently do not have the capability to expose their data to the Soil CRC VAS federation through web-services or API. This requires secure space on the Soil CRC VAS cloud-based server for each group, with a self-serve system that supports the following use-cases:
    - a group administrator login with the control to set-up other user logins with various permissions
    - a common information model, data structures and data standards to organise data so that they can be findable, understandable and reusable in the future
    - the capability to easily and simply add soil data from a spreadsheet, or other database, or by entering individual records to a group-controlled database in the soil data federation
    - the ability to add, edit or delete soil data records from an organisation's data sets
    - the ability to add and catalogue documents, images, videos, podcasts, or other non-spatial digital data
    - the ability to add georeferenced maps or GIS layers (e.g. shapefiles, Geotiffs, KML files) or web-services (e.g. WMS, TMS, WCS)
    - the ability to auto-fill and generate metadata records from the information entered
    - access controls to allow the organisation's data that is added to be viewed or accessed in different ways by different end-users. These different end users are generally: public open access (i.e. anyone, anywhere), the 'club administrators' (i.e. trusted users, e.g. farmer group staff or CMA staff), and individuals (i.e. data owners or custodians, e.g. farmers, landholders, agronomists). The controls are in relation to the location accuracy (e.g. randomised pixels, paddock, or exact location), embargoes (e.g. past data only or set withholding period) and the typical Creative Commons controls (attribution, non-commercial, no derivatives, share alike). A default access control is to provide contact details of the data custodian so that access to the data can be negotiated on a case-by-case basis (i.e. bespoke licensing).
    - the ability to change access controls for an organisation's data set or remove a data set

2. An information portal website accessed via a standard internet browser that allows users to:
    - explore spatial soil data in a web-based mapping application, together with other contextual information (including metadata). View the information according to the access controls set by the custodians
    - search an e-Library of resources and view the information according to the access controls set by the custodians
    - ability to send data access requests to the data custodian, for locked data
    - login for club views and private views of data and information according to the access controls set by the custodians
    - filter data by soil data fields (e.g. show pH ≥ 4.5, show %clay > 50%)
    - download data (or filtered data results) according to the access controls set by the custodians
    - create pdf files of map views or data tables
    - save bespoke map views as an emailable link
    - save bespoke map views as Google Earth format (*.kml), ESRI shapefile (*.shp) and/or GeoTIFF (*.tif)
    - report on all soil data available for a property, user-selected polygon, line with a designated buffer, and point with a radius
    - location summary of the average soil properties within a set radius around a point
    - graphs that can show trends over time where sufficient data exists
    - animations of spatial raster data such as changes in soil moisture over time, and animations of polygon data such as paddock soil properties over time
    - tools to allow interpolations of point data (i.e. gridding tools) and grid calculators to create new coverages of soil data for specific needs of the end-users
    - visualisations of soil data down a profile

1.4 Data contribution

Soil data contributions are generally separated into four categories: 1) the publicly available data in repositories of various organisations, 2) the data available from university research, such as undergraduate and postgraduate research projects, 3) the usually unpublished data collected by farmer groups and their members and 4) soil data from sources outside of the Soil CRC participants.

1.4.1 Public soil data

In the current Australian governments, research and university open data repositories, there are at least 23,000 soil data sets freely available (Table 2). It is recognised that this number is the result of an initial search and that the final number of data sets may well be in excess of 30,000.

Some of these data may be duplicated, and their geographic extents, formats, availability, interoperability and reusability are being assessed. A significant challenge for the project is how to show the soil data that is of most use to the end-users, as varied as they are!

Table 2. Public data sets found using the search keyword “soil”

Source	number	Website
Commonwealth Government	17,843	data.gov.au
CSIRO	190	data.csiro.au/collections
Geoscience Australia	114	services.ga.gov.au
ACT Government	5	www.data.act.gov.au
NSW Government	214	data.nsw.gov.au
NT Government	2	data.nt.gov.au
Qld Government	28	www.data.qld.gov.au
	254	qldspatial.information.qld.gov.au/catalogue
Tas Government	73	www.thelist.tas.gov.au/app/content/data
Vic Government	33	data.vic.gov.au
WA Government	530	data.wa.gov.au
Research Data Australia	2,890	researchdata.ands.org.au
Terrestrial Ecosystems Research Network	453	tern.org.au
National Library of Australia	634	www.nla.gov.au
Total	*23,263*	

