



Performance through collaboration

## FINAL REPORT

# VISUALISING AUSTRALASIA'S SOILS: A SOIL CRC INTEROPERABLE SPATIAL KNOWLEDGE SYSTEM

PHASE 1 (2019 – 2021)

PROJECT 2.3.001

Authors	Peter Dahlhaus, Andrew MacLeod, David Medyckyj-Scott, Bruce Simons, Christiane Bahlo, Amie Sexton, Helen Thompson, Megan Wong, Nathan Robinson, Rob Milne, Angela Neyland, Heath Gillett, and Alison Ollerenshaw.
Title	Visualising Australasia's soils A Soil CRC Interoperable Spatial Knowledge System Phase 1 (2019 – 2021)
ISBN	N/A
Date	June 2021
Keywords	Soil data, FAIR data principles, soil information models, social research, semantic web.
Publisher	N/A
Preferred citation	Dahlhaus P., MacLeod A., Medyckyj-Scott D., Simons B., Bahlo C., Sexton A., Thompson H., Wong M., Robinson N., Milne R., Neyland A., Gillett H. and Ollerenshaw A. (2021) Visualising Australasia's Soils: A Soil CRC Interoperable Spatial Knowledge System. Phase 1 (2019 – 2021). Project 2.3.001 Final Report. Cooperative Research Centre for High Performance Soils (Soil CRC). 119p.

## CONFIDENTIAL

Not to be distributed beyond the CRC for High Performance Soils (Soil CRC) Participants and Affiliates without the consent of the CEO.

## DISCLAIMER

Any opinions expressed in this document are those of the authors. They do not purport to reflect the opinions or views of the Soil CRC or its partners, agents or employees.

The Soil CRC 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 Soil CRC, 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.

## PEER REVIEW STATEMENT

The Soil CRC recognises the value of knowledge exchange and the importance of objective peer review. It is committed to encouraging and supporting its research teams in this regard.

The author(s) confirm(s) that this document has been reviewed and approved by the project's steering committee and by its program leader. These reviewers evaluated its:

- originality
- methodology
- rigour
- compliance with ethical guidelines

- conclusions against results
- conformity with the principles of the [Australian Code for the Responsible Conduct of Research](#) (NHMRC 2018), and provided constructive feedback which was considered and addressed by the author(s).

# PROJECT PARTICIPANTS



# CONTENTS

Project Participants .....	iv
Executive Summary.....	1
Introduction .....	4
Background.....	5
Methodology .....	13
Results.....	19
Discussion.....	40
Conclusion .....	44
Recommendations .....	45
Acknowledgements.....	47
References.....	48
APPENDIX A    Review of State soil repositories .....	52
APPENDIX B    Information model .....	64
APPENDIX C    Data entry headers.....	95
APPENDIX D    Data and Vocabulary mapping.....	101
APPENDIX E    VAS API .....	111

# EXECUTIVE SUMMARY

Visualising Australasia's Soils (VAS) is a foundational Soil CRC project that has an overall aim of making Australasian soils data from all sectors visible and reusable, according to agreed governance principles, so that Soil CRC participants can maximise the value of their research. The project commenced in February 2019 as a collaboration between three universities, one government agency, 13 farmer groups, two Catchment Management Authorities and one industry partner. This report details the first two years of the research project to 30 June 2021. Phase two of the project will run for three years to 30 June 2024.

Engaging with the farmer group partners has helped us understand their motivations for being involved in the project and their aspiring use-cases for the VAS system. The design of the system architecture has been based on the use-cases, resulting in a federated data supply model that includes a cloud-based data aggregator for the participant's soil data. Twelve of the 16 participants have provided 827 soil sites at which approximately 4400 samples have been taken with about 77,332 soil observations. The slow pace of data supply and the unexpectedly lengthy process of data curation required to load those data into the system has uncovered new challenges in data literacy and data harmonisation that will be a focus for the next phase of research. The barriers to soil data sharing will also be explored, including the barriers for academic researchers to share their data with the farmer groups and catchment managers.

The soil data portal was launched in December 2019. The current publicly available data includes some of the 1628 open public sector soil datasets, many of which have inadequate metadata and consequently, unknown value to soils research. A challenge for the next phase of the research is to show more of the public soils data in a relevant way and encourage the public data custodians to improve their data stewardship.

The co-development of governance frameworks for sharing data has been explored with a discussion paper. Adopting guidelines for data governance and stewardship has been deferred until the mechanics and functionality of the VAS system are fully trialled, and the subtleties of data stewardship and implications of the suggested governance structures are better understood by the project Steering Committee.

Training programs to improve data stewardship, including the collection and quality of data and metadata for inclusion in the knowledge system, was impacted by the global pandemic in 2020-21. Seven educational videos have been delivered to help the project partners use the VAS system. However, the aim of co-designing and delivering online educational materials for farmers and researchers to make best use of the data in the knowledge system was deferred to the next phase of the project. Tools to assist in spatially visualising soil data, searching and filtering data, downloading data sets and publishing data to the portal, have been trialled by project participants and their feedback will be used to enhance and grow the toolset in the next phase of the project. To date, only a couple of data custodians have temporal soil data (one with up to 13 years of data) to demonstrate the trends in soil monitoring over time.

In summary, phase one of the VAS project has successfully met the original aims even though they are not all to the level anticipated at the start of the project. However, the research has clarified the directions for the next phase of the research and fourteen recommendations have been made. These are grouped around the themes of strengthening the value proposition for the participants, improving data literacy and handling, improving the depth and breadth of the data, and developing a suite of more functional tools.

## OBJECTIVES

This project has two long-term objectives:

1. To make existing soils data more findable, accessible, interoperable and reusable (FAIR) to provide a range of benefits for research, on-farm decision making and policy development.
2. To integrate with other initiatives from the local to international scale that are aiming to liberate soils data and make it available according to the FAIR framework.

## RESULTS

The most significant result of this research to date is the clear demonstration that Australasian soils data, sourced largely from the private sector, can be made FAIR and shared subject to the access rules set by the data custodians. The project has implemented a functional and useful soil data federation system accessible via a soil data portal that includes public sector soil data.

After consulting with the project participants, a significant effort was made to design and build a data management system that meets their needs and can provide standardised data to researchers, subject to the data custodian's consent. A clear value proposition for the farmer groups and catchment managers is access to a trusted, supported, web-based spatial soil data management system that suits their purposes and is relevant to their location. However, a value case for wanting to share their data remains elusive.

An obvious lesson from this research is the generally poor level of data literacy among soil data custodians. Generally, data and metadata management and stewardship practices are often immature and *ad hoc*, with data spread across a variety of file formats and repositories, even within a single organisation. However, data that have been mapped into the system now conform to international data exchange standards and are interoperably available (subject to the data custodian's consent) for any future purpose, such as serving data (machine-to-machine) into other tools and applications, artificial intelligence engines, or decision support systems.

One barrier to data sharing by the farmer groups is that the data does not flow back to them from the researchers. Encouraging universities and research agencies to serve data to the soil data federation is considered equally important in the project. In that respect, phase one of the project failed, apart from a few datasets from Federation University being included. Third-party data could also be used to grow the breadth and depth of research data. Outside of the Soil CRC membership, there are other farmer groups, government agencies, businesses, and community groups who have expressed their desire to join the soil data federation.

Regarding the VAS portal, a major challenge still to be conquered is how best to show the massive volume of open public soil data that exists for Australasia, in a way that makes sense to the end-user. This and other challenges have been mapped for research in the next phase of the project.

<b>NEXT STEPS</b>	<b>TIMING</b>
This project moves into a second phase with Federation University, Manaaki Whenua – Landcare Research, University of Newcastle, Southern Cross University, 19 farmer groups and two catchment management authorities.	Phase two commences on 1 July 2021
Co-develop use-cases for sharing	30 July 2021
Co-develop and publish educational material on data literacy	31 December 2021
Implement soil moisture sensor feeds into VAS system	30 June 2022
Commence data input from Soil CRC projects	30 September 2022
Implement a data-sharing system with user access controls	30 December 2022
Report on data and text mining of legacy data	30 December 2022
Demonstrate significant modelling output using shared VAS data	29 September 2023
Finalise governance and stewardship	29 December 2023
Publish social research (project impact)	29 December 2023
Publish technical architecture	29 December 2023
Final reporting	30 June 2024



# INTRODUCTION

Visualising Australasia's Soils (VAS) is a foundational Soil CRC project that began in February 2019. The overall aim of the project is to make Australasian soils data from all sectors visible and reusable, according to agreed governance principles, so that Soil CRC participants can maximise the value of their research.

This document reports on the completion of the first phase of the project to 30 June 2021.

## PROJECT AIMS AND OBJECTIVES

The following six aims were proposed at the commencement of the project:

1. 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.
2. Co-develop, with CRC end-users, data stewardship governance frameworks for sharing data.
3. Collaborate with project participants to develop the capability to interoperably provide their data to the spatial knowledge system according to the rules that they set.
4. Co-design and co-develop training programs with project participants to improve data stewardship and governance, and to improve the collection and quality of data and metadata for inclusion in the knowledge system. Co-design and deliver online educational materials for farmers and researchers to make best use of the data in the knowledge system.
5. Co-develop simple, web-based tools to assist in spatially visualising soil data, searching and filtering data, downloading data sets and publishing data to the portal, to suit the needs of all Soil CRC participants.
6. Co-develop, with CRC end-users, dynamic models that are applied to the interoperably federated data to answer frequently asked questions such as finding temporal trends in soil performance indicators.

Meeting these aims will contribute to the proposed long-term objectives:

1. Make existing soils and agriculture data more findable, accessible, interoperable and reusable (FAIR) to provide a range of benefits for research, on-farm decision making and policy development.
2. Integrate with other initiatives from the local (e.g. farming data co-operatives) to international (e.g. International Union of Soil Sciences) scale that are aiming to liberate soils data and make it available according to the FAIR framework.
3. Co-develop and implement an enduring Australasian soils knowledge system that is based on principles of data democracy, self-sustaining and inherently useful for research and education.

# BACKGROUND

This project addresses the rapid increase in soil data supply, big data and its transformation into knowledge that has resulted from the adoption of digital technologies by Australasia's agricultural sector, government agencies and researchers. The VAS project has the potential to provide interoperable access to open data (typically government and research data), data provisioned by the Soil CRC members (e.g. grower groups, farmer data co-operatives, research organisations), data from Soil CRC research programs, and industry supplied data. Access to these data will greatly benefit Soil CRC research and industry partners by ensuring that all research builds on the best available and most current data sets.

The Soil CRC is in a unique position to improve Australian agricultural profitability by:

- bringing together soil data from disparate sources in both public and private sectors,
- making it FAIR (Wilkinson et al. 2016) in a 'pre-competitive space' (Antle et al. 2017a; Antle et al. 2017b)
- subsequently supplying federated, harmonised and standardised soil data and modelled derivatives to the 'competitive space' where it can be used to benefit those working in agricultural industries (Figure 1).

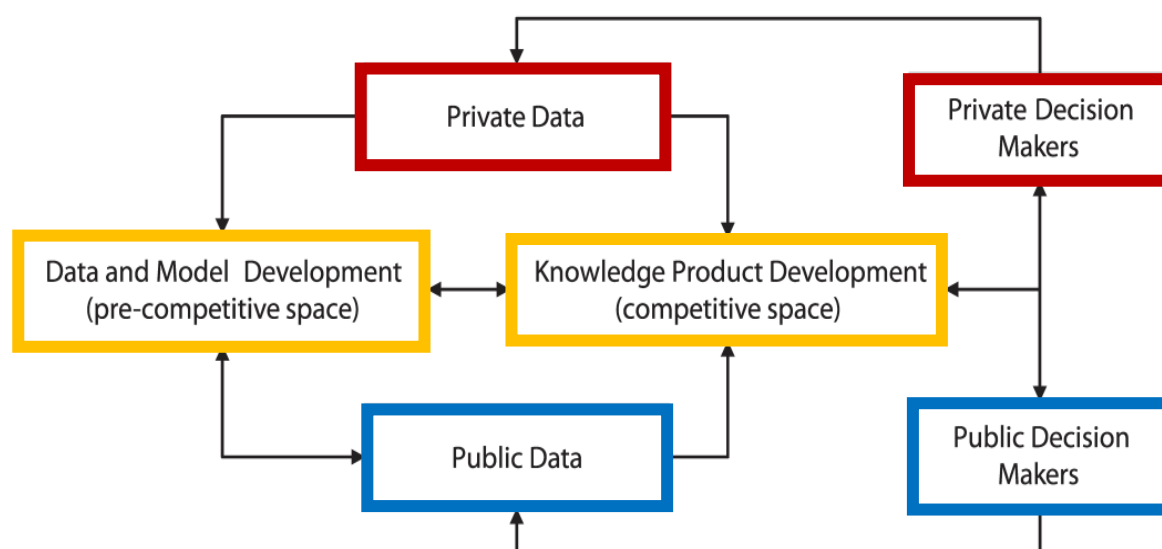


Figure 1. The high-level conceptual model for VAS from Antle et al. (2017a).

The VAS project supports many of the other Soil CRC programs, such as Program 1 (sub program 2), Program 2 (sub program 4) and Program 4 (sub program 3). It delivers research milestones in Program 2.3 and education milestones through training and online courses to improve data stewardship and governance by farmers, industry and researchers.

## SOIL DATA INFORMATION SYSTEMS

From the 1980s, soil data information systems that were historically paper-based index cards, aerial photographs and maps, were transformed to computerised databases and geographic information systems (GIS), generally improving soil data curation. Later adoption of Global

Positioning Systems (GPS) and mobile computing devices significantly improved soil data collection records. Since the ubiquitous adoption of the internet, there have been many initiatives around the globe to use digital technologies to better manage, access and visualise soils data, both in the public sector and private sector.

## PUBLIC SECTOR

Historically, soil data in the public domain were collected by state and commonwealth government agricultural departments at first to aid in the development of new farmlands (e.g. Leeper et al. 1936, Holmes et al. 1939, Downes 1949, Skene 1963). In later years, these public soil datasets have been supplemented by data collected by other government agencies and research institutions to better understand and manage a variety of issues – agricultural production and food security, land capability and suitability, soil conservation, soil contamination and pollution, water security and catchment management, geohazards risk management, infrastructure and urban development, mineral exploration, ecosystem services and the management of social and cultural heritage values. The legacy of databases constructed over time for various programs and projects by disparate departments, that came and went with the machinery of government changes, has generally resulted in poor continuity of soil data management and a plethora of legacy data archived<sup>1</sup>, lost, forgotten and ignored.

At present, Australian soil data in the public domain is held in disparate databases managed by federal and state governments and agencies, research institutions and universities, and a few non-government organisations and community groups. As a generalisation, the findability and accessibility of these data have improved considerably in recent years, but their interoperability and reusability remain poor (refer to Appendix A for a review). Most custodians have made soil data discoverable through online public metadata catalogues (e.g. data.gov.au, data.[state name].gov.au). Other national initiatives include the Australian Collaborative Land Evaluation Program (ACLEP), Australian Soil Resource Information System<sup>2</sup> (ASRIS 2020), the CSIRO National Soil Archive<sup>3</sup> (CSIRO 2020), and the data initiatives funded through the Australian Government’s National Collaborative Research Infrastructure Strategy (NCRIS) for example.

In 2021 the Australian Government committed to funding the National Soil Strategy in the federal budget, which includes a re-development of ASRIS “...to improve its ability to store soil data, track and report trends and changes in soil health, and monitor the impact of land management practices and environmental shocks over time.” (DAWE 2021). The upgrading of ASRIS is guided by the design of a National Soil Information Framework (NSIF) that has recently been released (NTT 2021).

In New Zealand, the situation mirrored that of Australia. Since the 1930s, Manaaki Whenua – Landcare Research – and its predecessor, DSIR Soil Bureau – have invested considerable time and effort establishing and describing soil profiles and undertaking soil surveys across the country. Other organisations such as Universities and local government also collected soils data but on a more ad hoc basis. A number of soil databases, for example, the National Soils Database (NSD), which held data for some 1,500 soil profiles, were created in the 1980s (MW-LR 2021b). However, funding for the collection of soils data has gone through peaks and troughs with data often being lost or unmanaged during periods where interest in soils and/or funding for research waned, and thus data collection declined. Maintenance of soil databases

---

<sup>1</sup> for example, the Victorian Soil Conservation Authority archives  
[https://prov.vic.gov.au/search\\_journey/select?keywords=Soil%20Conservation%20Authority](https://prov.vic.gov.au/search_journey/select?keywords=Soil%20Conservation%20Authority)

<sup>2</sup> <https://www.asris.csiro.au/>

<sup>3</sup> <https://www.clw.csiro.au/aclep/archive/>

also ceased with the result that soil datasets became orphaned, spread over multiple locations and in a variety of file formats.

With a renewed focus on the importance of soils information and associated funding successes in the last 10 years, funding for soils research has meant a renewed focus on soil data collection and the provision of soil information systems to both manage and publish data.

The nationally significant collections and databases Land Resource Information System (LRIS) programme has been working to revitalise the NSD creating the National Soils Data Repository which now contains some 4000 soil profiles (MW-LR 2021a). Data describing soil map units (polygons) is captured in S-map – a spatial soil database for New Zealand begun in the early 2000s and representing the response of Manaaki Whenua to the changing need for more quantitative soil information (Lilburne et al. 2012). S-map was designed as a digital product from the start. Data includes fundamental soil property data (e.g. depth, stoniness, clay and sand content) created from field observations and expert knowledge, as well as derived soil data based on models. S-map now covers 36.6 % of New Zealand, including 67.3 % of the ‘multiple-use land’ in the country. Using soil data requires significant expertise and is a barrier to wider use. Interoperability and reusability of soil data is poor; a challenge as other New Zealand agencies have started to collect soil data.

Globally, many current initiatives are working on soil data sharing, the main ones being the Global Soil Partnership Pillar 4 (FAO 2018) through their Global Soil Information System (GLOSIS); the Global Open Data for Agriculture and Nutrition (GODAN) Soil Data Working Group (GODAN 2018); the International Union of Soil Sciences (IUSS) Soil Information Standards Working Group (IUSS 2020); the ISRIC World Soil Information project (ISRIC 2021a); the OGC Agriculture Domain Working Group (OGC 2017); and the Research Data Alliance Agricultural Data Interest Group (RDA 2018). The Australian and New Zealand soil data research communities are connected with these initiatives.

## OTHER DATABASES

There are an unknown number of other soil databases outside of the open public sector group. Some examples include:

- The south-west Victorian Soil Health Knowledgebase
- Mallee Catchment Management Authority soil database
- West Gippsland Catchment Management Authority soil database
- East Gippsland Catchment Management Authority soil database
- Goulburn-Broken Catchment Management Authority soil database
- Lawsons Grains – 2016 CSIRO AgCatalyst competition soil data set
- Environmental Analysis Laboratory (EAL) at Southern Cross University
- The University of Sydney Institute of Agriculture database.

Other universities that have agricultural courses, such as the Tasmanian Institute of Agriculture, University of Southern Queensland, University of Newcastle, University of New England, University of Adelaide, Queensland University of Technology, University of Queensland, The University of Melbourne, Latrobe University, University of Western Australia, Curtin University, Central Queensland University, Charles Sturt University, and many others are likely to hold considerable volumes of soil data but it is not findable<sup>4</sup>, so presumed not to be centralised, structured or FAIR-compliant.

---

<sup>4</sup> Findable in public metadata repositories used by universities, such as Research Data Australia or Figshare.

Other initiatives include semi-private or private sector agricultural data brokers such as DataLinker (Rezare 2021) and the Soil Tech Project (Soil Tech Project 2021), and agricultural standards development bodies such as AgGateway (AgGateway 2021). Although some of these have developed data standards and schemas for interoperable data sharing, none have been accepted as international community standards.