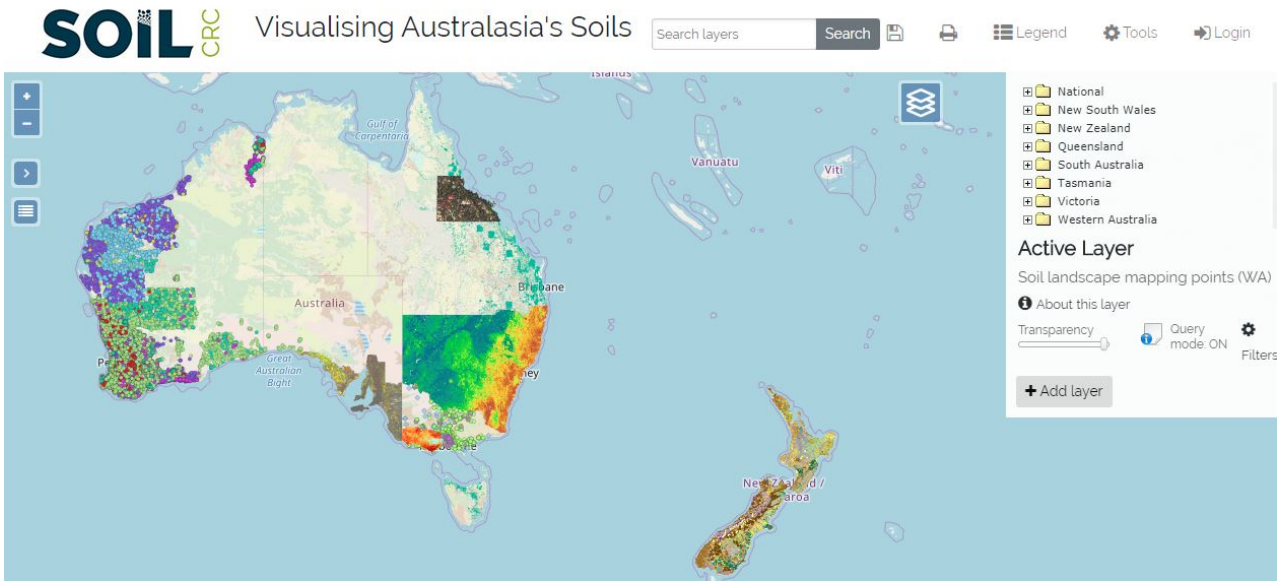

Figure 1. An example of some of the public data sets currently shown in the VAS portal.

1.4.2 University research soil data

In terms of soil data contributions, participating universities and research organisations also hold considerable quantities of legacy soil data that could be made available. To lead by example, three legacy collections have been identified by Federation University and are being prepared for exposure in the VAS soil federation portal:

- Federation University – profile descriptions and soil tests from three Honours student theses

- Federation University – digital soil maps (N Robinson PhD, 2016)
- Federation University/RMIT – Geotechnical soils data (P Dahlhaus M.App.Sci., 1980)

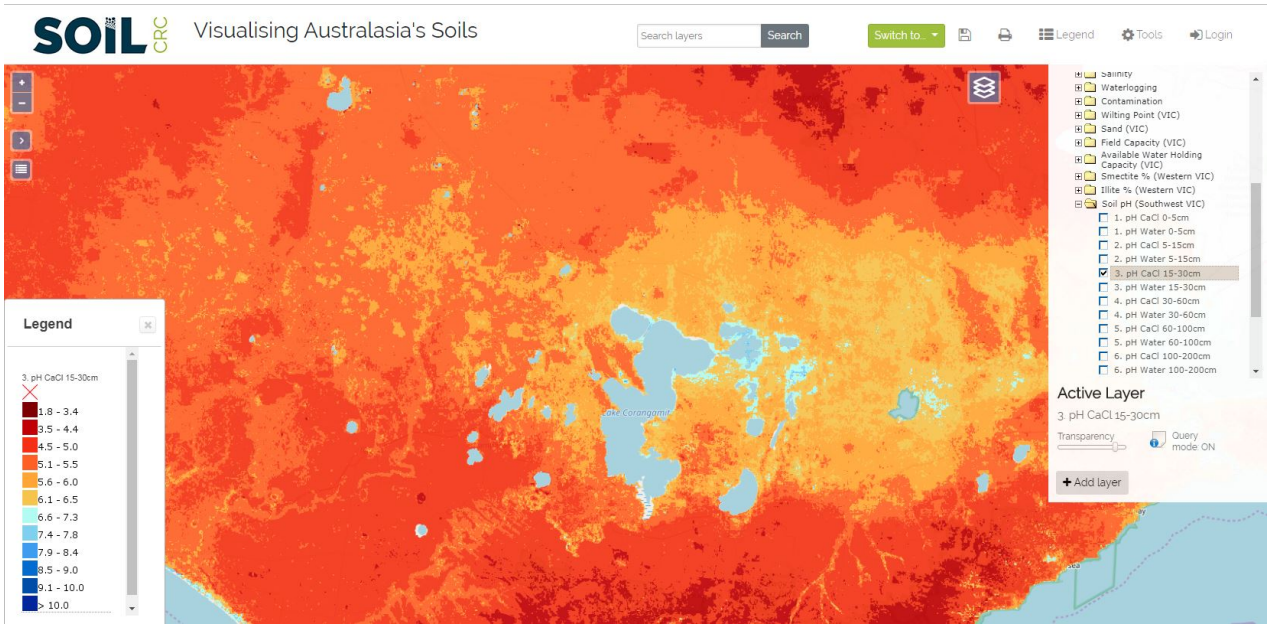

Figure 2. This digital soil map of pH_{CaCl} 15-30cm depth across south west Victoria was generated as an output from PhD research at Federation University (Dr N. Robinson, 2016).

1.4.3 Farmer group soil data

All project participants we have visited have agreed to trial some soils data in the VAS project. In all cases, initial data will not be publicly available until the VAS participants and research team have used these data to co-develop the self-serve input system with controls for data access, as well as visualisation and filtering tools to explore the data.

To date soil data has been supplied from:

- Southern Farming Systems (SFS) – Soil moisture probe data (~70 probes), approx. 200 million observations; plus SFS/Woady Yaloak Catchment Group – soil tests from ~ 95 farms (1994 – 2017) 900 paddocks, 1000 samples, approx 15,000 observations; plus SFS/Corangamite CMA – 100 soil monitoring sites (2015 & 2018), 600 samples,10,500 observations.
- Riverine Plains Inc. – 63 samples, 1385 observations.
- North Central CMA – ~100 samples
- Holbrook Landcare Network – 237 farms, approx 60,000 observations
- West Midlands Group – ~120 samples
- Gillamii Centre – ~230 samples
- Burdekin Productivity Services - ~460 sampled sites

1.4.4 Soil data from non-participants in VAS

There are many soil data sources that are not easily found in the public domain (even though they may be open data) such as environmental audits of contaminated sites, geotechnical investigations, geochemical surveys for mineral exploration and agricultural soils data from sources outside of the Soils CRC. Many of these are also included in the VAS data portal, as they are publicly available, or shown with data custodian consent.

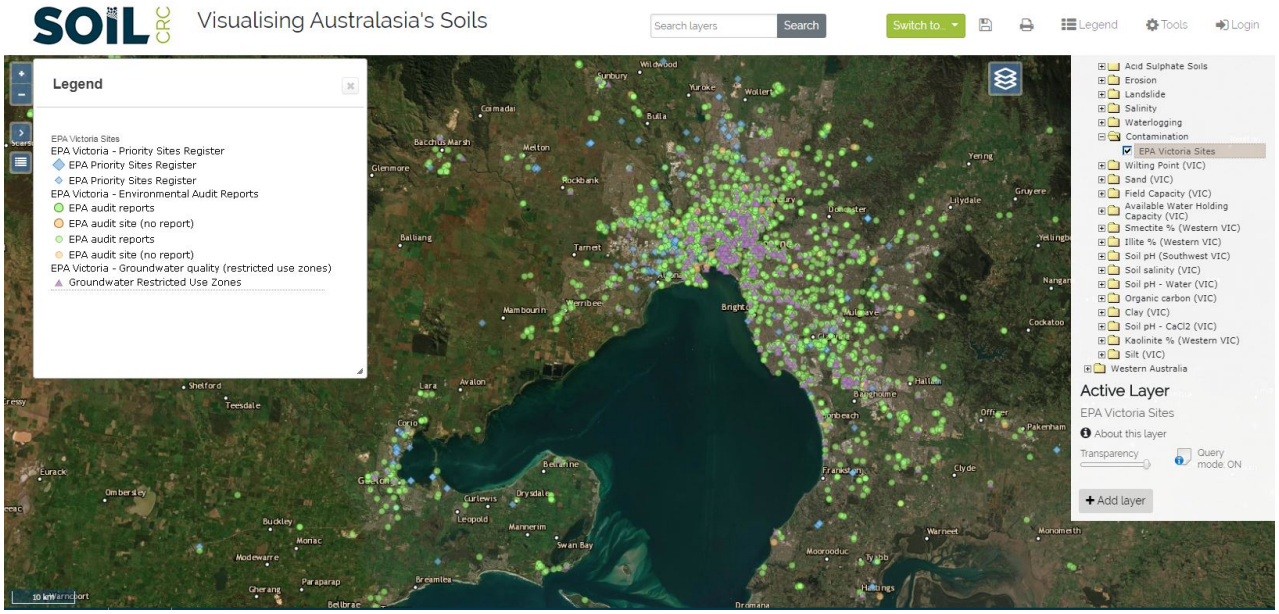

Figure 3. Audits of contaminated sites around Port Phillip Bay (EPA Victoria data, 2019)