Research institutions such as the Rural Research and Development Corporations<sup>5</sup>, Field Applied Research (FAR) Australia, Foundation for Arable Research (FAR) New Zealand, Thomas Elder Institute, etc., as well as fertiliser company laboratories, commercial soil testing service providers (such as Precision Agriculture P/L) and laboratories (such as Hills Laboratories, New Zealand), crop and pasture breeding laboratories, corporate farming enterprises, etc. are all highly likely to have soil data repositories. While some of these will be structured databases, they are not findable or accessible.

And there are a plethora of software applications that provide a level of soil data management (or functionality) that can be used by farmers and agricultural industry practitioners, most of which are coupled with farm management software, precision agriculture software and decision support systems. Many of these are available as mobile device applications (Apps), such as AgTrix (AgTrix 2021), agX® (Proagrica 2021), Back Paddock (Back Paddock Company 2021), FarmersEdge (FarmersEdge 2021), Farmer Pro (Trimble 2021), Farmlab (Farmlab 2021), FarmlQ (FarmlQ 2021), and MyEnviro (MyEnviro 2021) to name only a few examples.

## RESEARCH CHALLENGES

Compared to the abundance of international soil data information system initiatives and implementations, there is a combination of components that make the VAS research unique:

- the combined technical and social architectures that allow linking data on-the-fly from all sectors – public, industry, communities and private, according to the rules set by the data custodians
- the ability to serve soil data from disparate databases, sensors, web services and application programming interfaces (APIs), in an internationally standardised format which can be seamlessly consumed by existing interoperable data federations or decision support systems
- the spatial extent of the project to include data across Australia and New Zealand
- the co-design and co-development of data management tools with the project partners, that create value for them as end-users
- on-going evaluation of the project outputs and outcomes, using social science tools to understand end-user needs, as well as measure the practice change and value case for the end-users of the research.

The research challenges in creating and implementing a successful and enduring public-private soil data system are considerable. They include the social challenges of finding the value proposition for the participants and end-users (i.e. the providers, consumers and prosumers<sup>6</sup> of the soils data), the technical challenges of making it all work in an intuitive-to-use, seamless and effortless manner, and a business model that allows it to prosper.

---

<sup>5</sup> [https://www.agriculture.gov.au/ag-farm-food/innovation/research\\_and\\_development\\_corporations\\_and\\_companies](https://www.agriculture.gov.au/ag-farm-food/innovation/research_and_development_corporations_and_companies)

<sup>6</sup> A prosumer is both a provider and consumer

## JUSTIFICATION, NEEDS AND VALUES

Arguably, the most critical factor in achieving an enduring soil data information system is the value proposition for users. If the system can provide useful data and information as expected and needed by users, in an intuitive-to-use web portal that is openly available and free to use, then it will most likely continue to be supported by the data providers, consumers and prosumers. Hence, an important component in the VAS project is to understand the justification, needs and values for the project participants to share their data.

## GOVERNANCE

Another important factor in developing the VAS is to understand how the soil data system should be governed and how the data sharing should be governed. The intention is to implement a governance structure that will ensure:

- trust in the VAS entity and the way that it is run
- an enduring data sharing community beyond the life of the Soil CRC
- trust in the Soil CRC data and the community sharing these data
- that data custodians remain in full control of sharing their data (i.e. they set the rules)
- that shared data is not used in a way that disadvantages or penalises the data custodians
- that there are incentives to share i.e. clear rewards for the data providers.

These challenges are common to global data sharing systems and often seen as barriers to their implementation.

### *Data governance and stewardship*

The publication and global adoption of the FAIR Data Principles for Scientific Data (Wilkinson et al. 2016) have provided the basic requirements for soil data information models. While the FAIR guidelines have been rapidly implemented in many disciplines, including the geosciences (Stall et al. 2019), the soils community has been slower at adopting them, perhaps reflecting the number of private sector soil data custodians. Nevertheless, the FAIR principles have been rapidly and widely adopted by many governments and government agencies, especially in the European Union (Mons et al. 2017) and includes soil with agricultural data (e.g. FAO 2021a, ISRIC 2021b).

The largest advantage of adopting FAIR data guidelines is that it implements improved data management and stewardship, by moving the management regime from data anarchy to data democracy (Dahlhaus et al. 2017). If the soil data shared in the VAS project is to be used for input to models or other research applications, then it needs to conform to FAIR data principles (Table 1).

## INFORMATION MODELS

There are degrees of data FAIR-ness, as there are limitations in how to handle the distributed custodianship and heterogeneity of soil data. Not all (meta)data is equal in quality and completeness, which makes it difficult for the end-user to understand how to synthesise and harmonise the data for reuse in their models or research investigations.

An essential part of the solution lies in making disparate data as interoperable as possible, that is, making the data usable in a seamless manner regardless of its original heterogeneity. Data interoperability typically requires the transformation of data into a common representation readily understood by both the data supplier and consumer (Brodaric et al. 2018). Often this representation takes the form of a data standard developed and maintained

by key stakeholders, one that is typically used within two main technological environments: 1) within data networks that supply heterogeneous data to users, or 2) within workflow systems that chain together diverse software for modularised processing of data.

Table 1. The essentials of the FAIR Principles (sourced from: GO FAIR 2021).

<p><b>Findable:</b> Machine-readable metadata in an open catalogue are essential for automatic discovery of datasets and services.</p> <p>F1. (Meta)data are assigned a globally unique and persistent identifier</p> <p>F2. Data are described with rich metadata (defined by R1 below)</p> <p>F3. Metadata clearly and explicitly include the identifier of the data they describe</p> <p>F4. (Meta)data are registered or indexed in a searchable resource</p>	
<p><b>Accessible:</b> the conditions of data access, including authentication and authorisation, need to be clear.</p> <p>A1. (Meta)data are retrievable by their identifier using a standardised communications protocol</p> <p>A1.1 The protocol is open, free, and universally implementable</p> <p>A1.2 The protocol allows for an authentication and authorisation procedure, where necessary</p> <p>A2. Metadata are accessible, even when the data are no longer available</p>	
<p><b>Interoperable:</b> the data need to interoperate with other data, applications or workflows for analysis, storage, and processing.</p> <p>I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.</p> <p>I2. (Meta)data use vocabularies that follow FAIR principles</p> <p>I3. (Meta)data include qualified references to other (meta)data</p>	
<p><b>Reusable:</b> to optimise the reuse of data, metadata and data should be well-described and unambiguous so that they can be replicated and/or combined in different settings.</p> <p>R1. (Meta)data are richly described with a plurality of accurate and relevant attributes</p> <p>R1.1 (Meta)data are released with a clear and accessible data usage license</p> <p>R1.2 (Meta)data are associated with detailed provenance</p> <p>R1.3 (Meta)data meet domain-relevant community standards</p>	

Despite several past attempts and continued efforts, there is currently no international standard (schema) for soil data exchange. There have been proposed models through the European Union INSPIRE system, and there is a proposed International Organization for Standardization ISO/DIS 28258 standard for soil quality. A generalised soil data schema, known as soil markup language or SoilML model, was being progressed through the IUSS Working Group on Soil Information Standards (WG-SIS) (IUSS 2020) but has now become inactive. ANZSoilML is an Australian–New Zealand standard for the exchange of soil data that was developed as a contribution to the WG-SIS, in the absence of any other data schema (Simons et al. 2013).

At present, the most activity is in the Global Soil Partnership where the Global Soil Information System (GLOSIS), envisioned as a federation of national soil information systems, is being developed (FAO 2021b). This work is linked to the Global Soil Partnership Pillar 4 (FAO 2018). The other active initiative is the Federation of Earth Science Information Partners (ESIP) Soil Ontologies and Informatics group (ESIP 2021).

The challenge faced by the VAS project is to take varying data content and formats from a variety of data providers and make it available to potential users in a standard format, with standard content, via a standard mechanism. That is, to make it more FAIR. Trying to simplify and standardise the data discovery and delivery process makes the data import process, the data structure and content mapping more complicated.

## SYSTEMS ARCHITECTURE

Systems architecture for the Australian Government’s NCRIS funded data initiatives has been described by Box et al. (2015) who proposed several models of data sharing as illustrated below (Figure 2).

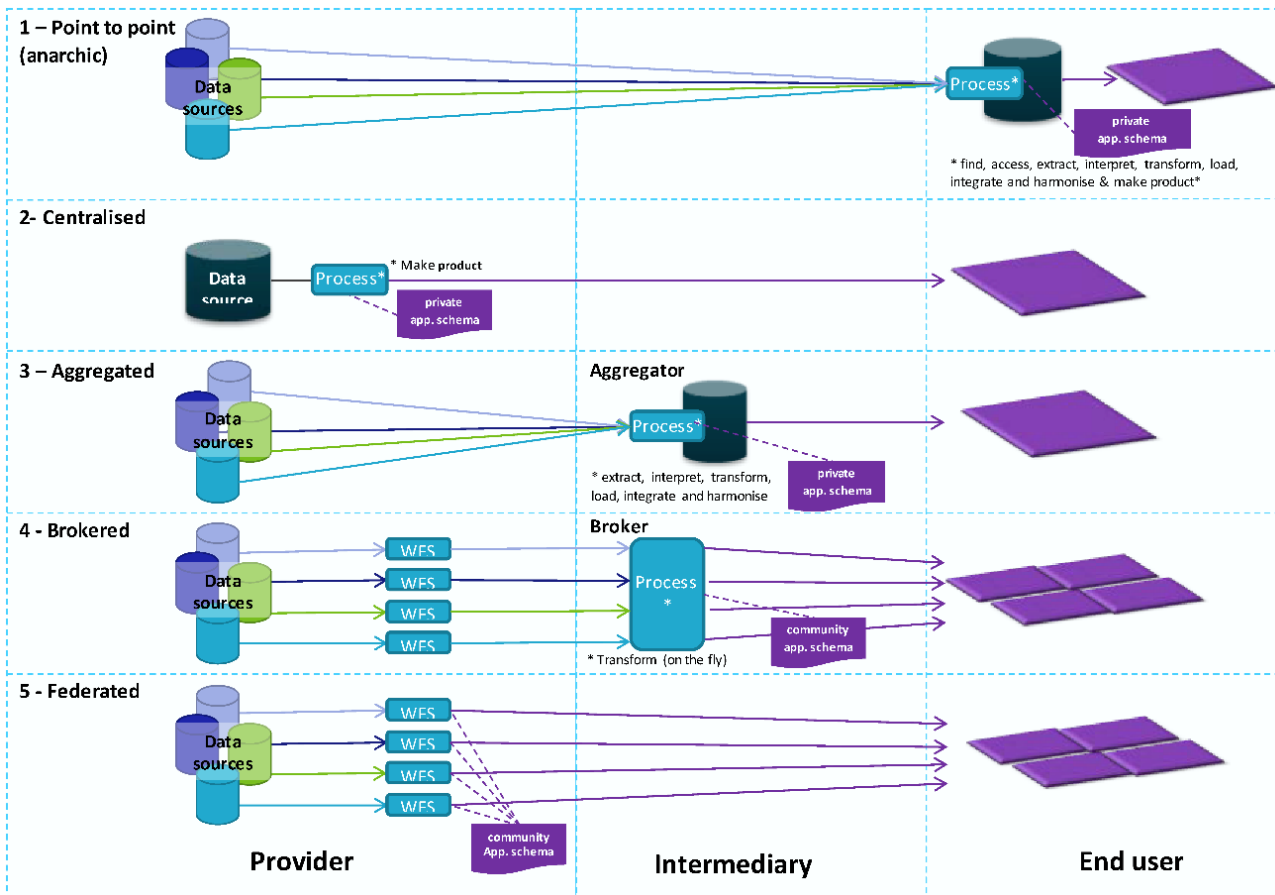


Figure 2. Data supply models (Box et al. 2015)

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 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 3). 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 (i.e. FAIR data). 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.



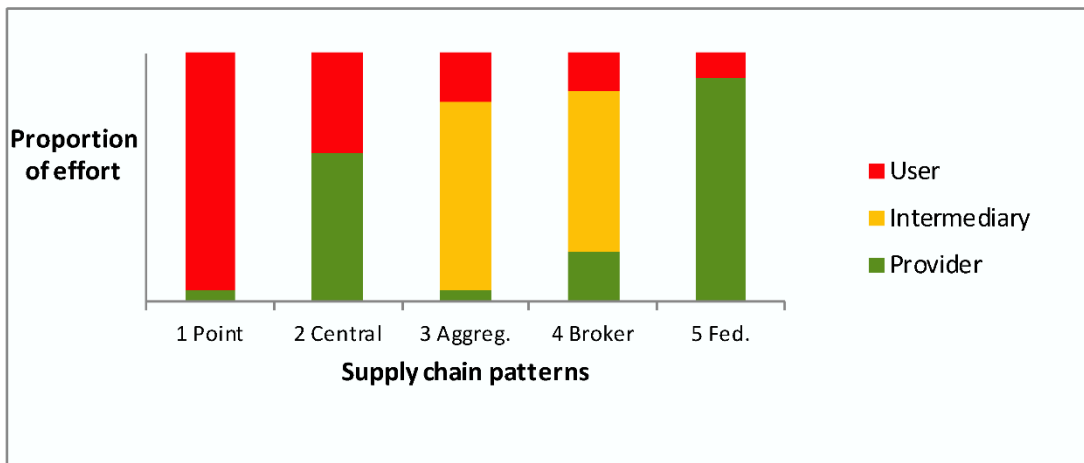


Figure 3. Relative effort for the stakeholders in each supply chain model (Box et al. 2015).

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 required data supply chain capability from the outset, especially the farmer groups. Therefore, in the interim, a combination of models will be required.

Where data providers are currently unable to directly deliver their data to a federated VAS system, Federation University acts as an intermediary. This involves taking the provider's data, aligning it with other VAS data, and delivering it as per the standards established for VAS data. As the system progresses it will move more toward a federated system, with more of this intermediary work undertaken by the data providers.

# METHODOLOGY

The methodology and methods adopted in the VAS project generally relate to the research challenges:

- project governance
- partner engagement, aspirations and use-cases
- systems architecture and implementation
- data supply chains and data stewardship.

## PROJECT GOVERNANCE

A Steering Committee provides governance for the VAS project. The establishment of the Steering Committee was written into the original project application and endorsed at the inaugural project workshop. The role of the Steering Committee was set out in the Terms of Reference adopted in November 2019. The committee comprises:

- 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-Tillage 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.

Their role is to provide expert advice to the project research and development team, both acting as a sounding board for the ideas and approaches that will be tested and refined, and as a body to direct the project development to ensure that the project aims and goals are met.

The Steering Committee was established in March 2019 and the first in-person meeting was held in Melbourne on 13 November 2019. A second Steering Committee meeting was held using online video facilities on 11 December 2020. Informal consultations with members of the committee have been held (*ad hoc*) as the need arises.

## PARTNER ENGAGEMENT

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. The main outcome of the workshop was an agreement that the project should deliver a clear case for rolling out the project as a fundamental foundation activity in the Soil CRC. In that respect, 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, time was devoted in the initial year of the project to visiting each of the project participants at their office to elicit their views on the value proposition for the VAS project (Table 2, Figure 4). 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 2. 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-Tillage 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



Figure 4a Project engagement. Clockwise from top left: Inaugural workshop at Ballarat, Riverine Plains at Mulwala, North Central CMA at Huntly and Holbrook Landcare Network at Holbrook.



Figure 4b. Participant engagement. Clockwise from top left: Herbert Cane Productivity Services Limited at Ingham, Gillamii Centre at Cranbrook, Burdekin Productivity Services at Ayr, MacKillop Farm Management Group at Naracoorte, Central West Farming Systems at Condobolin, West Midlands Group at Dandaragan and West Australian No-Tillage Farmers Association at Floreat.

The social engagement methods have been documented in a report available from the Soil CRC website (Sexton 2020). Each meeting took around three hours during which time the VAS team got to know the project partner organisation by exploring their soil issues and needs, and understanding their expectations, motivations, and concerns about the project. The team outlined the philosophy behind the project, explaining the vision and how it is likely to progress and explored how it could meet the participant's expectations, where it falls short, and how best to solve problems and allay concerns.

Towards the end of each meeting a short video interview was conducted with only one person (usually a social researcher) from the VAS team to encourage the participants to be at ease and to talk freely about their involvement in the project. The interview questions were:

1. Why are you participating in this project?
2. What are you hoping to get out of being involved?
  - short term / long term
  - individual / organisation / sector
3. Do you have any concerns about the project or your involvement?
  - Are there any challenges? Specific or general
4. What do you see as the opportunities and possibilities of this project to change things positively in the sector?
  - How?
  - What?
5. In terms of project management, what will help to make your involvement easy and positive?

The information from the meetings and interviews was then collated, analysed and summarised into the value propositions and use-cases (Sexton 2020).

## **SYSTEMS ARCHITECTURE AND IMPLEMENTATION**

The systems architecture was designed to meet the following general requirements, determined at the initial workshop:

- is a working portal with farmer group and industry group data across the nation
- is a free data portal that is intuitive-to-use and provides what is expected
- has functions that value-add to data, for example by showing trends
- has a greater emphasis on point data than raster data (although both are important)
- data custodians to be in full control of sharing their data (i.e. they set the rules)
- provides an effortless and seamless means of exposing data in the federation
- includes a self-serve system for data capture with access control
- provides training, helpdesk and user support
- the system to be technology neutral to adapt to all potential applications
- has the capability to feed data to existing decision support applications.

The systems architecture was designed to use a hybrid of data supply models (i.e. those in Figure 2) with a spatial web-mapping interface, or portal, to view the data. Open data is visible

to the public, but participants' data is made private by default, visible to nominated individuals via authentication in a standard internet browser, as preferred by all project participants.

To build the system, the following use-cases were required:

- 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 deliver data according to the FAIR data principles
- the capability to easily add soil data from a spreadsheet, or another 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
- the ability to change access controls for an organisation's data set or remove a data set
- the ability to filter data by soil data fields (e.g. show  $\text{pH} \geq 4.5$ , show % clay > 50 %)
- the ability to download data (or filtered data results) according to the access controls set by the custodians
- the ability to create pdf files of map views or data tables
- the ability to save bespoke map views as an emailable link
- the ability to save bespoke map views as Google Earth format (\*.kml), ESRI shapefile (\*.shp) and/or GeoTIFF (\*.tif)
- the ability to report on all soil data available for a property, user-selected polygon, line with a designated buffer, and point with a radius
- graphs that can show trends over time where sufficient data exists
- visualisations of soil data down a profile

A service-oriented architecture (SOA) based on tried and tested open source technology components has been successfully designed and deployed.

SOA is described by Mahmoud (2005) as *"...an architectural style for building software applications that use services available in a network such as the web. It promotes loose coupling between software components so that they can be reused. Applications in SOA are built based on services. A service is an implementation of a well-defined business functionality, and such services can then be consumed by clients in different applications or business processes."*

For data brokered by VAS, these services are delivered using the PostgREST API, which provides a REST-full machine-readable interface to the underlying data schema. As distinct from many other APIs, the data delivered via these services is structured to align with well-

known international standards for observational data and uses controlled vocabularies and linked data methods to improve interoperability and reusability. For software developers or data integrators, the endpoints and available functions are fully documented using the OpenAPI Specification (formerly swagger).

All access to the API is routed through a secure URL proxy (or router) where authentication and permission checks are handled. Access to private data is only permitted where a valid Bearer Token is provided. The system checks the permissions of the authenticated user against the access rules of the data or function being requested. Where datasets have been made publicly accessible by the group, a public token must still be used to access data via the API.