In other cases, soil data custodians have made data available through other soil data portals, such as the South West Victoria Soil Health knowledge Base which shows the soil health monitoring sites for the Corangamite and Glenelg Hopkins catchment management authorities. Other farmer groups, such as Perennial Pasture Systems and landcare groups, such as the Woady Yaloak Catchment Group also have elected to host their data in these portals.

1.5 The VAS portal

The initial version of the VAS portal was made available to the public in December 2019. The portal can be accessed at data.soilcrc.com.au

The current VAS portal is a work in progress and in many respects a taster of what is yet to come. There is still quite a bit to do, including co-designing and developing tools and visualisations with end-users. At the moment, almost all the data shown is public sector data and is generally brought in on-the-fly.

During 2020 the VAS portal will be enhanced through the addition of more data and tools that meet as many of the use-cases as feasible (refer to Section 1.3).

2 GOVERNANCE

The question arises of how the VAS portal should be governed. The intention is to implement a governance structure that will ensure:

- Trust in the Soil CRC data and the community sharing these data
- An enduring data federation community beyond the life of the Soil CRC
- That data custodians in full control of sharing their data (i.e. they set the rules)
- That data is not misused, and that there are no disadvantages or penalties in data sharing
- That there are incentives to share… not just funding, but a clear reward (value proposition)

2.1 The federated model

The vision of the VAS project is to enable findable, accessible, interoperable and reusable (FAIR) soils data to accelerate research and innovation for the Soil CRC partners and their communities. The VAS takes a unique approach in that data providers are part of a community - a Federation - that works towards the shared goal of making soils data more FAIR.

There are several models of data sharing as illustrated by Box et al. 2015 (Figure 4).