## DATA SUPPLY AND STEWARDSHIP

Soil data supplied to the VAS system can be generally categorised into four categories:

1. The publicly available (open) 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.
4. Soil data from sources outside of the VAS participants.

Data from public (open) sources are accessed using the interoperable technologies that are available from those suppliers, such as web services or API endpoints.

For the data supplied by the project participants, Manaaki Whenua – Landcare Research (MWLR) and Federation University (Federation) were the only ones who supplied data using interoperable technologies. All other participants generally supplied data as spreadsheets via email or a data delivery service (e.g. Dropbox). In a few cases, the data were in digital document files (\*.pdf) or GIS format files (\*.shp). The general process for handling the participant data is shown below (Figure 5):

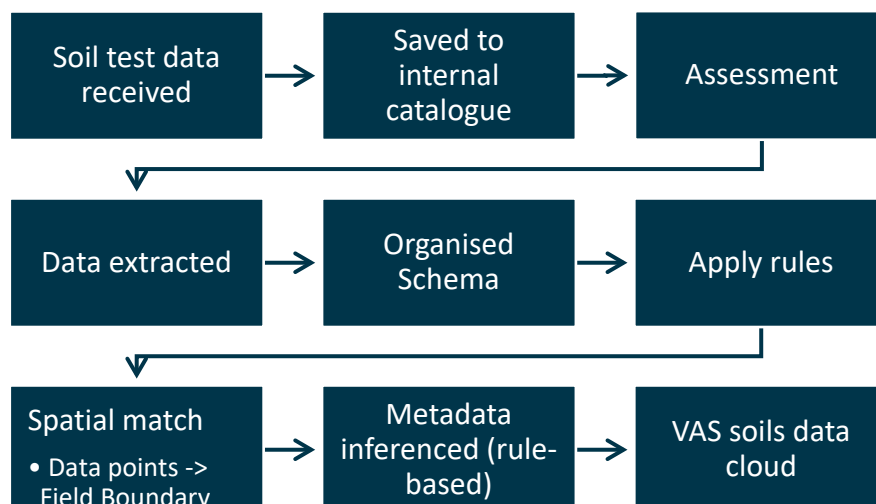


Figure 5. The general flow of soil data processing from participants.

The participating groups were each offered \$10,000 to compensate for their time in the project. This was issued via a subcontract with Federation University (the lead organisation of the project) that allocated \$5,000 on receipt of a sample data set, representative of the type of

data that the participants might wish to include in the VAS system. On receipt of the data, the VAS team assessed them and if required, contacted the data custodian or supplier to clarify any necessary details for FAIR metadata requirements, such as data ownership, data licensing, sample locations, test methods, units of measure, or sample dates.

Once their data were loaded into the system, the participant group were issued with a unique login and password to the portal. The remaining \$5,000 was provided once the group provided feedback on the portal functionality and suggested improvements to the system.

## RESULTS

### VAS GOVERNANCE

The question of how the VAS data system should be governed was outlined in a discussion paper delivered to the Steering Committee in November 2019 (project Milestone 4) and published by the Soil CRC in December 2019. It is available in the Members Area of the Soil CRC website (Dahlhaus 2019).

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

However, it is recognised that not all the participants have the technical capability from the outset to interoperably serve their data according to the required standards. Therefore, as an interim solution, a soil data aggregator has been constructed (refer to the Technical Architecture section on page 21) which allows the participants to self-load their data into a secure cloud-based database. Their data can then be provisioned to the federation, according to international data standards.

The discussion paper recommended that the VAS adopt the Agricultural Research Federation (AgReFed) governance structure as an exemplar for the 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.

### VALUE PROPOSITIONS

The main result from the social engagement is a much deeper appreciation of the diversity of the partner organisations, in their geographies, services and resources. Their soil issues varied across the spectrum of farming systems, climates, terrains and soils. One implication of this diversity is that the VAS project participants vary considerably in the resources and capability that they can contribute to the project.

Some farmer groups have a dozen staff members, including technology experts, while others may cover a similar geographical spread with a fraction of one person's time and rely heavily on government extension officers or consultant agronomists to undertake their research and provision services. Some groups have significant member subscriptions, while others have none. And some groups are strongly focused on field trials or service provision, while others are more focussed on extension and education. Hence, the variation in responses (and expectations of the VAS project) from across the groups is much greater than anticipated.



From the meetings with each of the project participants, the following general desires and use-cases for their participation in the project were expressed.

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 (a \*.pdf file) 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 member or stakeholder queries.
  
2. Having access to locally relevant data to enhance 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 soil moisture trends, groundwater levels, seasonal forecasts, terrain/drainage, 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, greenness (NDVI) maps, feed-on-offer (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.

While not all of these functions or tools are in scope for the first phase of the project, they can inform the direction for further developments in later phases of the VAS.

## TECHNICAL ARCHITECTURE

The VAS data portal was launched in December 2019 (project Milestone 5) and is accessible through the Soil CRC website (or [data.soilcra.com.au](http://data.soilcra.com.au)). At present, the public view shows some of the openly available soil data from Australia and New Zealand. The participants have access through their login where they manage their own data, loading it via a self-serve system, and use tools to filter, search, graph, download data and print reports.

## SYSTEMS ARCHITECTURE

The technical architecture of VAS is based on achieving semantic interoperability, meaning that there is potential for soil data to be served from VAS and consumed by other systems (such as AgReFed), according to the access controls set by each data custodian; and for VAS to consume data from any other interoperable system, according to the access controls set by their data custodians. The general overview of the VAS systems architecture is illustrated below (Figure 6).

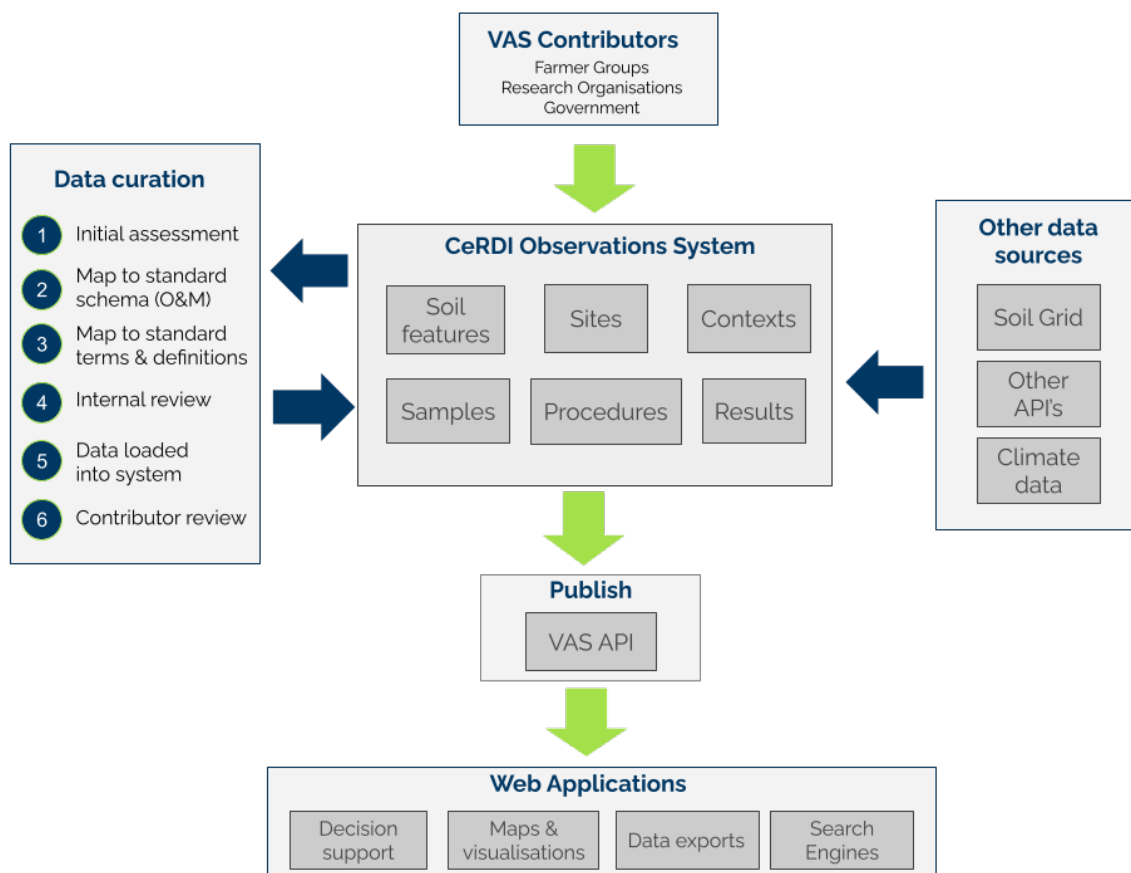


Figure 6. A schematic overview of the VAS system architecture.

The constructed architecture includes an aggregator (the CeRDI Observations System) that allows participants to use a self-serve system to upload their data into a secure cloud-based data host, where it is mapped to the international data standards. This component places an unanticipated heavy burden on the VAS project team (an intermediary as shown in Figure 3), which is inevitable until the data literacy of the project participants increases and the data management via the self-serve system becomes entirely seamless.

## INFORMATION MODEL

The most significant effort that has gone into VAS has been in co-designing, co-developing and constructing the observations system (the data aggregator) for the private (and public<sup>7</sup>) sector data that cannot yet be independently served to the soil data federation with the required degree of data stewardship (Figure 7). In simple terms this means providing a data curation service (Figure 6 left-hand side) that transforms a spreadsheet of soil test data loaded by a custodian into FAIR data by mapping it to a compliant information model (Figure 7).

The primary purpose of the observations system is to collect and store observation and measurement data and publish this data for use by researchers, project participants, and potentially the broader public. It is based on the International Standard ISO19156 and OGC Observations and Measurements (O&M) model to store field and laboratory environmental data in a domain independent structure.

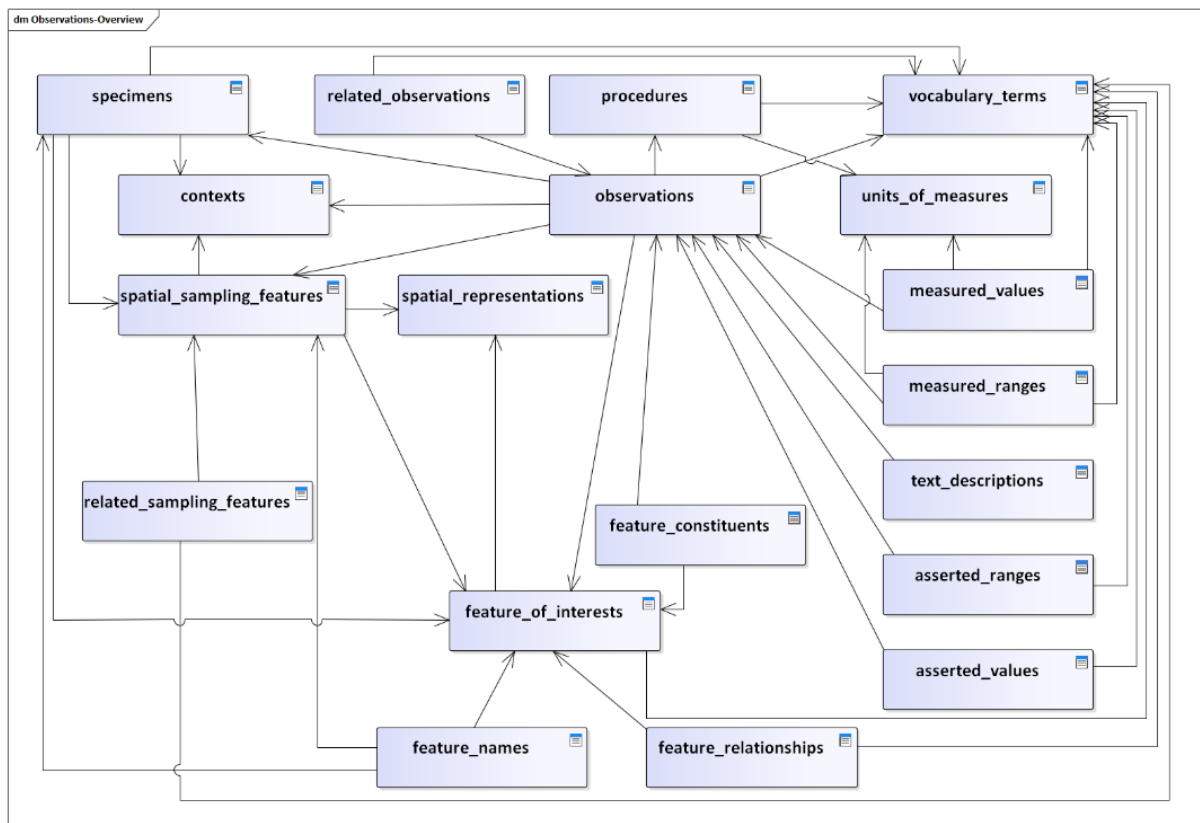


Figure 7. The database tables relating to the VAS observations system.

<sup>7</sup> The catchment management authorities may regard their data as public.

The observation model states:

‘An **Observation** is an action whose **result** is an estimate of the value of some **property** of the **feature-of-interest**<sup>8</sup>, obtained using a specified **procedure**’.

The key insights are:

- to separate the *observation act* from the *procedure* (which may be used for other observations) and the *feature-of-interest* (which has many properties, the values of each of which may be estimated more than once, at different times or using different procedures), and
- to recognise that the outcome of an Observation is a *result*, the value of which constitutes an estimate of a value of a property (which may be a value or range of values if a measurement, or a term, a term range or a description if an assertion).

In addition to standardising the data structure, the observations system makes use of existing domain-specific controlled vocabularies and ontologies to standardise the semantic content.

The fully documented information model is appended (Appendix B).

### *Controlled vocabularies*

The ‘vocabulary\_terms’ and ‘units\_of\_measures’ tables (Figure 7) are central to providing controlled terms to the concepts stored in the database. These two tables play the role of ‘look-up tables’, with additional associated information and links to external references and ontologies for the concepts. The intention is that external parties manage these terms using external applications that then make the vocabularies available via web services.

For vocabularies, these may be terms from traditional look-up tables such as terms associated with laboratory methods, such as ‘pH<sub>1:5 water</sub><sup>9</sup>’, or with descriptive terms such as ‘well-drained’ or ‘poorly-drained<sup>10</sup>’. By comparison, the units of measure table is based on the Quantities, Units, Dimensions and Types Ontology (QUDT). It allows specifying the label (e.g. ‘metre’), its abbreviation (e.g. ‘m’), alternative labels (e.g. ‘meter’), and the kind of measure the unit of measure relates to (e.g. ‘length’). The table also caters for the identifiers (Uniform Resource Identifiers, or URIs) for the unit of measure and its quantity kind. Appendix D provides more detail.

## PORTAL FUNCTIONALITY

The data portal functionality in Phase 1 includes a basic data input self-serve system and commonly requested data visualization or output functions. The intention is to further co-design and co-develop these functions in the following phase of the project.

### *Data input*

Given the commitment to good data stewardship, much of the effort has been in working with data custodians to make their data FAIR. Using the initial data that the project participants submitted, data entry checklists and templates were iteratively refined to cover all the variation in the various data sets (Appendix C). These data entry tools are accessed by the ‘submit data’ function in the Partner Dashboard upon login (Figure 8).

---

<sup>8</sup> For VAS, the feature-of-interest is the soil feature, such as a soil layer, soil horizon, soil profile, or soil body.

<sup>9</sup> <http://registry.it.csiro.au/def/soil/au/scma/4A1>

<sup>10</sup> <http://registry.it.csiro.au/def/soil/au/asls/soil-prof/soil-water-drainage>

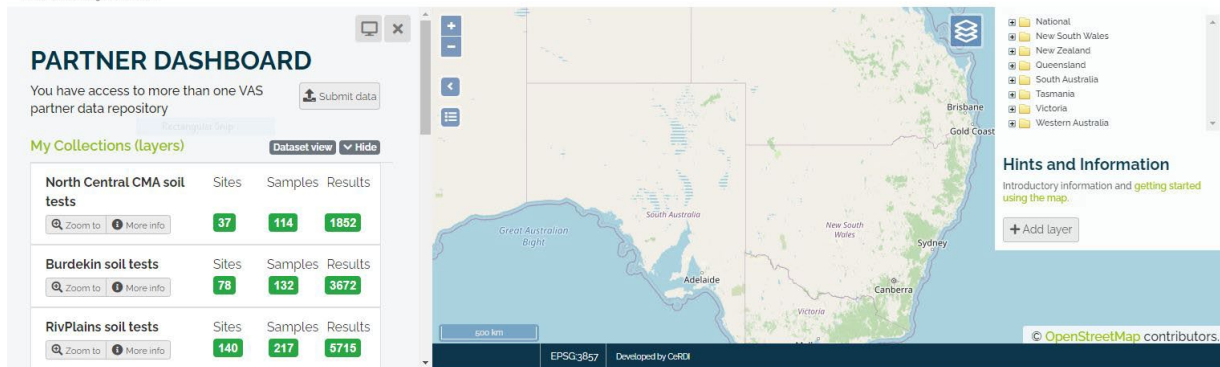


Figure 8. Partner dashboard (administrator view).

Data entry has been unexpectedly complex because of the variety of data types and formats submitted into the observations system (refer to Appendix D). The current process relies on manually checking and manipulating the data into the data schema and FAIR compliance (Figure 9). However, it is anticipated that as the data literacy of the project participants increases and their data disparity decreases, the process will be ultimately seamless.

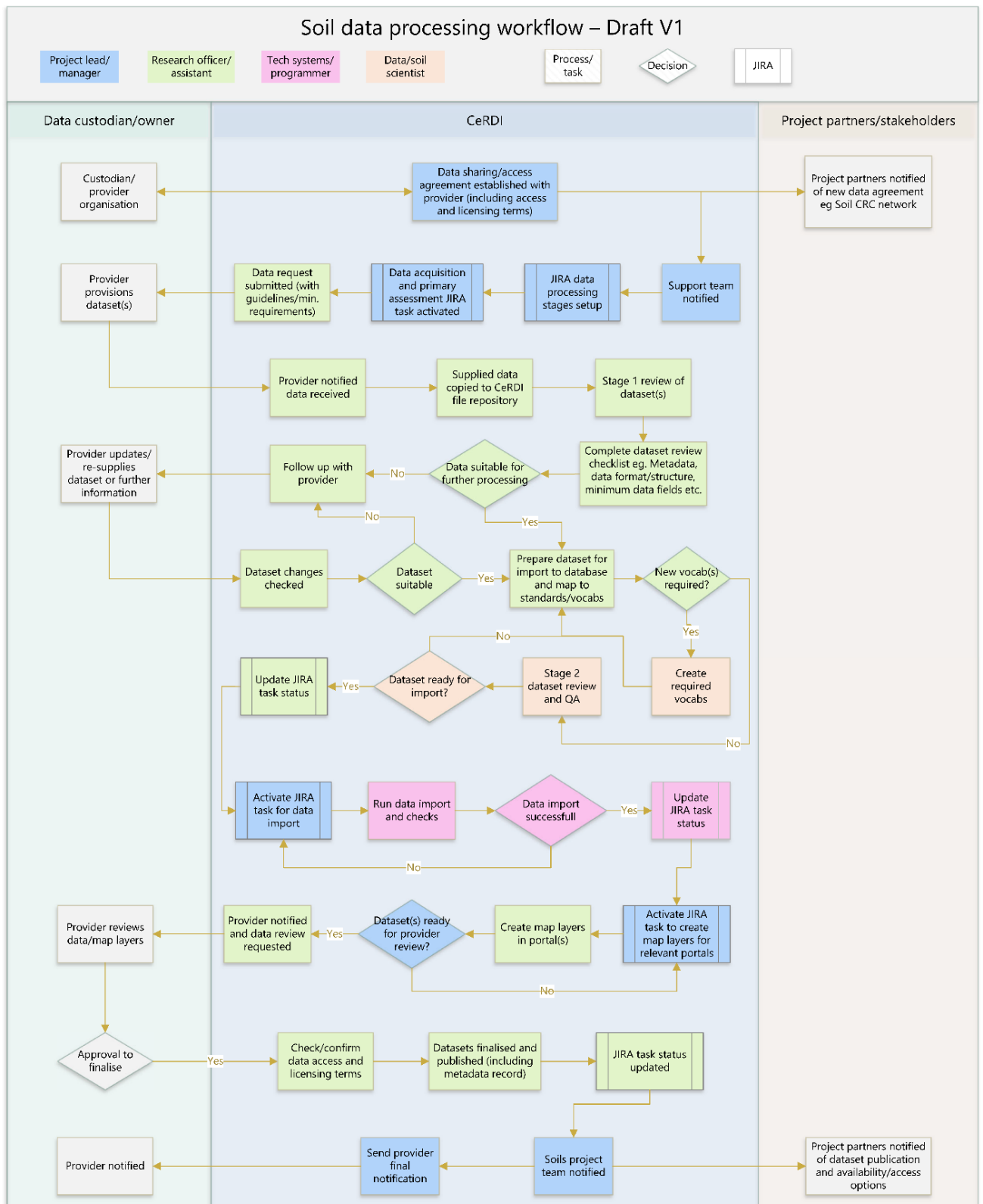


Figure 9. The data entry processing workflow for the VAS team.

### Data visualisation and output

Via the login, partners have access to basic tools to filter and graph their data. Filters can be used to select data by parameter, date and value (Figure 10). For sites that have test values recorded over different dates, the trends over time can be viewed. Similarly, for multiple samples taken at a single location, their values can be graphed by depth (Figure 10).

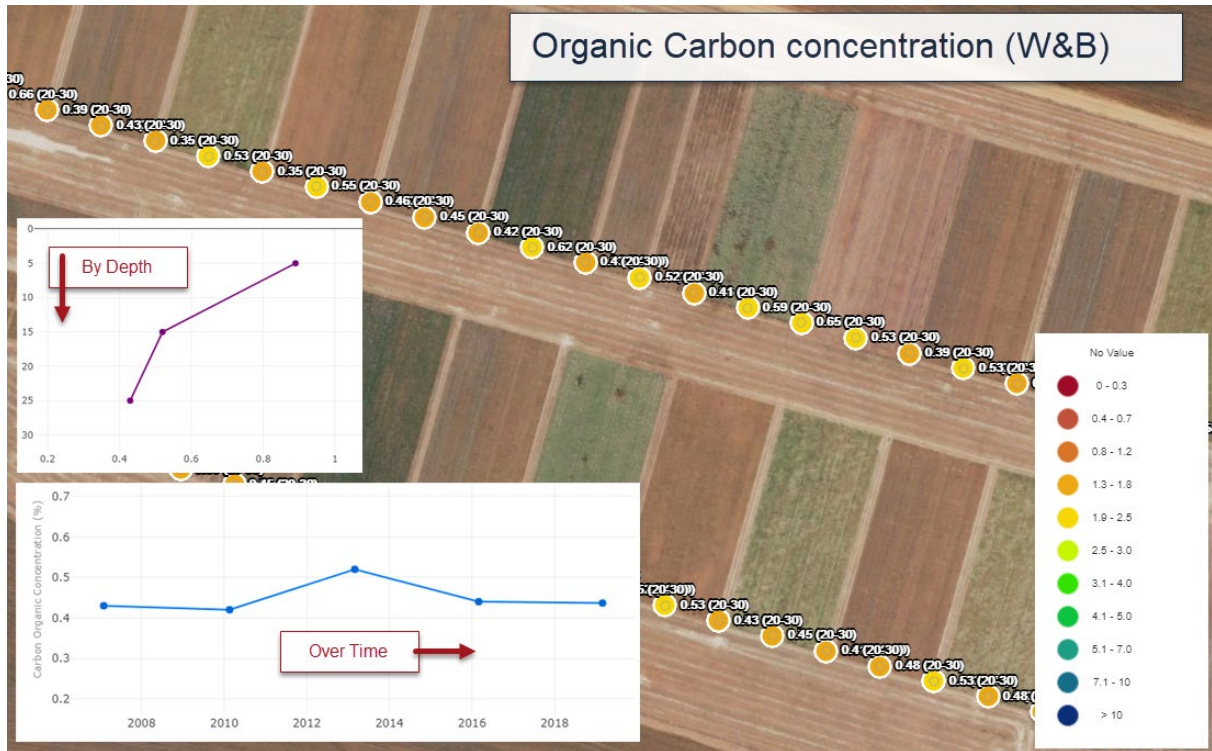


Figure 10. An example of VAS data that varies over time and by depth.

Ultimately, as these data have grown in number, the distribution of soil properties over regions can also be seen, filtered by property, time and custodian (Figure 11).

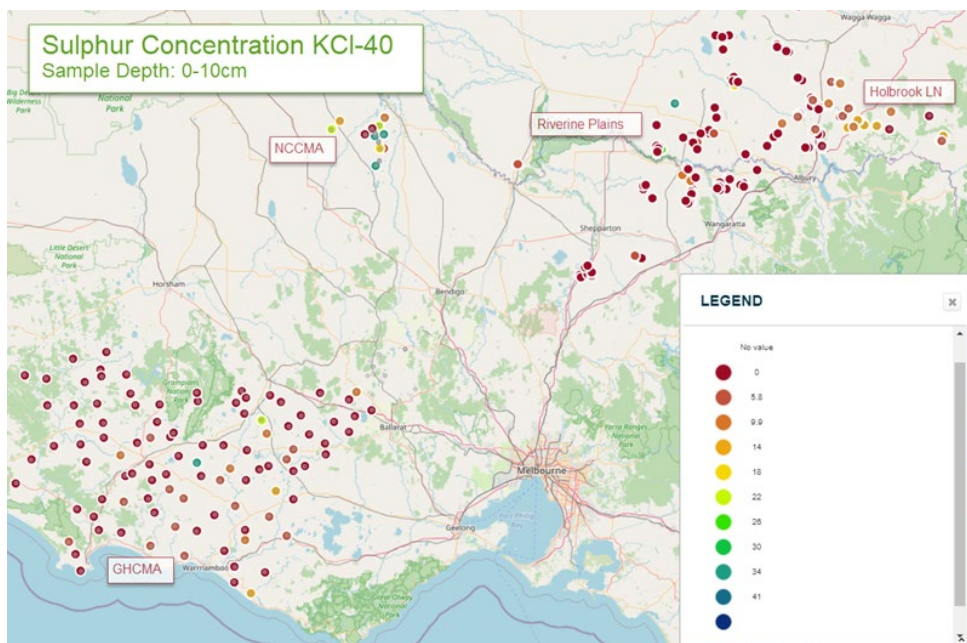


Figure 11. An example of a soil data parameter from four different custodians shown for a region.

### *Educational videos*

Educational videos have been developed to demonstrate key functionality and how to use the VAS system. The videos are screen recordings with accompanying audio instructions and have been edited and processed into a dedicated YouTube channel. There are currently seven videos, varying from 30 seconds to 2 minutes in length, each focussing on a different piece of functionality and how it can be accessed. Additional topics are under development and it is expected that common themes will emerge covering aspects of the self-service system and public mapping portal.

The current videos are:

1. Login and Access the My Soils Dashboard
2. Accessing and viewing individual soil test data
3. Charts: How to view soil property trends over time
4. Charts: How to View Soil Property by Depth
5. Charts: How to customise and save
6. Exporting Soil Test Data
7. Advanced Options Filtering and Styling.

## **DATA CONTRIBUTIONS**

The data in the VAS system is sourced from both the public sector and the private sector.

### **PUBLIC SECTOR (OPEN) DATA**

When the VAS system was initially constructed, open data from the public sector data repositories were used to build and test the portal functionality and data interoperability. These included Australian national-scale soil data from ASRIS, as well as other potentially useful data such as climate, land use and some satellite imagery or satellite-derived layers sourced from the Bureau of Meteorology (BoM), CSIRO, GeoScience Australia (GA), and the National Computational Infrastructure (NCI). An early addition also included four data sets from New Zealand – land cover, land use capability, national land cover and soil classification.

State soil data was then added on an *ad hoc* basis due to the variability and complexity of the data. This led to a systematic search of available soil data for Australia, which was undertaken using bespoke programming scripts<sup>11</sup> developed by Chris Bahlo for her PhD research. The searches were initially run during May and June of 2020 and then re-run during March 2021 to obtain a current set of results.

The search was conducted on 22 open data catalogues (Table 3). Where available (11 catalogues), the data search used API endpoints, the remainder were queried using online portals. Four portals allowed the export of search results as comma-separated value (\*.csv) files or other formats. Tabular file export was used where possible, and the remaining seven data catalogues were queried manually from within their respective online portals. Two catalogues were found to have no relevant datasets.

A previous search of public data catalogues showed that full text searches for a search term (i.e. 'soil') yielded too many results (>23,000), many of which were unrelated to the subject of the search. On the other hand, searches for keywords (e.g. 'agricultural soil data') were too restrictive, because catalogue search engines attempted a case-sensitive match on only

---

<sup>11</sup> <https://github.com/narrawin/datasearch>



exact, whole words. The most reasonable search yield was obtained by running a full text search and then filtering this using a case-insensitive, partial match on keywords. This method was used to obtain search results for the automated searches which directly queried the catalogue APIs (where available).

Table 3. The inventory of soil datasets available in Australian public catalogues.

Catalogue	URL	Metadata access method	Datasets found
Figshare	<a href="https://figshare.com/">https://figshare.com/</a>	Manual	0
Auscope	<a href="https://www.auscope.org.au/">https://www.auscope.org.au/</a>	Manual	0
Soil Sites SA	<a href="https://apps.environment.sa.gov.au/soils_forms/search">https://apps.environment.sa.gov.au/soils_forms/search</a>	Direct link	1
data.qld.gov.au	<a href="https://www.data.qld.gov.au/dataset">https://www.data.qld.gov.au/dataset</a>	CKAN	2
ACT data	<a href="https://www.data.act.gov.au/browse">https://www.data.act.gov.au/browse</a>	Manual	2
AgReFed	<a href="https://www.agrefed.org.au/ExploretheData">https://www.agrefed.org.au/ExploretheData</a>	Direct link	4
ASRIS	<a href="https://www.asris.csiro.au/">https://www.asris.csiro.au/</a>	Manual	9
Dryad	<a href="https://datadryad.org/search">https://datadryad.org/search</a>	Manual	13
SEED	<a href="https://datasets.seed.nsw.gov.au/dataset">https://datasets.seed.nsw.gov.au/dataset</a>	CKAN	27
Data.Vic	<a href="https://discover.data.vic.gov.au/">https://discover.data.vic.gov.au/</a>	CKAN	30
data.nsw.gov.au	<a href="https://data.nsw.gov.au/data/dataset">https://data.nsw.gov.au/data/dataset</a>	CKAN	30
Geoscience Australia	<a href="https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/home">https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/home</a>	CSV export	45
data.wa.gov.au	<a href="https://catalogue.data.wa.gov.au/">https://catalogue.data.wa.gov.au/</a>	CKAN	59
TAS TheList	<a href="https://www.thelist.tas.gov.au/app/content/data/index">https://www.thelist.tas.gov.au/app/content/data/index</a>	CSV export	70
CeRDI	<a href="http://data2.cerdi.edu.au/">http://data2.cerdi.edu.au/</a>	CKAN	97
data.sa.gov.au	<a href="https://data.sa.gov.au/data/dataset">https://data.sa.gov.au/data/dataset</a>	CKAN	101
TERN	<a href="https://portal.tern.org.au/">https://portal.tern.org.au/</a>	CSV export	108
data.nt.gov.au	<a href="https://data.nt.gov.au/">https://data.nt.gov.au/</a>	CSV export	161
CSIRO	<a href="https://data.csiro.au/collections">https://data.csiro.au/collections</a>	OpenAPI	215
Qspatial	<a href="http://qldspatial.information.qld.gov.au/catalogue/custom/search.page">http://qldspatial.information.qld.gov.au/catalogue/custom/search.page</a>	API	252
RDA	<a href="https://researchdata.edu.au/">https://researchdata.edu.au/</a>	getRIFCS API	516
data.gov.au	<a href="https://data.gov.au/search">https://data.gov.au/search</a>	Magda API	517
Total	22 catalogues		2259

Manual searches within portals were guided by search and filter options available. In some cases, keyword searches were unavailable, in which case the results set was refined using other options, such as subject area or data provider.

Not all desired metadata was available in each catalogue. Licencing or distribution details were not always provided or in a form that could be easily extracted and were simply recorded as 'unknown' in the analysis. Information about dates and spatial coverage was also not always provided. No attempt was made at this stage to obtain any missing metadata through manual effort via the respective portals since the value of doing so was unclear.

After metadata cleaning, the remaining 2103 data sets (from the 2259 listed in Table 3) were compiled in a spreadsheet. Sorting, filtering and counting functions were used to analyse the

result set. 839 datasets were listed in two or more catalogues (up to four). By this method, 475 duplicate datasets were removed, leaving **1,628 unique datasets**. A visual representation of the relationship between the data catalogues is shown in Figure 12.

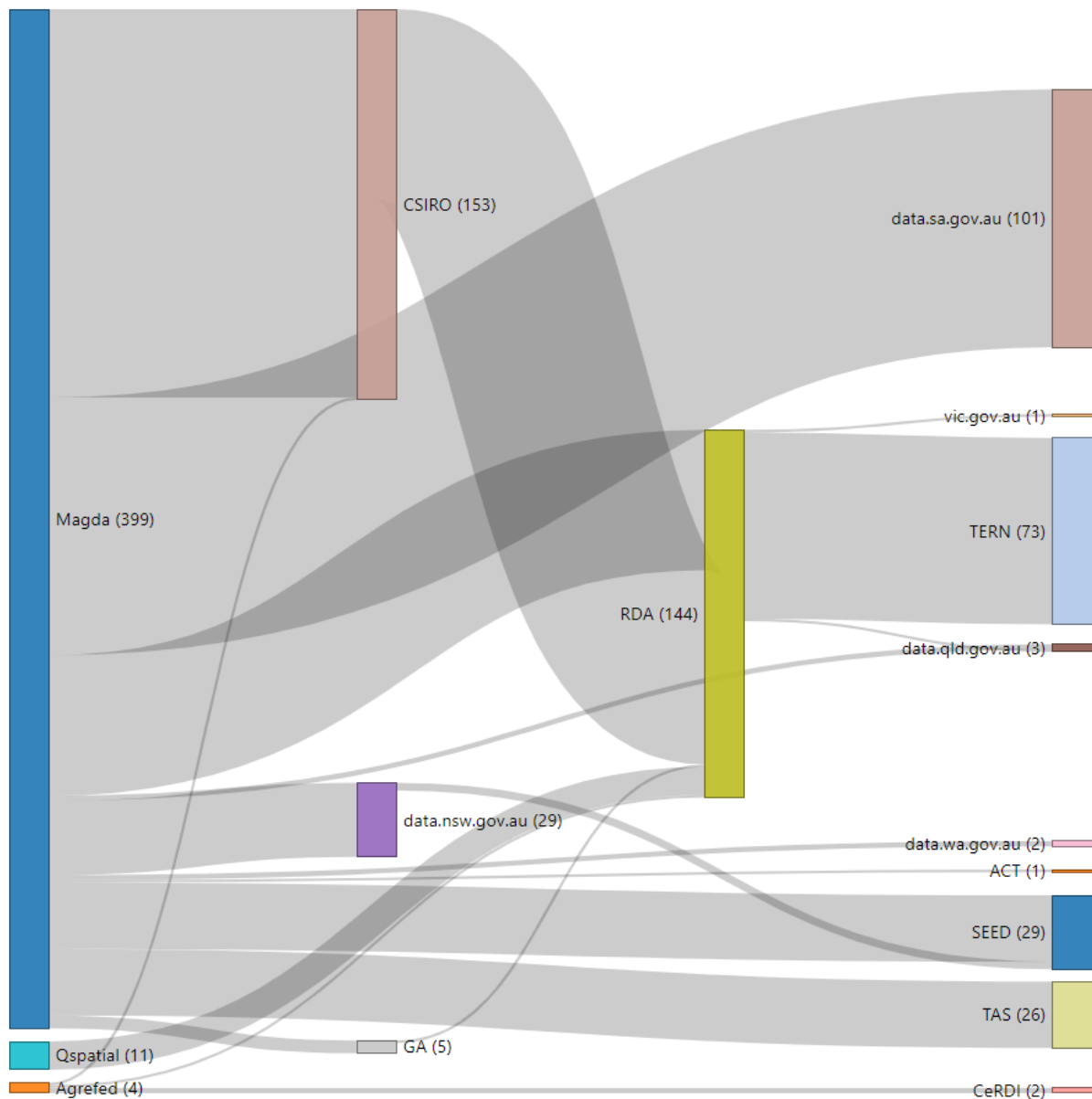


Figure 12. Duplications of soil datasets between catalogues

Note 1. A limitation of this graph is that the datasets without duplicates in other catalogues are not shown, and it does not make clear where datasets are common to more than two catalogues.

Note 2. Magda is an Australian federated open-source data catalogue designed as a set of microservices to pull data from many different sources into one easily searchable catalogue. <https://magda.io/>

An alternative representation is a Venn Diagram, shown in Figure 13, where the number of soil datasets is indicated by the size of the circle, and the amount of overlap indicates common datasets. However, it is impossible to accurately show the correct amount of overlap in a two-dimensional picture, therefore the overlap is approximate only.

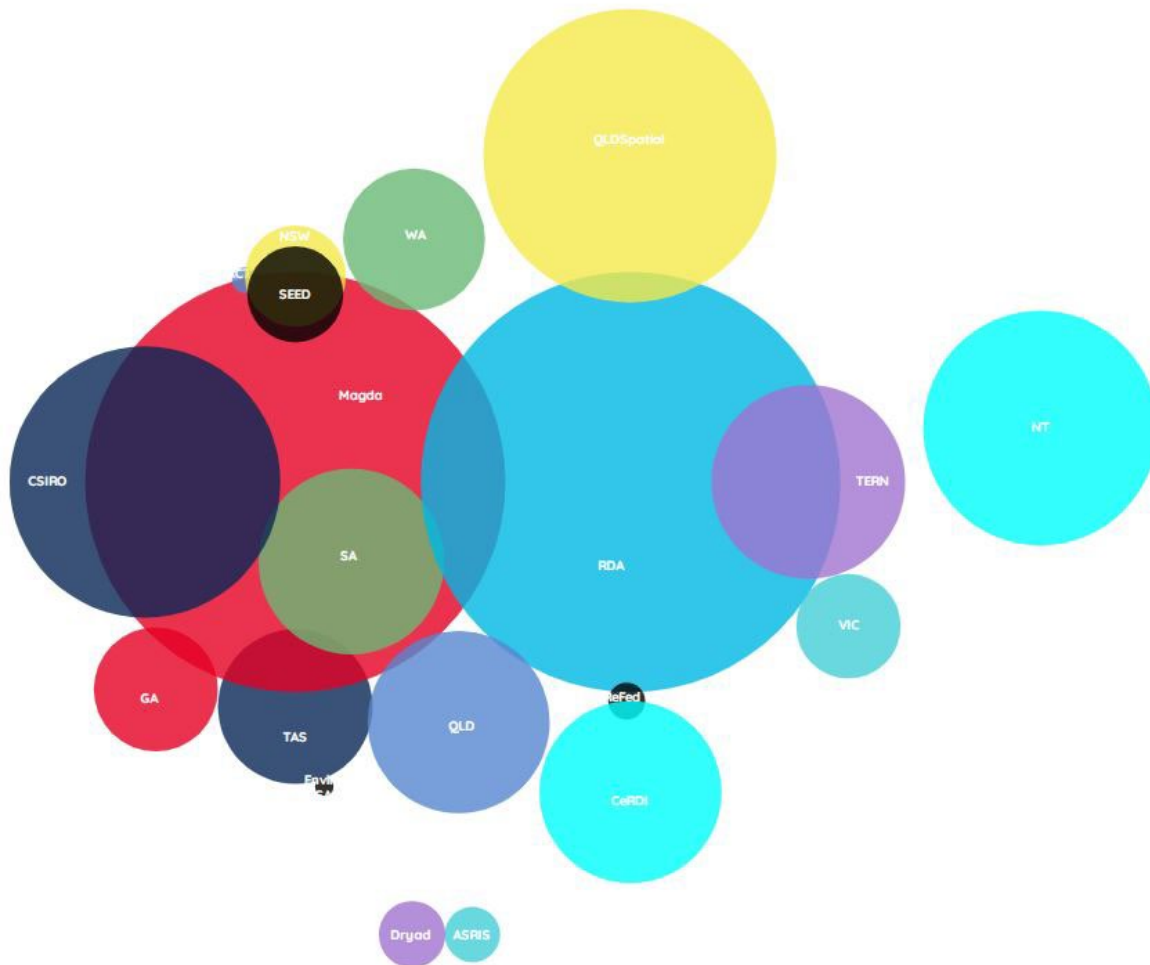


Figure 13. Venn Diagram representation of the overlap of soil datasets in public catalogues.