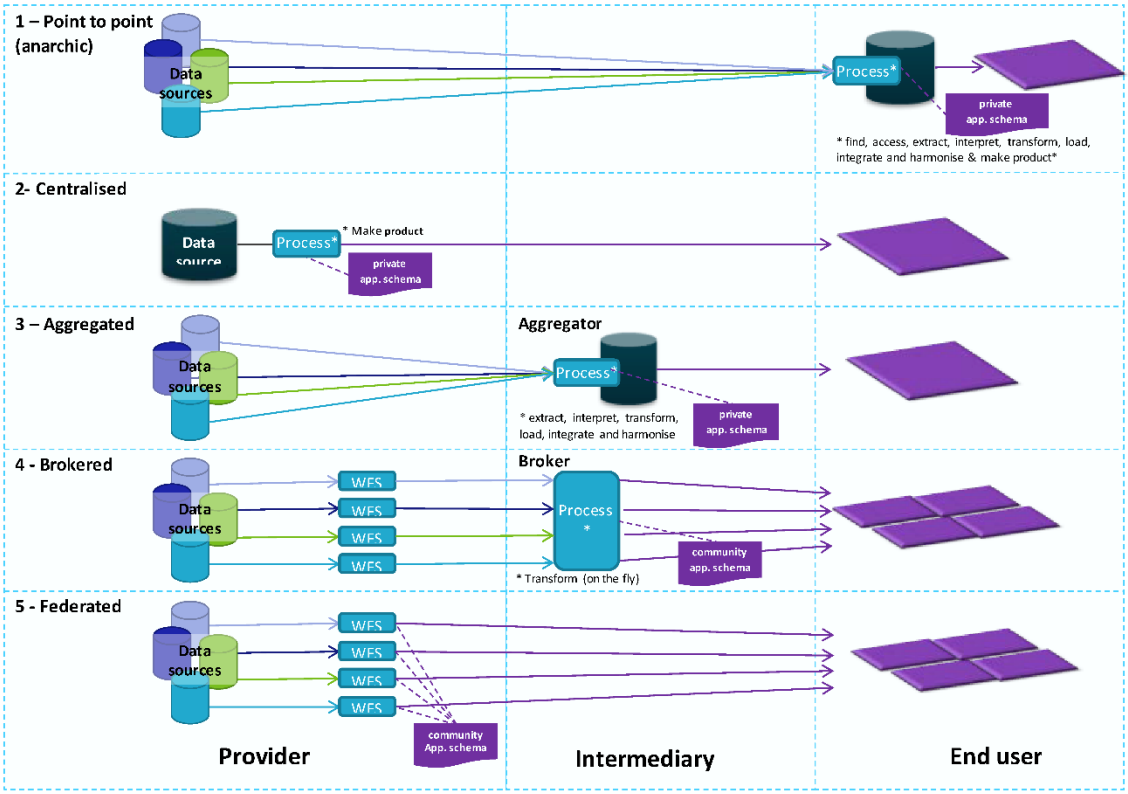

Figure 4. Data supply models (Box et al. 2015)

In the first three models depicted above, the data providers supply their data to a system that manipulates it into a common format and displays it as aggregated data in a web portal. In these cases, the data custodians relinquish control of their data, perhaps under a bespoke licensing

agreement that may involve payment for data. In both the brokered model and the federated model, data custodians maintain control of their data and expose it to the end-users according to the rules that they set. In other words, the custodians remain in full control of their data, including who gets to access it and under what licensing conditions.

Each model has some advantages and disadvantages depending on whether you are the data provider, consumer or intermediary (Figure 5). While the federated model puts the controls in the hands of the data provider, it also creates a significant effort to be able to serve data with rich metadata in the agreed formats and schemas to the users. By comparison, simply handing over the data to the consumer places a burden on the end-user to understand the data formats, structures, provenance, veracity, units, errors, licensing, etc.

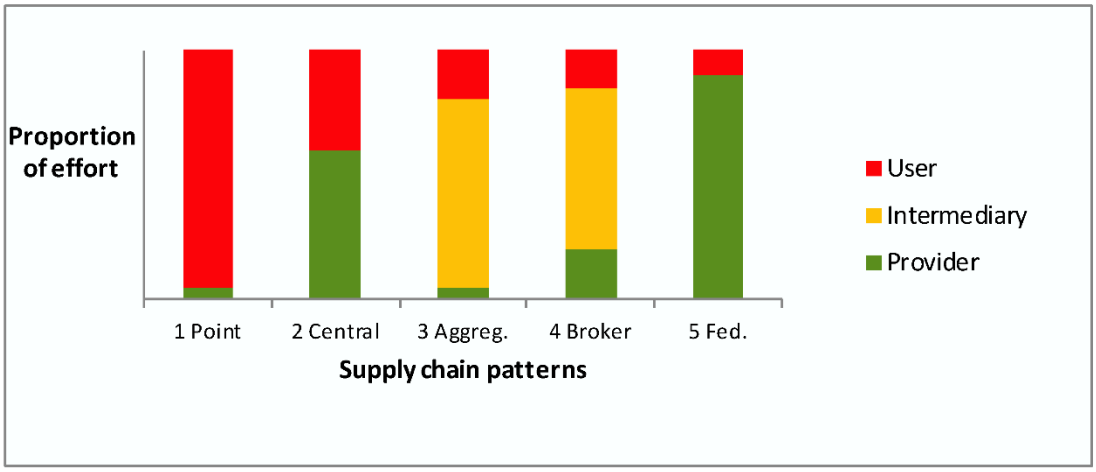

Figure 5. Relative effort for the stakeholders in each supply chain model (Box et al. 2015)

The aspiration for VAS is as a fully federated model, however it is recognised that not all the participants (i.e. federation members) will have the capability from the outset. Therefore, in the interim, a brokered model will be offered as a step towards a true federation as discussed more fully in the next section (Section 3.1 of this discussion paper).

It is expected that the VAS will result in a collegiate community that shares ideas, experiences, and examples on how to adapt existing technologies (especially those sourced from the open source communities) to maximise the benefits to the federation. This component will also reach into the international collaborative space, since some of the research participants are already active in the Open Geospatial Consortium (OGC), World Wide Web Consortium (W3C) and global initiatives to improve the sharing of agricultural data (e.g. FedUni, MWLR & UTas).

2.1.1 AgReFed as an exemplar for VAS

In 2019, the Australian Research Data Commons (ARDC) established the Agricultural Research Federation (AgReFed – www.agrefed.org.au). AgReFed has been established as a true federation in which the foundation members – ARDC, Federation University, CSIRO, University of Adelaide, University of Western Australia, WA Department of Primary Industries and Regional Development, University of New England – have established governance mechanisms through which the federation will be steered and governed. These include a Federation Council and Technical Committee. Importantly, the federation is not 'owned' by any of the members. AgReFed Provider Communities are independent, autonomous entities who retain control over their own data, data services, repositories and operations. The Federation Council determines AgReFed strategy and policy for the alignment of these Provider Communities through AgReFed

to achieve the collective AgReFed vision.