Due to the disparate data file formats, the exact number of soil data observations in the 1,628 datasets has not been determined. The metadata reports 73,000 observations of soil physical and chemical characteristics for one set in New South Wales; 179,490 observations (112,618 of which are laboratory analyses and the remainder are soil textures) for a South Australian dataset; and 96,000 for a Queensland dataset. Australian soil data from many of the public databases included in the above analysis was documented by Searle (2014), who reported approximately 2.5 million soil observations.

For several states, soil data is provided through information portals that are not easily queried or do not make data available in an easily downloaded useable form such as a spreadsheet. Victoria, New South Wales and Tasmania all have examples (Appendix A). The Victorian Resources Online (VRO) is an example of a site where a considerable quantity of high-quality soil data is locked into inaccessible formats<sup>12</sup>. Similar situations exist for the NSW eSpade portal and Tasmania's portal known as The List.

<sup>12</sup> meaning not easily downloaded into spreadsheets or structured data tables for input into other applications. See for example [http://vro.agriculture.vic.gov.au/dpi/vro/coranregm.nsf/pages/cor\\_soil\\_pits](http://vro.agriculture.vic.gov.au/dpi/vro/coranregm.nsf/pages/cor_soil_pits)

The datasets are available in a variety of file formats, with some being available in numerous file formats, selectable at download.

The count of file formats for each dataset is tabulated (Table 4) and the count for the various file formats is listed in Table 5 and illustrated in Figure 14.

Table 4. Number of datasets with the number of file formats available per set.

File formats per dataset	Count of datasets
0	949
1	268
2	53
3	149
4	24
5	21
6	108
7	52
8	3
9	1
<b>Total</b>	<b>1628</b>

Most of the file formats are commonly used. Although some are generated using proprietary software programs (generally the ESRI<sup>13</sup> suite of GIS products), they can be read by open-source tools, such as QGIS.

Table 5. List of file formats

File format	Count	Percentage of total
esri shape	355	22 %
pdf	217	13 %
html	210	13 %
wms	169	10 %
tif	150	9 %
zip	146	9 %
json	130	8 %
km	130	8 %
geojson	126	8 %
data	64	4 %
api	60	4 %
doc/docx	57	4 %
wfs	56	3 %
mpk	53	3 %
csv	45	3 %
xml	43	3 %
xls/xlsx	40	2 %
esri rest	37	2 %
vnd	19	1 %
geodatabase	16	1 %
access	3	<1 %
postgres	1	<<1 %
software	1	<<1 %

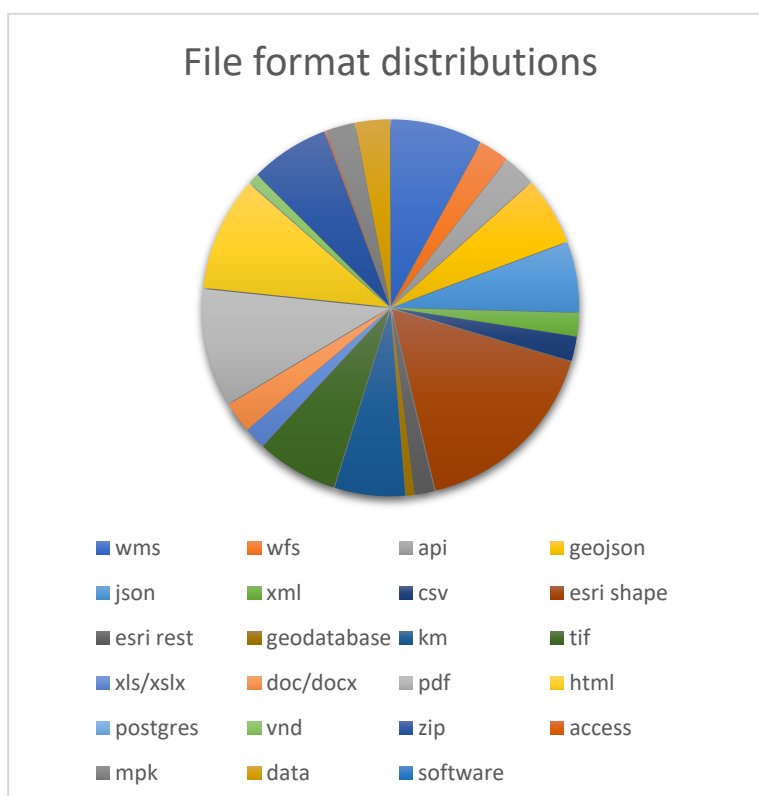


Figure 14. Graphical representation of the file formats

<sup>13</sup> ESRI is the Environmental Systems Research Institute, a well-known GIS company (ArcGIS).

Dataset licences in relation to data catalogues are shown in Figure 15, illustrating the high number of ‘unknown’ results<sup>14</sup>. This could be reduced if the metadata is available by manually accessing each dataset in each data catalogue portal to see if the information exists. Most that have included the details are listed as Creative Commons (CC)<sup>15</sup>.

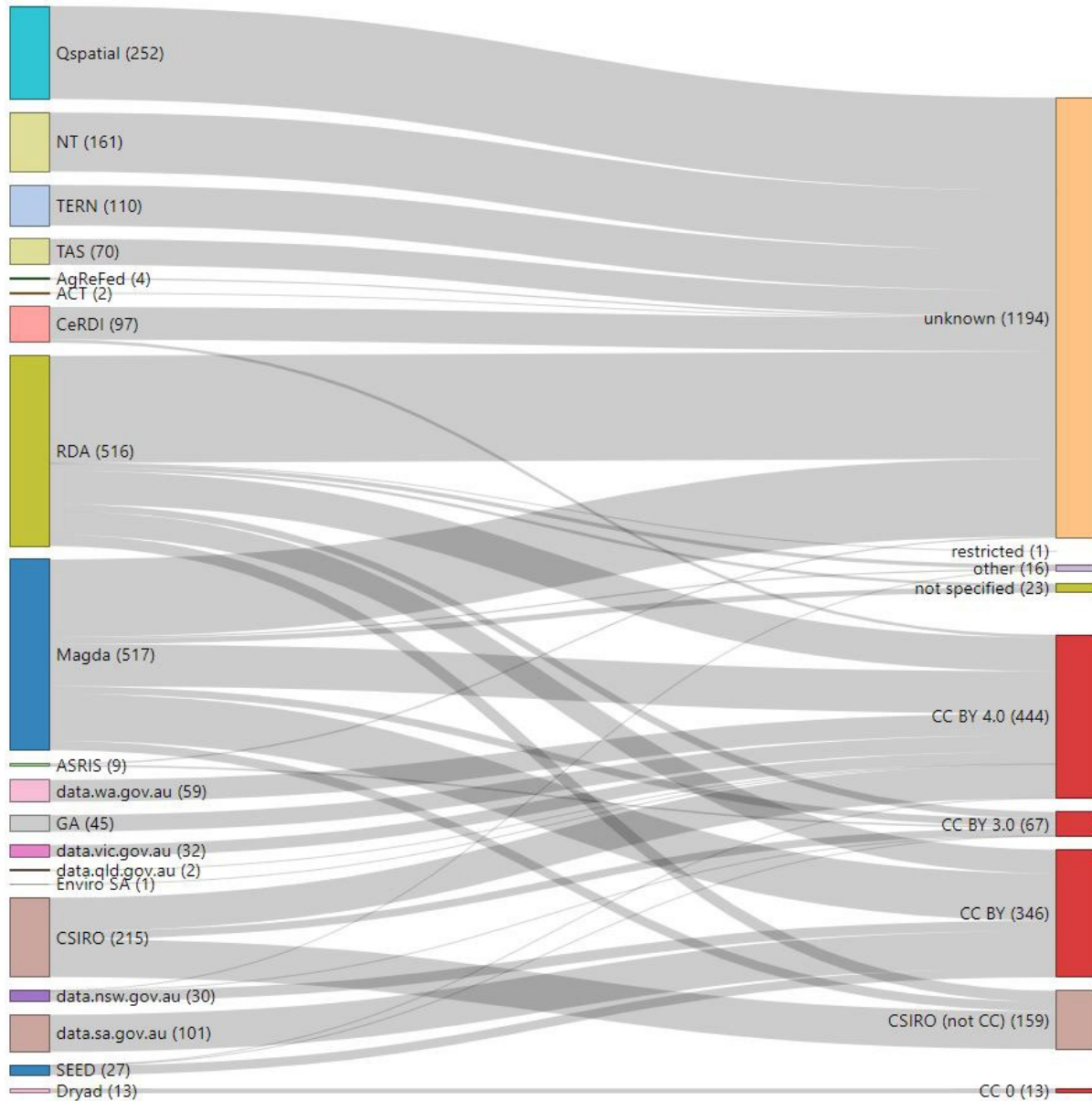


Figure 15. Licence details for public soil datasets.

While most of the state government departments have migrated considerable volumes of soil data into digital databases, there are many hundreds of thousands of legacy soil sites, samples and observations that remain in archives. In Victoria, archives of the Soil Conservation Authority, EPA Victoria (Audit Reports), Geological Survey of Victoria, State Electricity Commission, State Rivers and Water Supply Commission, Country Roads Board, Gas and Fuel Commission, and many water boards, drainage committees, and irrigation

<sup>14</sup> This is not uncommon with geospatial type APIs and soils data is not unusual in this respect.

<sup>15</sup> <https://creativecommons.org/about/cclicenses/>

boards, all have considerable volumes of good-quality public soil data that may never be made openly accessible. A similar situation exists in other states.

Attempting to spatially sort a definitive map of public soil data sites from the various datasets would be a massive undertaking. However, research undertaken by CSIRO (Searle 2014) for the Terrestrial Ecosystem Research Network (TERN), an NCRIS funded platform, assembled 281,202 soil data sampling sites across Australia (Figure 16).

Another method to display this is shown in Figure 17 which is a 'heatmap' of the number of soil sample sites in the TERN database in 2014. Since that time, the number of sites has increased by some thousands, but not tens of thousands (Ross Searle, *pers. comm.* 19 April 2021) and the distribution is largely unchanged. Both figures 16 and 17 illustrate the clustered density of soil sites around the nation's population centres.

A more recent analysis of the Australian public soil data was published by Searle et al. (2021) and includes a map (Figure 18) of the distribution of soil observations in relation to landscapes. Using this method, developed by Brendan Malone, landscapes in yellow (Figure 18) have a bigger range of observations across a particular landscape, compared to other landscapes. This map illustrates that sampled soil sites in the farming areas in the Murray-Darling Basin and the wheat belt of Western Australia have a bigger range of observations per location than may be initially interpreted from the previous maps. The analysis does not have a temporal component. In fact, around 99.5 % of the public soil data locations in Australia have only ever been visited once.

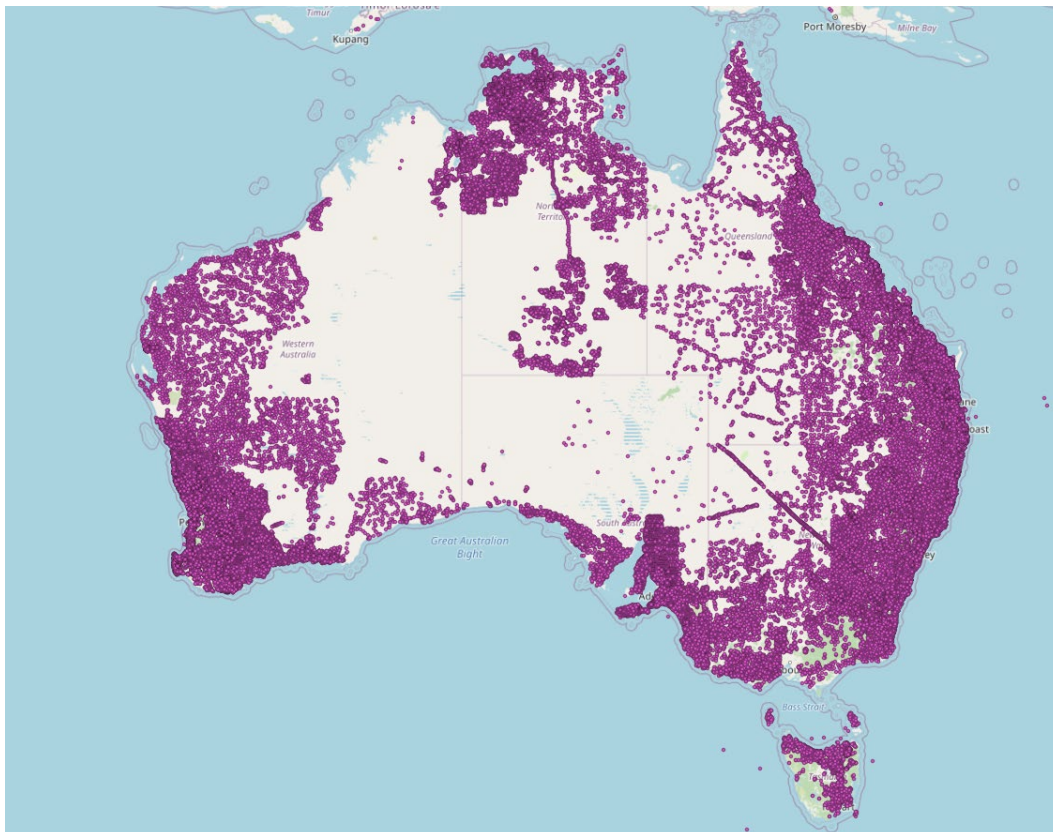


Figure 16. A spatial distribution of much of the public soil data for Australia (data source: Searle (2014) for the TERN database)

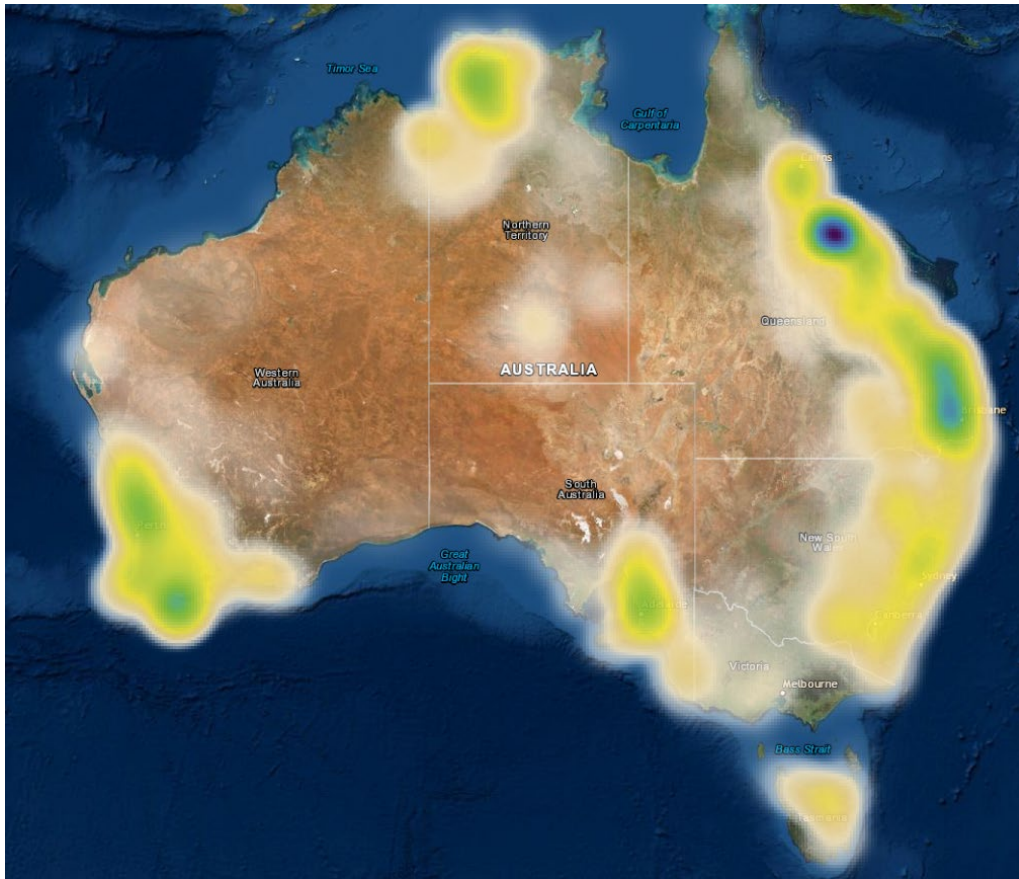


Figure 17. A heatmap of the number of sites in the TERN database in 2014.

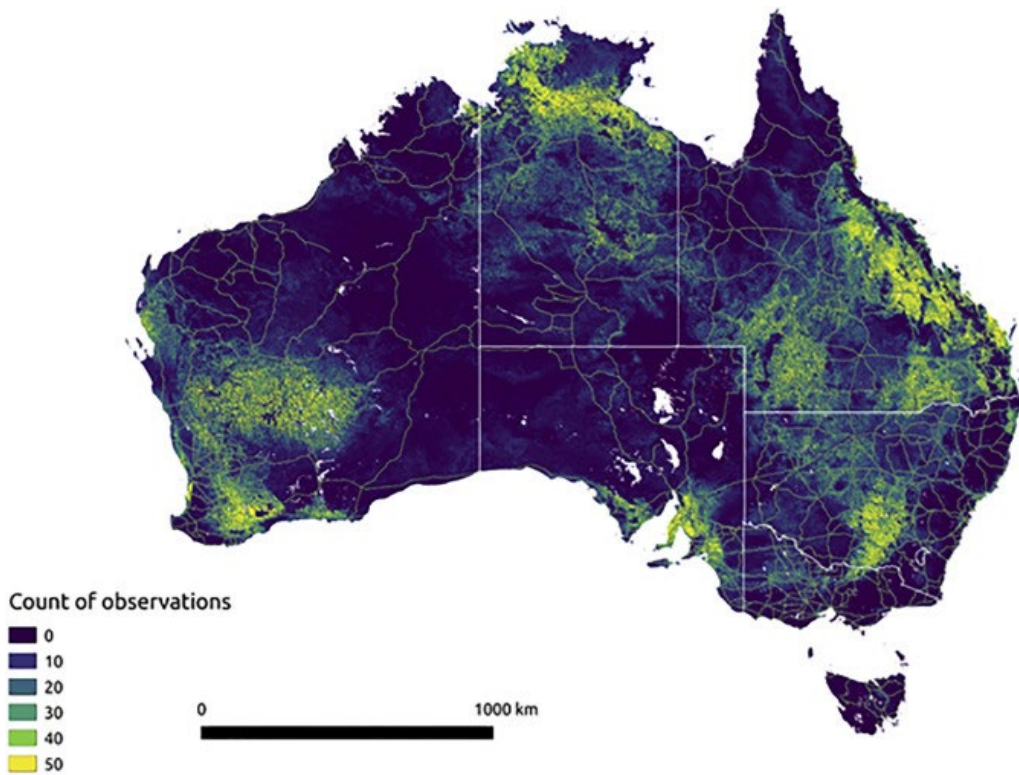


Figure 18. The spatial distribution of soil observations in relation to landscapes. It illustrates a comparative measure of landscapes based on their range of observations (sourced from Searle et al. (2021)).