A substantial effort has gone into the AgReFed Governance arrangement, and this can be found in three main documents:

- The Guidelines for the development of a Data Stewardship and Governance Framework (Box et al. 2019)
- White Paper for the enactment phase (Box et al. 2019)
- Policy Suite for the enactment phase (Wong et al. 2019)

It is proposed here that the AgReFed model serves as an exemplar for the VAS project. The most obvious benefit is that the VAS project can leverage the intellectual effort invested in the AgReFed project. A question is whether the model may conflict with the Soil CRC philosophy or intention for the VAS project.

2.2 VAS ownership and branding

Given the distributed system and the proposed governance of the federation, the concept of portal ownership and branding must be carefully considered. The VAS federation has soil data providers, soil data consumers and soil data prosumers[1], resulting in the data being drawn from multiple disparate custodians and is attributed as such. It is not all Soil CRC data. The vast majority is drawn in from other custodians. Some data will be interoperably served from and consumed by other platforms (not just VAS) according to the access controls set by the data custodians.

In addition, it is inevitable that over the life of the Soil CRC, the VAS will transition from a research data cloud to a platform, where soil data users access and co-create shared tools, workbenches, and data pipelines, and have access to complementary FAIR data (e.g. agricultural, climatic, environmental and economic data). Because it is a distributed model, the software stack/platform cannot be strictly specified, as many members and prosumers have already invested in their chosen infrastructure. Hence the VAS platform cannot be prescriptive in terms of which technologies, applications, tools, methodologies, approaches and platforms to adopt, since it will be unique to each member's technology stack. The core VAS technology is built on open source software and open standards. It is not a bespoke Soil CRC product *per se*, it is a data federation.

Since it is a Soil CRC funded initiative, it is logical that the portal has the look and feel of the Soil CRC branding. The question arises on how to acknowledge the federation members. How should the federation be branded to acknowledge all the contributing members?

In addition, almost all of the project participants desire that the portal is embedded in their organisation's website, since that is the logical trusted place for their members to find soil data and information. Besides, having the portal embedded in their websites fulfils the value proposition of supporting group membership, as it offers value and services to members. It is proposed that a 'widget' is offered that allows the participating project partners that opportunity.

[1] A prosumer is a both a provider and consumer of data.

3 DATA STEWARDSHIP

Data stewardship is at the core of the VAS soil data federation. In the past five years there has been a considerable shift towards data democracy in scientific journals and scientific research organisations, culminating with the publication of the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016). Since their publication, these principles have been rapidly adopted by many disciplines, with the geosciences leading the way (Nature Geoscience 2019) and many Australian Universities now mandating FAIR data (e.g. Curtin University, Federation University).

FAIR guidelines stipulate that scientific data shall be:

Findable - This requires that soil data and metadata are easy to find by both humans and computers. Persistent identifiers and machine-readable metadata are essential for automatic discovery of relevant datasets and services.

Accessible - For accessibility, the limitations on the use of data, and protocols for querying or copying data must be made explicit for both humans and machines. This requires data stewardship model for custodians to clearly set the rules under which access to their data, or parts of their data, is possible. FAIR does not mean open, as some data may be made accessible at a cost, or under stipulated data agreements.

Interoperable - Put simply, interoperability means that a computer can interpret the data, so that they can be automatically combined with other data. Data interoperability is usually specified as technical, syntactic, schematic and semantic interoperability in increasing order of complexity.

Reusable - This means that data and metadata are sufficiently well described for both humans and computers, so that they can be replicated or combined in future research. To be reusable requires the data to be tagged with a licence (e.g. Creative Commons), metadata and a standard, so that future discoverers of the data can clearly understand the data provenance, quality, and conditions of reuse.

As with the Governance model, a considerable effort has been invested in Data Stewardship arrangements for the AgReFed project. It is recommended that, to the extents possible, the AgReFed guidelines are adopted as the framework for the VAS project. A more detailed explanation can be found in Box et al. 2019 The Guidelines for the development of a Data Stewardship and Governance Framework

The key components of the proposed VAS Data Stewardship framework are the FAIR Guiding Principles for scientific data management and stewardship (FAIR) and the Core Trust Seal Data Repositories Requirement (Core Trust Seal). There are established tests for data FAIRness that have been published (e.g. The FAIR self assessment tool and Go FAIR). The Core Trust Seal lists the requirements to meet the Core Trustworthy Data Repository Requirements co-developed by the International Science Council (ISC) World Data System and the Data Archiving and Networked Services Data Seal of Approval.