## MAPS AND RASTERS

As shown in Table 5, at least 50 % of the public soil datasets are in spatial formats, such as mapped polygons and modelled grids.

### *Polygonised maps*

The majority of the older maps or traditional maps are polygons of soil series (Figure 19), land systems, soil capability or suitability, and soil constraints or risks (Figure 20).

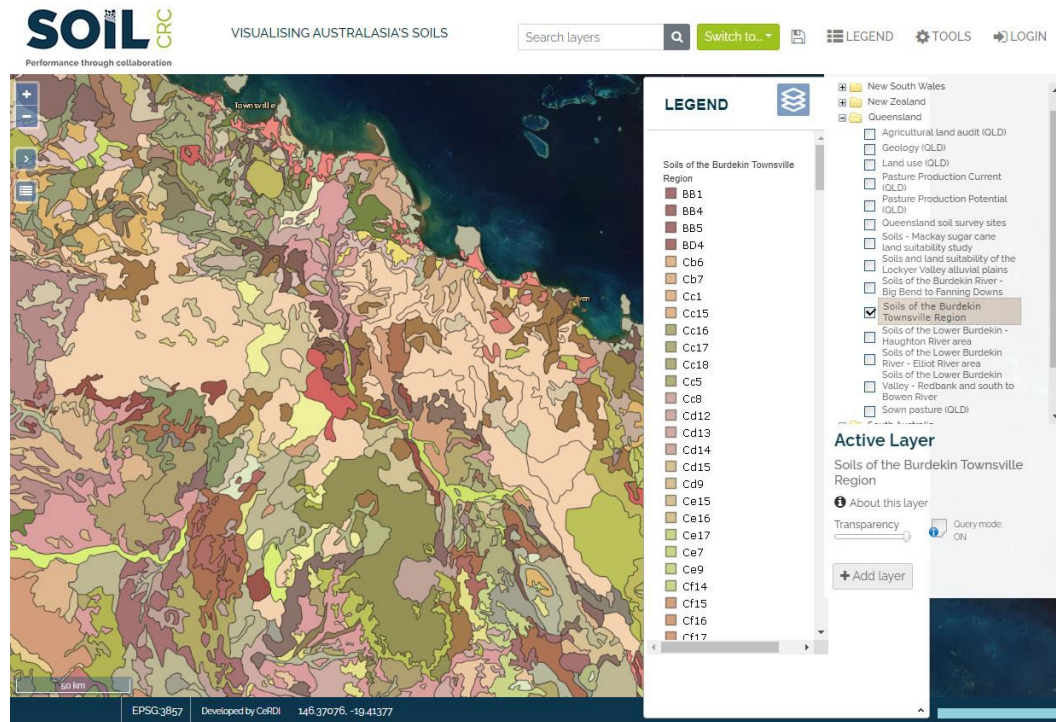


Figure 19. A soil series map for the Burdekin – Townsville region, Queensland.

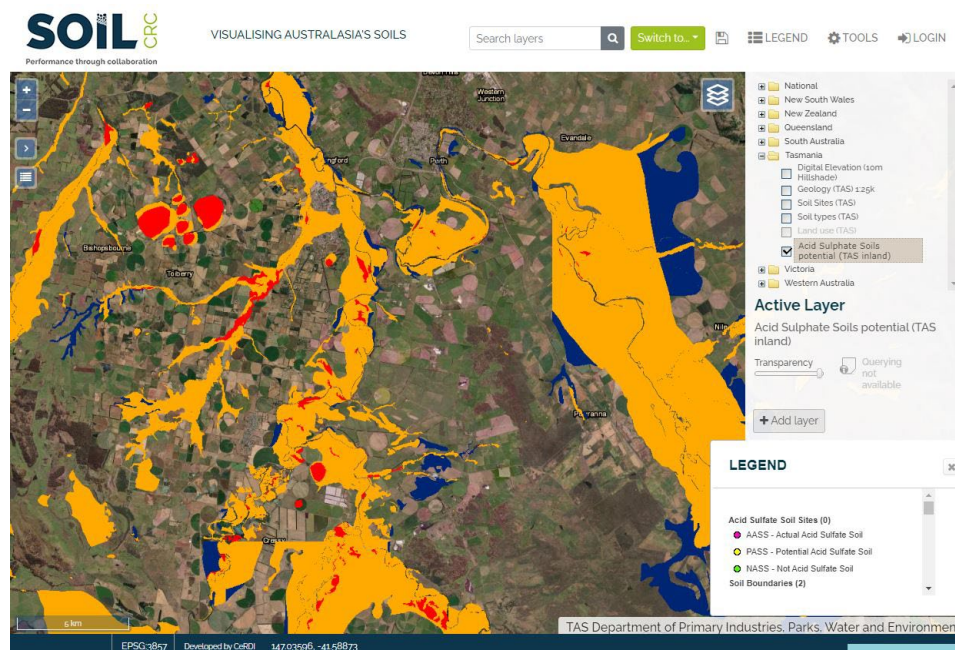


Figure 20. Map of acid sulfate soils potential for part of Tasmania.



Most polygonised maps, such as shown in Figure 19, obviously require linking to the descriptive resource to understand the encoded polygon identifiers. This is yet to happen, as linking all the required descriptive resources is a substantial task. Even where the map polygons are self-explanatory, such as the hazard map shown in Figure 20, they also require a link to the explanatory notes to appreciate the limits to the mapping and map usage.

### *Soil and Landscape Grid of Australia*

The Soil and Landscape Grid of Australia (SLGA)<sup>16</sup> is arguably the best known of the digital soil mapping products. It was originally completed in 2014 as a partnership led by CSIRO and included the University of Sydney, Geoscience Australia, most of the state and territory government agricultural departments and the Terrestrial Ecosystem Research Network (TERN).

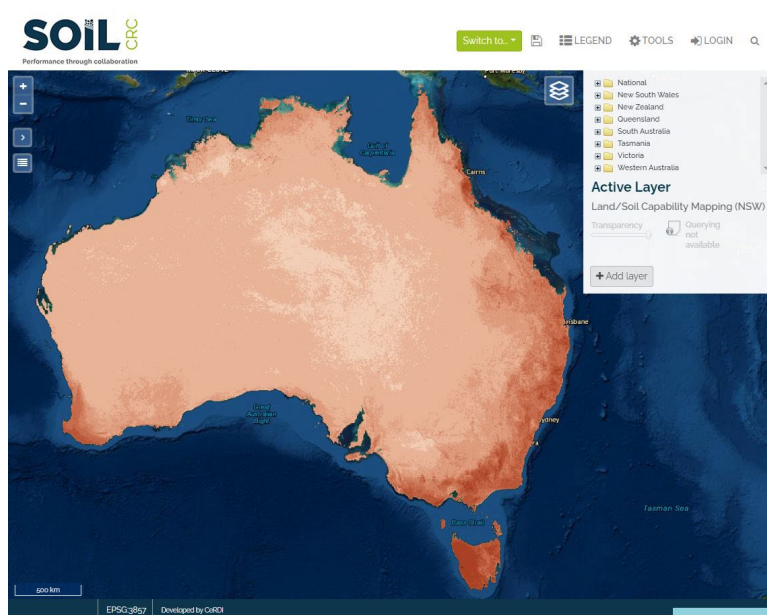
The SLGA was constructed using the available data mostly from public sector databases and was modelled, using cubist data mining methods, as a 3 arc-second (approximately 90 m x 90 m) grid of the entire Australian continent. Soil attributes provided include Bulk Density (Whole Earth), Organic Carbon, Clay, Silt, Sand, pH Soil Water, pH CaCl<sub>2</sub>, Available Water Capacity, Total Nitrogen, Total Phosphorus, Effective Cation Exchange Capacity, Depth of Regolith, Depth of Soil, and Coarse Fragments.

Landscape attributes provided include Slope (%), Slope Relief Classification, Aspect, Relief 1000 m Radius, Relief 300 m Radius, Topographic Wetness Index, Topographic Position Index, Partial Contributing Area, Multi-resolution Valley Bottom Flatness (MrVBF), Plan Curvature, Profile Curvature, Prescott Index, Solar Radiation (SRad) Net Radiation January, SRad Net Radiation July, SRad Total Shortwave Sloping Surface January, and SRad Total Shortwave Sloping Surface July.

The SLGA soil attributes are modelled for six depths (0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm and 100-200 cm) to comply with the Global Soil Map requirements. The values are provided as 'Expected Value', and 5th percentile confidence limit and 95th percentile confidence limit. The provision of these data in international standard data exchange format makes them visible in any interoperable data portals (Figure 21).

All products are freely available and are released under a Creative Commons Attribution Licence (CC BY).

Figure 21. Modelled organic carbon (5 – 15cm) interoperably displayed from the Soil and Landscape Grid of Australia data portal.



<sup>16</sup> <https://www.clw.csiro.au/aclep/soilandlandscapegrid/>

### Other digital soils models

Other models worthy of mention are those produced by the various states such as the Victorian and New South Wales models for soil parameters (Figure 22). In general, these are of finer resolution (20 m x 20 m, or 30 m x 30 m) and have higher accuracy than the SLGA. While confidence limits will have been generated by the digital soil mapping methods used, they are not easily accessible.

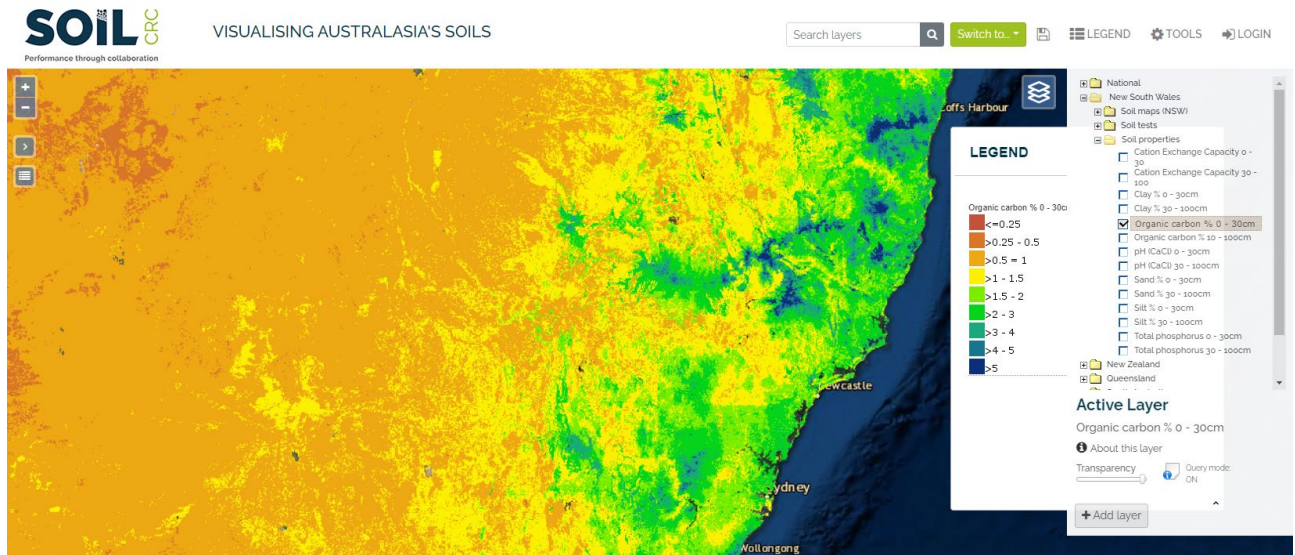


Figure 22. Soil organic carbon (% at 0 – 30 cm depth) modelled for New South Wales.

There are many other digital soil products from Australia that have been, and are being, produced by academic researchers. One of these included in the VAS is PhD work by Robinson (2016) which produced fine-scale digital soil mapping for western Victoria which included pH and clay mineralogy type.

Until recently, New Zealand has not created digital soil map layers. The exception is the soil carbon stock for the New Zealand mainland for 0-30 cm dataset (2012)<sup>17</sup>. Recently, MWLR has commenced work on the creation of a series of digital soil maps and covariate layers starting with a modelled soil pH layer (0-220 cm). The opportunity exists to add these to VAS in phase 2.

<sup>17</sup> <https://www.pggrc.co.nz/files/1500851629564.pdf>

## VAS PROJECT DATA

After visiting the groups, most have supplied some example data that has been loaded into the aggregator. The general nationwide spread of data is shown in Figure 23.

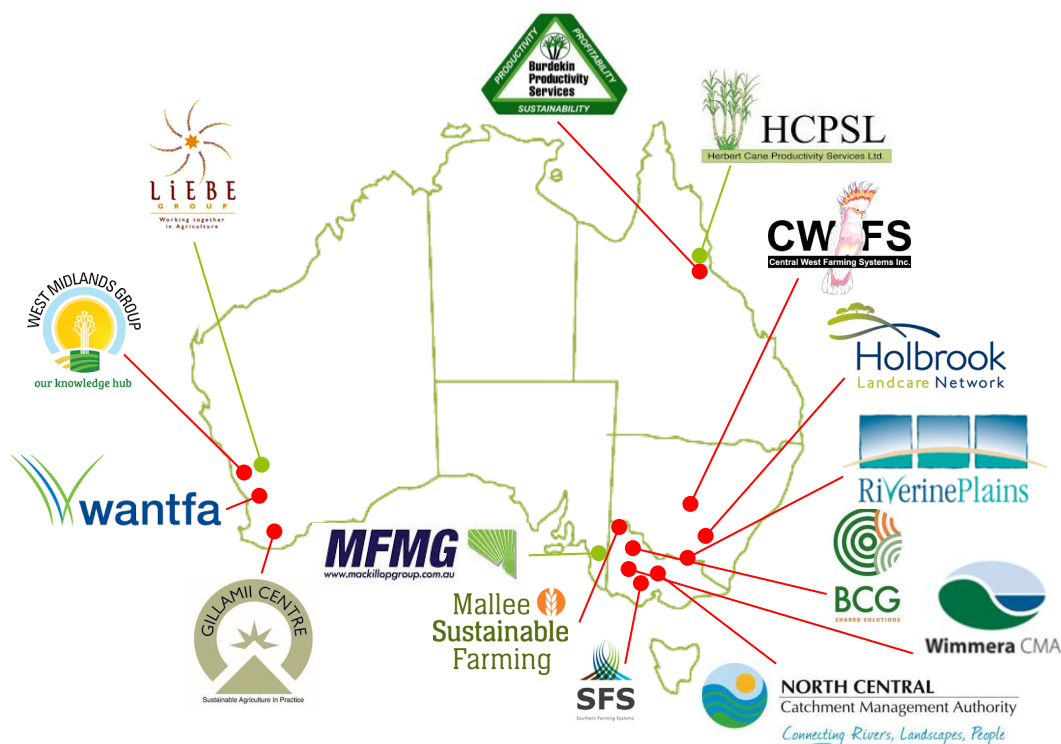


Figure 23. The geographic spread of the Australian participants in the VAS project.

(Note: those marked with a red line have loaded some data to the federation portal)

Data sharing by the groups is entirely at their discretion. While very little data has been shared between or beyond the groups to date, the current tally is shown in Table 6.

Table 6. Current tally of soil data uploaded by the VAS participants.

Group	Number			Data Status
	Sites	Samples	Observations	
1	37	114	1852	Loaded
2	78	132	3672	Loaded
3	140	217	5715	Loaded
4	31	277	3574	Loaded
5	128	1592	28827	Loaded
6	48	264	3626	Loaded
7	15	240	4035	Loaded
8	200	700	13968	Loaded
9	133	523	4355	Loaded
10	2	165	2708	In process
11	15	~160	~5000	In process
Totals	827	~4384	~77332	

The counts of observed properties are listed in Table 7.

Table 7. Counts of observed properties.

Observed property	Count	Observed property	Count
Aluminium concentration	471	Magnesium concentration	90
Ammonium (NH <sub>4</sub> ) concentration	2585	Manganese concentration	1098
Ammonium and nitrate as N concentration	42	Mineral soils field texture grade	1812
Anaerobic respiration	345	Mineralisation	45
Biological mass	315	Molybdenum concentration	75
Boron concentration	1046	Nitrate N concentration	2667
Calcium concentration	112	Nitrification	30
Calcium to magnesium ratio	881	Nitrogen concentration	1554
Carbon - organic concentration	4548	Nitrogen fixation	15
Carbon to nitrogen ratio	60	Organic nitrogen concentration	15
Cation exchange capacity	124	pH	6650
Chloride concentration	984	Phosphorus buffering index	547
Clay content	212	Phosphorus concentration	2915
Cobalt concentration	60	Phosphorus retention index	196
Colour	1908	Plant available water	15
Composition	390	Potassium concentration	2134
Concentration of organic carbon as humus or decomposed materials < 0.053mm	300	Potassium to magnesium ratio	12
Concentration of particulate organic carbon	300	Potassium to sodium ratio	22
Concentration of recalcitrant organic carbon	300	Sand content	212
Copper concentration	1076	Selenium concentration	15
Depth	400	Silicon concentration	171
Effective cation exchange capacity	623	Silt content	212
Electrical conductivity of soil	3553	Sodium concentration	97
Exchange acidity	12	Soil bulk density	1494
Exchangeable aluminium	2681	Soil organic matter concentration	267
Exchangeable calcium	2693	Soil texture	42
Exchangeable cation ratio	12	Soil water content	288
Exchangeable hydrogen	163	Soluble salt concentration	132
Exchangeable magnesium	2693	Specific gravity	15
Exchangeable potassium	2994	Sulphur concentration	2362
Exchangeable sodium	3222	The lower depth boundary	2983
Field texture	352	The upper depth boundary	2983
Grass tetany risk index	179	Zinc concentration	1154
Iron concentration	1280	<b>Total</b>	<b>70456</b>

From the above table, the most popular soil tests are soil acidity/alkalinity (pH), organic carbon, salinity (electrical conductivity), exchangeable ions, phosphorus, nitrate, ammonium, sulphur and potassium.

This does not represent the total number of soil tests held by the farmer groups and catchment management authorities involved in the VAS project. The participants have many tens of thousands of legacy soil tests usually stored in hardcopy reports and maps, or in a variety of digital data file formats on legacy storage devices. While possible to move those legacy data into the VAS federation, for some groups it remains doubtful because of the substantial effort required and the unknown or uncertain value of the data.

## DISCUSSION

The research lessons from this first phase of the VAS project are discussed in this section. They provide a much clearer understanding of the research challenges and set the direction for the next phase of the project.