Certification as a Core Trust Seal repository complies with International Standards Organisation (ISO) standard ISO 16363. Despite being challenging to attain, aspiring to Core Trust Seal certification starts with a self-assessment that appraises a repository against a list of requirements considered as the essential elements of trustworthiness. This allows data custodians to reflect on the state of their data repositories and potential for improvement.

3.1 Data stewardship in the VAS federation

As a generalisation, it could be expected that the federation members may adopt, over time:

- The FAIR data principles and guidelines (e.g. the FAIR data test)
- CoreTrustSeal repository requirements (if they are serving data from their own repositories)
- Adoption of, or alignment with, global standard data schemas and services, including existing services to serve and consume controlled vocabularies
- Spatial data on the Web Best Practices (W3C/OGC)

While the ultimate intention is to establish VAS as a truly federated system architecture, not every data provider, consumer or prosumer currently has the capability to establish the technical infrastructure required to serve and consume accredited FAIR data services from accredited trusted repositories in international standard formats. Most of the VAS industry partners (i.e. the grower groups and CMAs) are currently unable to meet those requirements.

Hence, to facilitate the implementation of technology stacks for these data prosumers, it is intended to construct a reference architecture that would allow a provider (node) to jump into the 'software stack' wherever they needed to, for example:

- For small groups starting with nothing (e.g. a farmer group, community group, SME), this would provide a service-oriented data platform – 'of last resort'. Designed to be easily implementable using open-source components (e.g. PostGIS/PostgREST), the service-oriented platform would comprise a self-serve system for exposing data to the federation, with switchable access control rules to determine the data accessibility. To the extents possible, metadata will be autogenerated with the data submission.
- For a provider with an existing service (i.e. many of the Universities, and open data providers such as the CSIRO and State agencies), there are multiple options, e.g. implement best practices for Linked Data (increasing FAIR-ness) with the use of persistent identifiers, controlled vocabularies, etc. or make use of a 'mediator' component of the reference architecture to 'standardise' the outputs.

It is envisaged that into the future, some of the data prosumers may choose to invest in other options for their data management and provision (e.g. establish a data cooperative, or proprietary system tailor-made to meet their requirements). It is the intention of VAS to assist where possible to ensure that data stewardship guidelines are respected. Documented guides (the recipe and lists of ingredients) can be grown as the federation community co-designs and co-develops a shared toolkit to lower the overheads and barriers to becoming a prosumer within the VAS, while improving the FAIR-ness of the soil data being served from trusted repositories and consumed by the end-users.

4 DISCUSSION

The material presented in this discussion paper was discussed by the VAS project Steering Committee at their meeting on 13 November 2019 in Melbourne.

4.1 VAS soil data federation governance

In terms of governance, the Steering Committee recommended that the decision is deferred until more information and experience is gained with the VAS portal and current development of the VAS project. In other words, the Steering Committee remain a *de facto* governing body until the members of the nascent soil data federation – that is, the current VAS project participants – gain more first-hand knowledge, understanding and experience in using the portal.

In these discussions, the uncertainties in the governance of the portal were expressed in terms of the risks, such as the likelihood of data being hacked by external miscreants, and the consequences of such events. In addition, the risks associated with legacy data being orphaned if federation member organisations (as data providers) cease to exist or no longer provision web-services. Since the overall intent is for farming system groups to have improved soil data management capacity, these risks should be able to be managed and mitigated by the VAS governance rules in the future.

Apart from the risks, other points of discussion were raised in relation to the rules for allowing data providers from outside the Soil CRC to join the soil data federation. For example, it could be envisaged that other farmer groups, advisors, businesses or researchers may see the advantages and value proposition in joining the VAS soil data federation as prosumers. The benefits of high volumes of high-quality FAIR soil data being provisioned are clear to all, whether they are participants of the Soil CRC or not.

Further discussions by the Steering Committee resolved that the provision of FAIR data in the VAS federation should not be solely the responsibility of the current VAS project participants but should extend to all soil data being collected in the projects funded by the Soil CRC. In that respect, the current call for Soil CRC projects includes a project to address this. Also raised was the concern that the flow of data should not be one-way – from the farming groups to the university researchers - but a two-way contribution where university research partners also provision their FAIR soil data to the soil data federation. There is believed to be a considerable volume of high-quality data in legacy research projects that could be made available by universities.

4.1.1 Branding

In relation to branding and implementation, it was acknowledged by the Steering Committee that almost all the project partners desire that the portal is embedded in their website. In this scenario, a farming group member opens the data map on their farming group website and sees the data for their region first, according to the access rules of the data custodians. The web portal map is branded like the host organisations website. The intention is to allay mistrust and suspicion around storage and sharing of data and accessing data from a member organisation is more trusted. The Steering Committee acknowledge that strong acknowledgement of the Soil CRC is still required as is a central point for VAS access (data.soilcrc.com.au).

On the VAS portal, rotating logos for partners could implemented to recognise the contributing partners. The inclusion of data from others (CSIRO, TERN, state agencies) should also be acknowledged and may require the implementation of service agreements.

Further discussions with the Soil CRC is required to resolve the eLibrary and web-mapping components and two-way linkages between Soil CRC and VAS portals.

4.2 VAS soil data stewardship

In principle, the Steering Committee endorsed the suggested VAS soil data stewardship intention to build FAIR data tests and Core Trust Seal requirements into data management practice. For the farmer groups and catchment management authorities contributing data, the goal is to make it less onerous, initially through the implementation of the cloud-based technical architecture and self-serve system.

The discussion extended to a broader conversation around data management for the Soil CRC and the requirement for a cloud repository or data container. While this logically would become a component of the VAS soil data federation, there were suggestions that it could also provide a safety back-up system for federation member data, if appropriate.

By aspiring to the Core Trust Seal certification, assessment of data FAIRness, and assisting with the development of metadata will vastly improve the chances of legacy data being rescued and made reusable.

5 CONCLUSIONS

The following conclusions are drawn in relation governance and data stewardship for the VAS project:

1. The preferred model for the VAS project is based on federated architecture in which the data remains in the control and management of the data custodians, who have the control over who can access the data and the licensing conditions. This overcomes many of the barriers to sharing soil data.
2. Although the aspiration for VAS is as a fully federated model, it is recognised that not all the participants (i.e. federation members) have the technical capability from the outset to interoperably serve their data according to the required standards. Therefore, as an interim solution, a brokered model will be offered in which participants can use a self-serve system to provision their data through a secure cloud-based data host, according to the international data standards.
3. While it is recommended that the governance of the VAS soil data federation is based on the AgReFed model, the VAS Steering Committee have deferred the decision on governance until a greater first-hand knowledge, understanding and experience with the VAS portal and its operation is acquired.
4. It is logical that the VAS has a central access point (i.e. data.crc.com.au) but it is recognised that most participants desire that the portal is embedded in their organisation's website, since that is the logical trusted place for their members to find soil data and information. Embedding the VAS portal in their websites also fulfils a value case in supporting group membership, as it offers value and services to members.
5. The VAS project is strongly committed to good data stewardship and will work with data custodians in implementing the guidelines to make their data findable, accessible, interoperable and reusable (FAIR), including standard assessments for FAIRness. Also encouraged is the certification of data repositories with the Core Trust Seal, hence complying with International Standards Organisation (ISO) standard ISO 16363 as a trusted repository.

ACKNOWLEDGEMENTS

This research is funded by the CRC for High Performance Soils and supported by the Cooperative Research Centres program, an Australian Government initiative.

Additional funding and in-kind support are provided by:

Federation University Australia

Manaaki Whenua Landcare Research

University of Southern Queensland

University of Tasmania

Birchip Cropping Group

Burdekin Productivity Services

Central West Farming Systems Inc.

CSIRO

Gillamii Centre

Herbert Cane Productivity Services Ltd

Holbrook Landcare Network

Nutrien (formerly Landmark)

Liebe Group

MacKillop Farm Management Group

Mallee Sustainable Farming

NSW Department of Primary Industries

North Central Catchment Management Authority

Riverine Plains Inc.

Southern Farming Systems

Western Australia No Till Farmers Association

West Midlands Group

Wimmera Catchment Management Authority

The author acknowledges the time, ideas and hospitality contributed by the farming groups and catchment management authorities, and colleagues who contributed time and ideas to this project.

CRC FOR HIGH PERFORMANCE SOILS

University of Newcastle
University Drive
Callaghan NSW 2308

T +61 2 4921 5473
E enquiries@soilcrc.com.au
W soilcrc.com.au

SOiL CRC
Performance through collaboration

Australian Government
Department of Industry, Science, Energy and Resources
Business Cooperative Research Centres Program