### THE ELUSIVE VALUE PROPOSITION

In the social engagement undertaken while visiting the participating farmer groups and catchment managers it was apparent that access to a trusted, supported, web-based spatial soil data management system for their purposes was highly desired. Therefore, the value proposition for use by the farmer groups and their members is reasonably clear, especially if the data management system includes an intuitive toolkit of functions that makes their data relevant to their geographic patch and supports their stakeholder communities. In that respect, one of the most requested improvements to the VAS system is to be able to include the contextual data alongside the soil data, such as paddock histories, treatments, yields, climate records, etc. so that decision support tools can be auto populated with their latest data.

However, for data custodians to share their data with users outside of their trusted networks requires an obvious and worthwhile value proposition that remains elusive at this stage. While considerable effort has been made through the VAS project to explore what that may entail (Sexton 2020), it will need to be demonstrated through a highly attractive application before custodians would be convinced of the value of data sharing. For that reason, this is included as one of the key components in the second phase of the VAS project, commencing on 1 July 2021.

The intention in the next phase is to encourage soil data visibility and sharing through at least one large and significant project, co-designed and co-developed by the project participants, that delivers on one or more Soil CRC Milestones. The intention is to demonstrate the value of harmonising digital soil data to create aggregated de-identified views, in a novel and pragmatic way. For example, to benchmark areas, show trends over time, propose fit-for-purpose soil performance indicators, or identify significant research gaps. The aim is to encourage collaborative new Soil CRC projects based on the evidence drawn from data analytics supplied by the VAS.

### THE ISSUES OF DATA LITERACY AND DATA HARMONISATION

Perhaps the most obvious lesson to emerge from the first phase of the VAS project is the generally poor level of data literacy among soil data custodians. In the public sector, the data literacy skills are quite varied, from world-leading FAIR data custodians (e.g. CSIRO, TERN, AgReFed) to custodians with little knowledge of acceptable global practice (e.g. some state government agencies and universities). Among the private sector VAS participants, some have adequate data management practices (especially those who have adopted data management Apps) but others in the project do not.

As a generalisation, data management and data stewardship practices are often immature and *ad hoc*, with data spread across a variety of file formats and repositories, even within a single organisation. In many cases legacy data is no longer discoverable or readable, because of the inability to access legacy data formats (e.g. floppy disks, legacy databases, lost passwords, failed storage devices, or failed computing systems). Most participants recognise the unfortunate reality that high-quality soil data, usually collected in a publicly funded project

by a trained soil scientist, agricultural scientist or agronomist and processed in a laboratory accredited by the National Association of Testing Authorities (NATA), may be lost, forgotten or ignored because of inadequate data management systems in the office, and due to staff turnover.

Another general observation is that both public sector and private sector data custodians could significantly improve their metadata. Since metadata is key to making data FAIR, the current metadata quality is a significant barrier to data access and reuse. The unforeseen laborious task (shown in Figure 9) of mapping the datasets contributed by the project participants to the data schema of the VAS soil data aggregator revealed that almost all the participants' soil data required clarification to make it FAIR. The required metadata for data ownership, licensing, project details, test dates, and test methods, were the most lacking.

However, for those data that have been mapped into the system, they now conform to international data exchange standards and are interoperably available (subject to the data custodian's consent) for any future purpose, such as serving data (machine-to-machine) into other tools and applications, artificial intelligence engines, or decision support systems.

Data literacy is an issue that will need to be addressed if the VAS is to endure as a long-lived soil data federation for Australasia.

### *Data harmonization*

Arguably, the most unexpected lesson from the VAS is that soil data are more incongruent than we expected. For example, there are 32 different test methods for soil phosphorus recognised in Australia (Rayment and Lyons 2010) with terms such as extractable, total, available, index, saturation, and ratio. Descriptive measures include %, mg P/kg, µg P/kg, X/C, PBI, colour, and common assays are Colwell, Olsen and DGT (Robinson 2021). The selection of soil tests methods is tailored to suit their purpose and landscape factors ('horses-for-courses'), such as farm management decisions, analysing crop or pasture trials, or soil health monitoring.

While all these variations are recorded in the VAS (via the 'data curation' process, refer to Appendix D for details), to model these data still requires harmonisation of the incongruous data so that the comparisons of values are valid. This may be possible using pedotransfer functions, or algorithms to compare varied methods of chemical analyses, but it often requires additional information. To harmonise say, organic carbon measured in per cent with that measured in tonnes per hectare also requires the soil bulk density to be known<sup>18</sup>.

When the data is extended to include the complementary data (treatments, yields, etc.) the harmonisation is even more challenging since there are few data schemas and standard vocabularies available to map these complementary and supplementary data to.

Hence, to make full use of the substantial volume of data available, high-quality metadata and credible harmonisation functions will be essential.

## **GROWING THE DATA SOURCES**

A barrier to data sharing in the first phase of the VAS project is the assumption that the data only flows one way, that is, from the farmer group participants to the Soil CRC researchers. For data sharing to be encouraged, and to show trust, the soil data should also flow from the researchers to the project participants. Therefore, encouraging universities and research agencies to serve data to the soil data federation is considered equally important in the project.

---

<sup>18</sup> E.g. <https://www.agric.wa.gov.au/soil-carbon/measuring-and-reporting-soil-organic-carbon>

In that respect, phase one of the project failed, apart from a few datasets from Federation University being included.

Attempts to persuade the other participating universities to contribute soil data were stymied by the perceived barriers of intellectual property (IP) and data curation, even though the identified research datasets are already publicly available through the Australian Digital Theses repositories or published as supplementary data associated with journal papers or appendices in research reports. The universities and research agencies can, usually through their information librarian experts, make a reasonable volume of legacy soil data findable and accessible<sup>19</sup>. We acknowledge that it is more challenging to make these data interoperable and reusable, but unless the flow of data is two-way, the farmer groups and catchment managers can justifiably challenge the reasons to share their data.

From the point of view of the end-users, the most common request from the contributing soil data custodians is to be able to include contextual data with their soil data. Since almost all of their soil data is captured for particular projects, be it agricultural trials, soil health monitoring or land management, it is accompanied by complementary data such as soil treatments, yields, weather observations, biomass cuts, and supplementary data like reports, maps, videos, field day fliers, etc. These data provide context to research interpretations and in some cases may provide rich data sources for new research projects, such as calculating soil carbon flux balances. Logically, these complementary and supplementary data be considered as a research asset in addition to soil data.

Third party data could also be used to grow the breadth and depth of research data. Outside of the Soil CRC membership, there are other farmer groups, government agencies, businesses, and community groups who have expressed their desire to join the soil data federation. All have data to share and in all cases are involved in other projects with Federation University and therefore use similar technology platforms. These additional data are listed in Table 8 and all sites are in Victoria, apart from the company data that spans the south-eastern states of Australia.

*Table 8. Other known data that could be potentially included in the soil data federation.*

Data custodian	Sites	Samples	Observations	Time range
Landcare group 1	13 locations	37	480	2007
Landcare group 2	9 locations	27	350	2009
Farmer group 1	50 farms, 414 paddocks	845	13120	1994 to 2017
Farmer group 2	24 farms, 73 paddocks	119	1200	1995 to 2020
Individual farmer	1 farm, 29 paddocks	169	2551	1993 to 2018
Company	~200,000 locations	229,854	~700,000	2013 to 2021
Totals	~250,000	231,051	~720,000	1993 to 2021

It is likely that other Soil CRC research members similarly share data with their research collaborators. Consideration should be given to enabling these data to add to the VAS federation. It will require agreement with the governance and stewardship guidelines, but the rewards for Australasian soil research would be significant.

<sup>19</sup> For example in the Research Data Australia metadata catalogue <https://researchdata.edu.au/> or FigShare <https://figshare.com/>

## TOOLSETS

Regarding the VAS portal, a major challenge still to be conquered is how best to show the massive volume of open public soil data that exists for Australasia, in a way that makes sense to the end-user. In the current public view, there are a limited number of soil data sets visible, since it is not clear whether adding all the possible (1,628+) datasets would be of value.

For many of the current spatial layers, the mapped features are cryptically described or subject to misinterpretation without access to the explanatory notes, even for a soil scientist. One option being considered is to show the available datasets for a user-selected polygon or map frame and allowing the user to select from a list to add to their map portal view. It does not overcome the obscurity but does provide the user with the knowledge of what soil information exists for a chosen region (i.e. makes all data findable). In the next phase of VAS a variety of solutions will be tested with the project participants and Soil CRC researchers to reach a consensus on the best method.

A disappointing outcome of the first phase of the VAS project is that the data self-serve system does not yet have the functionality that we had expected. In particular, the development of the fine-grained access controls was technically limited by the inadequate quality of the metadata and the heterogeneity of the datasets. Nevertheless, the lesson from this research is that with improved data literacy, the goal of seamless self-serve data management and access control can still be achieved.

Early in the next phase of the VAS, the intention is to co-design and extend the educational tools to build data literacy within the participants and end-users using practical examples of the advantages and pitfalls in data use. For example, clarifying how to easily assign data ownership, licencing and metadata, and understand the data obligations of contracts. It is thought that these lessons could be best developed as short videos (preferably humorous), podcasts, and web-based tutorials. Simultaneously, the technical team will work to increase the robustness of the self-serve data input system and the data export and reporting tools, aiming for low user effort, seamless integration, intuitive-to-use, and easy functionality for everyday uses.

In the long-term, the intention is to extend the data visibility by encouraging the soil data federation members to provide wider access to their data. This will require a significant research effort in building a data sharing framework that has sufficient granularity but remains efficient, easy to use and administer, and has the trust of the custodians.

Apart from the data management components, the development of the basic functionality to view data, login, use data filters, graph observations, save data views and so on, has worked well and provided the project participants with a taste of what might be possible. As these functions will increase confidence in the value of the project, it is a relatively important component to expand upon.

Therefore, in the next phase, in collaboration with participants willing to pilot new initiatives, we will introduce existing tools (e.g. calculators, decision support, etc.) that value-add to their data. These tools may be already used by the participants (outside of the VAS system)<sup>20</sup>, or 'off-the-shelf' tools developed by others and open sourced. In addition, functionality to have automatic data entry from sensor systems (e.g. soil moisture probes and satellite systems) can be developed. Soil data from legacy projects could be added by text and data mining from past theses and research projects.

---

<sup>20</sup> For example, Excel spreadsheet functions or macros, online calculators or lookup charts.



# CONCLUSION

The VAS research project aimed to establish a soil research data federation 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. Phase one of the VAS project has successfully met the original aims even though they are not all to the level anticipated at the start of the project.

Although there are at least 1600 open public soil datasets, many are limited for use by their inadequate metadata and FAIRness, making it difficult to judge their value to soils research. A remaining challenge is to show more of the public soils data in a relevant way and encourage the public data custodians to improve their data stewardship.

The collaboration with project participants to provision their data to the spatial knowledge system according to the rules that they set has been reasonably successful. At the end of phase one, 12 of the 16 participants have provisioned 827 soil sites at which around 4384 samples have been taken with about 77,332 soil observations. This process has uncovered new challenges in data literacy and data harmonisation that will be a focus for the next phase of research. Data sharing outside of a custodian's group has not yet happened. The barriers to sharing will also be explored, but part of the solution will be building trust by encouraging the researchers to share their data with the farmer groups and catchment managers.

The co-development of data stewardship governance frameworks for sharing data has been deferred for the time being, until the mechanics and functionality of the VAS system are fully tried, and the subtleties of data stewardship and implications of the suggested governance structures are better understood. Although the milestone was met by providing a discussion paper, the aim has not yet been achieved.

Delivering training programs with project participants to improve data stewardship and governance, to improve the collection and quality of data and metadata for inclusion in the knowledge system was an aim that was, to some extent, thwarted by the global pandemic in 2020-21. But apart from the restrictions on face-to-face meetings, a major delay was due to the slow pace of data supply and the unexpectedly lengthy process of data curation required to load those data (once they were received) into the cloud-based data aggregator. Some 'how-to' educational videos have been delivered to assist the project partners with the use of the VAS system, but the aim of co-designing and delivering online educational materials for farmers and researchers to make best use of the data in the knowledge system was not achieved.

The aim of developing simple web-based tools to assist in spatially visualising soil data, searching and filtering data, downloading data sets and publishing data to the portal, was achieved. These tools have been trialled by project participants and their feedback will be used to enhance and grow the toolset in the next phase of the project.

The final project aim, to co-develop, with CRC end-users, dynamic models that are applied to the interoperably federated data to answer frequently asked questions such as finding temporal trends in soil performance indicators, was considered ambitious from the outset. Meeting this aim depended on the provisioning of temporal datasets with full metadata, which rarely occurred. The aim has been partially achieved through data provided by a couple of data custodians (one with up to 13 years of data) demonstrating the potential.

# RECOMMENDATIONS

From the lessons of the first phase of the VAS research, the following recommendations are intended to guide the second phase of the project.

## VALUE PROPOSITION

The following recommendations will be implemented to improve the value proposition for the data custodians to use the VAS portal and share their data more broadly:

1. In collaboration with each project participant, co-design and implement a couple of simple use-cases for their data that can demonstrate a tangible benefit of their participation in the VAS project.
2. Encourage soil data visibility and sharing through at least one large and significant project, co-designed and co-developed by the project participants, that delivers on one or more Soil CRC Milestones. The intention is to demonstrate the value of harmonising digital soil data to create aggregated deidentified views, in a novel and pragmatic way, for example to benchmark areas, show trends over time, propose fit-for-purpose soil performance indicators, or identify significant research gaps. The aim is to encourage collaborative new projects based on the evidence drawn from data analytics supplied by the VAS.
3. Extend the social research to measure the impact of the portal (practice change, for example) and inform future portal developments, and through linking with the Soil CRC Program 1 project being led by Dr Hanabeth Luke, et al.
4. Finalise the portal governance and business model, and data governance and stewardship components. The aim is to ensure that the future of the VAS data federation is supported by the members and enduring.

## DATA LITERACY AND HANDLING

1. Co-design and extend the educational tools to build data literacy within the participants and end-users using practical examples of the advantages and pitfalls of data use. In particular, clarifying how to easily assign data ownership, licencing and metadata, and understand the data obligations of contracts. These could be best developed as short videos (preferably humorous), podcasts, and web-based tutorials.
2. Extend the data visibility by encouraging the soil data federation members to provide wider access to their data. This will require a significant research effort in building a data sharing framework that has sufficient granularity but remains efficient, easy to use and administer, and has the trust of the custodians.

## DATA INCLUSION

1. Increase the breadth and depth of the soil data across the federation, especially by including data provisioned by the researchers and research organisations (Soil CRC partner universities and government research agencies), as well as all 21 of the farming group partners in the Soil CRC.
2. Start including the data collected by Soil CRC projects by linking with the Soil CRC Program 2 project being led by Dr Nathan Robinson, et al.

3. Include data feeds from other agricultural data federations, such as the Agricultural Research Federation, TERN Landscapes project, National Soil Information Framework, National Soil Monitoring Program, National Soil Carbon Monitoring Program, Australian AgriFood Data Exchange, etc. using semantic web technologies, especially where the farming groups, project participants and researchers can find value.
4. Expand on the legacy data collection, sensor data feeds, and trial consuming data feeds from farm Apps and tools commonly used by the project participants, and serving data feeds to the farm Apps and tools.
5. Test different ways to include all available data, including that from third party participants from the private sector and/or community sector.

## TOOLSETS

1. In collaboration with participants willing to pilot new initiatives, introduce existing tools (e.g. calculators, decision support, etc.) that value-add to their data, develop automatic data entry from sensor systems (e.g. soil moisture probes and satellite systems), and soil data from legacy projects (e.g. text and data mining from past theses and research projects).
2. Increase the robustness of the self-serve data input system and the data export and reporting tools, aiming for low user effort, seamless integration, intuitive-to-use, easy functionality for everyday uses.
3. Expand the visualisation capability of the data to encourage discoveries in the data for both soil practitioners and researchers.

# ACKNOWLEDGEMENTS

The authors acknowledge the guidance provided by the project Steering Committee members and thank them for their time and effort in reviewing reports, attending meetings and making time for the *ad hoc* discussions about the project. It would not be a success without their valuable input.

We thank all the project participants, especially those who personally met with us to discuss their involvement in the VAS project and those who worked with our team to load their data into the system, and then tested the login and functionality of the portal.

Soil CRC administrative team are wonderfully supportive and efficient in resolving issues and answering our many queries. Jodi, Julie and Katherine are singled out for their unwavering and superb support to the VAS project team.

The collaboration with Peter Wilson, Linda Gregory, Simon Cox, Paul Box, Ross Searle and their colleagues at CSIRO has been critical to the success of the project to date, especially in their support and discussions related to the technical, information, and social architectures.

Cam Nicholson (Nicon Rural Services), Ben Fleay and Kirsten Barlow (Precision Agriculture), Rob Shea (Perennial Pasture Systems), Casandra Schefe (AgriSci P/L), Steve Williams (AgVic), John Friend (NSW DPI), Chris Pitfield (Corangamite CMA), Troy Clarkson (DAWE), Anna Whitton and George Ellis (DISER), David Lamb and Birgita Hansen (Food Agility CRC) have all contributed their ideas to the VAS project. Finally, the CeRDI PhD students: Basharat Ali, Rekha Attanayake, Rob Clarke and Peter Weir are thanked for their inputs and discussions that also helped shape the VAS concept.

## REFERENCES

- AgGateway. (2021). "AgGateway." (web page) Retrieved 25 February 2021, from <https://www.aggateway.org/Home.aspx>
- AgTrix. (2021). "AgTrix " (web page) Retrieved 25 February 2021, from <https://www.agtrix.com/>
- Antle, J.M., Basso, B., Conant, R.T., Godfray, H.C.J., Jones, J.W., Herrero, M., Howitt, R.E., Keating, B.A., Munoz-Carpena, R., Rosenzweig, C., Titttonell, P. and Wheeler, T.R. (2017a). Towards a new generation of agricultural system data, models and knowledge products: Design and improvement. *Agricultural Systems* 155: 255-268. DOI: <https://doi.org/10.1016/j.agsy.2016.10.002>.
- Antle, J.M., Jones, J.W. and Rosenzweig, C. (2017b). Next generation agricultural system models and knowledge products: Synthesis and strategy. *Agricultural Systems* 155: 179-185. DOI: <http://dx.doi.org/10.1016/j.agsy.2017.05.006>.
- Back Paddock Company. (2021). "Back Paddock Company. Unlocking the power of agronomy using science and technology to increase the dollars." (web page) Retrieved 25 February 2021, from <https://www.backpaddock.com.au/>.
- Box, P., Simons, B., Cox, S. and Maguire, S. (2015) A Data Specification Framework for the Foundation Spatial Data Framework. CSIRO, Sydney, Australia. 72pp.
- Brodaric, B., Boisvert, E., Chery, L., Dahlhaus, P., Grellet, S., Knoch, A., Létourneau, F., Lucido, J., Simons, B. and Wagner, B. (2018). Enabling global exchange of groundwater data: GroundWaterML2 (GWML2). *Hydrogeology Journal* 26/3: 733-741. DOI: 10.1007/s10040-018-1747-9.
- Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. *PLoS Comput Biol* 17(6): e1009041. <https://doi.org/10.1371/journal.pcbi.1009041>
- Dahlhaus, P., Thompson, H. and MacLeod, A. (2017). Towards data democracy in digital agriculture, Australian Society of Agronomy. 'Doing more with less', Proceedings of the 18th Australian Agronomy Conference, 24 – 28 September 2017, Ballarat, Australia: 4.
- 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. Milestone 4 report. Cooperative Research Centre for High Performance Soils (Soil CRC), Callaghan, NSW. 22p.
- DAWE (2021). Budget 2021-22. National Soil Strategy, Australian Government Department of Agriculture, Water and the Environment. . Factsheet 2.
- Downes, R.G. (1949) A soil, land-use, and erosion survey of parts of the Counties of Moira and Delatite, Victoria. CSIRO Bulletin. 243. Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia. 66p.
- ESIP. (2021). "ESIP. Soil Ontologies and Informatics." (web page) Retrieved 20/06/2021, from [https://wiki.esipfed.org/Soil\\_Ontologies\\_and\\_Informatics](https://wiki.esipfed.org/Soil_Ontologies_and_Informatics)
- FAO. (2018). "Global Soil Partnership." (web page) Retrieved 7 May 2018, from <http://www.fao.org/global-soil-partnership/en/>
- FAO. (2021a) Farm data management, sharing and services for agriculture development. . Food and Agriculture Organisation of the United Nations Rome, Italy. 162p.

- FAO. (2021b). "Global Soil Partnership. Areas of Work. Soil information and data. Welcome to GLOSIS | Global Soil Information System. ." (web page) Retrieved 18/06/2021, from <http://www.fao.org/global-soil-partnership/areas-of-work/soil-information-and-data/en/>
- FarmersEdge. (2021). "FarmersEdge." (web page) Retrieved 21/05/2021, from <https://www.farmersedge.ca/>
- FarmIQ. (2021). "Farm IQ. Bring it all together. Farm management software designed to drive sustainable, productive and profitable outcomes by bringing all your farm information into one place. ." (web page) Retrieved 18/06/2021, from <https://farmiq.co.nz/>
- Farmlab. (2021). "Farmlab. Astronomy and project management software to streamline your business." (web page) Retrieved 25 February 2021, from <https://www.farmlab.com.au/>
- GO FAIR. (2021). "FAIR Principles." (web page) Retrieved 4/1/2021, from <https://www.go-fair.org/fair-principles/>.
- GODAN. (2018). "Soil Data " (web page) Retrieved 7 May 2018, from <http://www.godan.info/working-groups/soil-data>.
- Gudivada VN., Rao DL. and Gudivada AR. (2018) Information Retrieval: Concepts, Models, and Systems. Chapter 11 in Handbook of Statistics (Editors: Venkat N. Gudivada, C.R. Rao), Volume 38, Pages 331-401, Elsevier <https://doi.org/10.1016/bs.host.2018.07.009>.
- Holmes, L., Leeper, G.W. and Nicolls, K. (1939). Survey of the Country Around Berwick. *Proceedings of the Royal Society of Victoria* 52/1: 177:238.
- ISRIC. (2021a). "ISRIC World Soil Information." (web page) Retrieved 25 February 2021, from <https://www.isric.org/>.
- ISRIC. (2021b). "Our Approach for Generating Open Soil Data." (web page) Retrieved 21/05/2021, from <https://www.isric.org/explore/soil-information-provisioning>
- IUSS. (2020). "IUSS Working Groups." (web page) Retrieved 25 February 2021, from <https://www.iuss.org/organisation-people/organisation/working-groups/>.
- Leeper, G.W., Nicholls, A. and Wadham, S.M. (1936). Soil and pasture studies in the Mount Gellibrand area, Western District of Victoria. *Proceedings of the Royal Society of Victoria* 49 77-138.
- Lilburne, L.R., Hewitt, A.E. and Webb, T.W. (2012). Soil and informatics science combine to develop S-map: A new generation soil information system for New Zealand. *Geoderma* 170: 232-238. DOI: <https://doi.org/10.1016/j.geoderma.2011.11.012>.
- Mahmoud, Q.H. (2005). Service-Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application Integration (EAI). USA Oracle Corporation Technical Article.
- Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B. and Wilkinson, M.D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use* 37: 49-56. DOI: 10.3233/ISU-170824.
- MW-LR. (2021a). "National Soils Database. National Soils Data Repository (NSDR)." (web page) Retrieved 25 February 2021, from <https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database/the-nsdr-improvement-programme/>
- MW-LR. (2021b). "Soil Data and Maps. National Soils Database." (web page) Retrieved 25 February 2021, from <https://soils.landcareresearch.co.nz/soil-data/national-soils-data-repository-and-the-national-soils-database>
- MyEnviro. (2021). "MyEnviro. Connecting you to your environment through data. ." (web page) Retrieved 18/06/2021, from <https://myenviro.co.nz/>

NTT. (2021) Design of a National Soil Information Framework. Final Report. Consulting report for the Department of Agriculture, Water and the Environment 9 April 2021 v 0.01. NTT Australia Digital Pty Ltd Canberra p.

OGC. (2017). "Agriculture DWG." (web page) Retrieved 30 January 2017, from <http://www.opengeospatial.org/projects/groups/agriculturedwg>.

Proagrica. (2021). "agX® The language of agriculture." (web page) Retrieved 25 February 2021, from <https://proagrica.com/products/agx/>

RDA. (2018). "Agricultural Data Interest Group (IGAD)." (web page) Retrieved 7 May 2018, from <https://www.rd-alliance.org/groups/agriculture-data-interest-group-igad.html>

Rezare. (2021). "DataLinker and the Farm Data Standards." (web page) Retrieved 25 February 2021, from <https://www.datalinker.org/>

Robinson, N. (2016). Assessing productive soil-landscapes in Victoria using Digital Soil Mapping. PhD, Federation University Australia.

Robinson, N. (2021) GRDC Online Farm Trials. Soil data. Presentation slide deck (unpubl.). Federation University Australia Mt Helen, Ballarat. 23p.

Searle, R. (2014). The Australian site data collation to support the *GlobalSoilMap*. *GlobalSoilMap: Basis of the global spatial soil information system* D. Arrouays, N.J. McKenzie, J.W. Hempel, A.C. Richer de Forges and A.B. McBratney. London, UK, CRC Press, Taylor & Francis Group: 127-133.

Searle, R., McBratney, A., Grundy, M., Kidd, D., Malone, B., Arrouays, D., Stockman, U., Zund, P., Wilson, P., Wilford, J., Van Gool, D., Triantafyllis, J., Thomas, M., Stower, L., Slater, B., Robinson, N., Ringrose-Voase, A., Padarian, J., Payne, J., Orton, T., Odgers, N., O'Brien, L., Minasny, B., Bennett, J.M., Liddicoat, C., Jones, E., Holmes, K., Harms, B., Gray, J., Bui, E. and Andrews, K. (2021). Digital soil mapping and assessment for Australia and beyond: A propitious future. *Geoderma Regional* 24: e00359. DOI: <https://doi.org/10.1016/j.geodrs.2021.e00359>.

Sexton, A. (2020) Visualising Australasia's Soils: Social Engagement and Collaboration Learnings. June 2020. Federation University Australia Cooperative Research Centre for High Performance Soils (Soil CRC). 24p.

Simons, B., Wilson, P., Ritchie, A. and Cox, S. (2013). ANZSoilML: An Australian - New Zealand standard for exchange of soil data: EGU2013-6802.

Skene, J.K.M. (1963) Soils and Land Use in the Deakin Irrigation Area, Victoria. Soil Survey Technical Bulletin 16. Department of Agriculture, Victoria, p.

Soil Tech Project. (2021). "From Soil Science to Soil Management." (web page), from <https://www.soiltechproject.org/>

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E. and Wyborn, L. (2019). Make all scientific data FAIR. (Comment). *Nature* 570: 27-29. DOI: 10.1038/d41586-019-01720-7.

Trimble. (2021). "Farmer Pro." (web page) Retrieved 21/05/2021, from <https://agriculture.trimble.com/product/farmer-pro/>

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooff, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR

Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018.  
DOI: 10.1038/sdata.2016.18.



# APPENDIX A REVIEW OF STATE SOIL REPOSITORIES

**Note: this review has been contributed by Andrew MacLeod**

This Appendix provides an overview of Australian State and Territory foundational soil data. The focus is on soil site surveys and analysis rather than soil series maps or other mapping products.

## QUEENSLAND

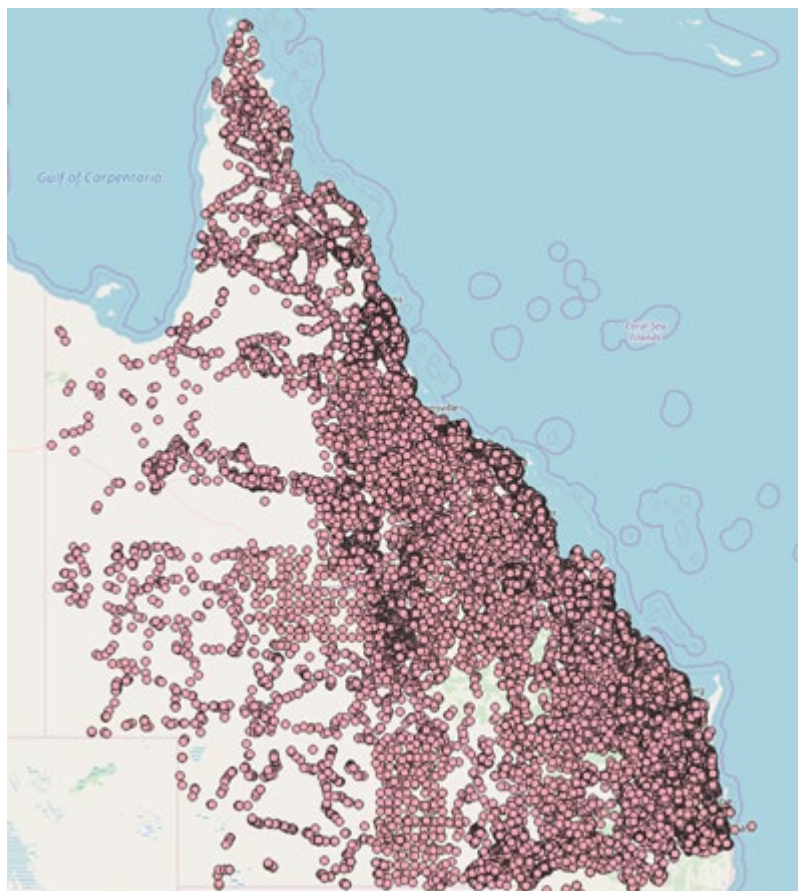


Figure A-1. Distribution of public soil data sites in Queensland

Queensland Globe / [data.qld.gov.au](http://data.qld.gov.au) allows full download of SALI soil sites (via QSpatial) via:

- [data.qld.gov.au/dataset/queensland-soil-survey-sites-sali\\_site](http://data.qld.gov.au/dataset/queensland-soil-survey-sites-sali_site) (metadata)
- <http://qldspatial.information.qld.gov.au/catalogue/custom/detail.page?fid={0B7EB39F-73C8-4079-9CEE-27C43A11821C}> (download)

Download is via spatial datamart-type interface. You request the dataset in the format you would like, and a link is emailed to you. Federation University email guard has a problem with the links and an email address from outside the university was required.

The dataset contains **93,355** records (sites).

The data is mainly location and summary info with a link to a site report in \*.pdf format. A sample record is tabulated below (Table A-1).

While there are lots of other data and information in the site report \*.pdf files, most notably soil profiles, it was not possible to determine if this data is available in a more interoperable way. This requires an enquiry to the data custodian.

Table A-1. Queensland Government sample public soil data record

OBJECTID	59217
PROJECT_CODE	MISSQ
SITE_ID	46
OBSERVATION_NUMBER	1
FIRST_OBSERVED	18/02/2004
LATITUDE	-25.10062
LONGITUDE	147.79443
LOCATION_MEASUREMENT_METHOD	Averaging GPS
LANDFORM	Hillslope
SOIL_NAME	
AUSTRALIAN_SOILS_CLASSIFICATION	Tenosol
LABORATORY_DATA_AVAILABLE	No
SITE_REPORT_URL	resources.information.qld.gov.au/soils/reports/sites?project=MISQ&site=46

Separately, a **soil chemistry API** is available via:

- [data.qld.gov.au/dataset/sali-soil-chemistry-api](https://data.qld.gov.au/dataset/sali-soil-chemistry-api) and
- [soil-chem.information.qld.gov.au](https://soil-chem.information.qld.gov.au)

It appears that this was the outcome of a Queensland Government GovHack event.

The API is well documented and functionally rich via ODATA specification. It allows fairly complex filtering and sorting and returns JSON:

- <https://soil-chem.information.qld.gov.au/swagger/>

This was integrated into the VAS demo map portal:

```
var sali_api_url = "https://soil-chem.information.qld.gov.au/odata/SiteLabMethodResults";
var filter = "SiteId eq " + siteId + " and ProjectCode eq " + projectCode + " and ObservationNumber eq " + obsNumber + """;
var expand = "SiteLabMethod,Sample($select=UpperDepth,LowerDepth)";
var api_url_request = encodeURIComponent(sali_api_url + "?$filter=" + filter + "&$expand=" + expand);
```

## SOUTH AUSTRALIA



Figure A-2. Distribution of public soil data sites in South Australia

Soil Sites SA has its own website:

- <https://data.environment.sa.gov.au/Land/Data-Systems/Soil-Sites-SA/>

There is a link on the page above to download all publicly available Soil Site SA data:

- <https://apps.environment.sa.gov.au/soils.zip>

This consists of a package of nine \*.csv files

- Sites
- Structures
- Horizons
- Samples
- Observations
- segregations
- land\_uses
- mottles
- course\_frgs

There are a total of **19,019** sites with coordinates. A sample record is tabulated below.

Table A-2. South Australian Government sample public soil data record.

site id	153721
id	026-19_2110
obs id	89856
id_1	1
obs type	Undisturbed soil core
nature	
described by	MASD
date described	04/JUL/1978
date confidence	Best available. Data has been derived using recognised techniques by a qualified soil science or land resource assessment professional.
latitude	-35.00257052
longitude	138.9597931
location method	Hard-copy 1 to 50k topographic map
notes	

The Samples \*.csv file has ~180,000 rows with results as below:

**sample id,lower depth,upper depth,horizon id,texture,method name,lab result value**  
227441,110,65,307427,Medium clay,4A1-pH Soil/water-,9.2  
227441,110,65,307427,Medium clay,12A1\_ZN-Extractable Zn-mg/kg,0.11  
227441,110,65,307427,Medium clay,12A1\_MN-Extractable Mn-mg/kg,3.34  
227441,110,65,307427,Medium clay,12A1\_FE-Extractable Fe-mg/kg,12  
227441,110,65,307427,Medium clay,12A1\_CU-Extractable Cu-mg/kg,1.55  
227441,110,65,307427,Medium clay,15\_NR\_NA-Exchangeable Na+-meq/100g,9.99  
227441,110,65,307427,Medium clay,15\_NR\_MG-Exchangeable Mg++-meq/100g,15  
227441,110,65,307427,Medium clay,15\_NR\_K-Exchangeable K+-meq/100g,2.4

## WESTERN AUSTRALIA

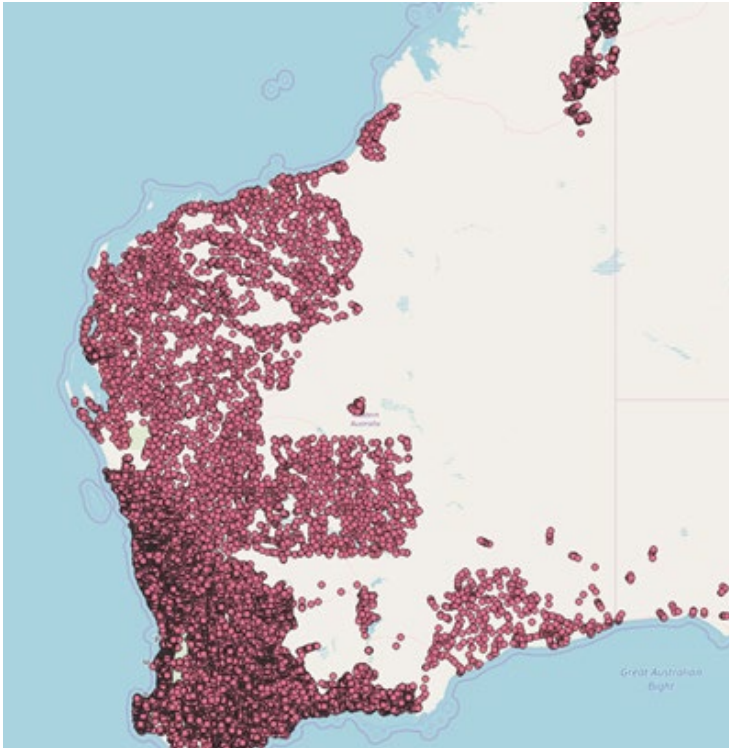


Figure A-3. Distribution of public soil data sites in Western Australia

These data are sourced from [data.wa.gov.au](https://data.wa.gov.au)

- <https://catalogue.data.wa.gov.au/dataset/soil-landscape-mapping-soil-sites>

Downloads are via slip data service which requires registration (free):

- [https://maps.slip.wa.gov.au/datadownloads/SLIP\\_Public\\_Services/Soil\\_Landscape/Soil\\_Landscape.mpk](https://maps.slip.wa.gov.au/datadownloads/SLIP_Public_Services/Soil_Landscape/Soil_Landscape.mpk)

This results in a 6.4GB file download. The \*.mpk file can be unzipped with 7zip or by changing the file extension. It contains a number of (what appear to be) versions of ESRI Geodatabases. These contain several data layers that can be opened in QGIS.

16	crop_rooting_depth_1	38131	MultiPolygon
9	dpird017	1296	MultiPolygon
10	dpird027	137001	MultiPolygon
23	dpird064	34221	MultiPolygon
20	ease_of_excavation_1	50183	MultiPolygon
18	land_instability_hazard_1	7073	MultiPolygon
21	microbial_purification_1	40675	MultiPolygon
19	site_drainage_potential_1	22372	MultiPolygon
26	sl_mapping_wasg	137001	MultiPolygon
25	sl_points_wm	70528	MultiPoint
24	sl_projects	582	MultiPolygon
22	sl_rangelands_april2018	30432	MultiPolygon
0	soil_flood_risk_1	16570	MultiPolygon
1	soil_phos_exp_risk_1	40673	MultiPolygon
2	soil_salinity_risk_1	20553	MultiPolygon
3	soil_subacid_risk_1	20764	MultiPolygon
4	soil_subcomp_risk_1	42313	MultiPolygon
17	soil_water_storage_1	44613	MultiPolygon
5	soil_watERO_risk_2	24931	MultiPolygon
7	soil_watlog_risk_1	33044	MultiPolygon
6	soil_watrep_risk_1	43517	MultiPolygon
8	soil_windero_risk_1	51785	MultiPolygon
12	subsurface_acidity_1	13779	MultiPolygon
14	subsurface_alkalinity_1	26408	MultiPolygon
11	surface_acidity_1	25849	MultiPolygon
13	surface_alkalinity_1	13447	MultiPolygon
15	surface_salinity_1	18132	MultiPolygon

A GeoPackage containing **70,328** sites can be downloaded. A sample record is shown below.

Table A-3 Western Australian Government sample public soil data record.

OBJECTID	23169
id	126832
unique_id	501_NEG_0122_1
o_type	soil pit
latitude_g	-29.077428
longitude_	119.945984
zone	50
easting_gd	786792
northing_g	6779851
wa_soilgrp	406
map_unit	279Ca
s_date_des	1988-09-09T00
o_pos_acc	25 - 100 m
sp_wasg_na	Red shallow sandy duplex
sp_ag_soil	Shallow sandy duplexes
sp_zone	279
sp_max_dep	30
sp_layers	2
o_ref_type	Morphology, little or no Lab data, Geocode reliable

This mapping portal <https://maps.agric.wa.gov.au/nrm-info/> allows these sites to be displayed based on how much information is available but it does not seem possible to get to this underlying information.

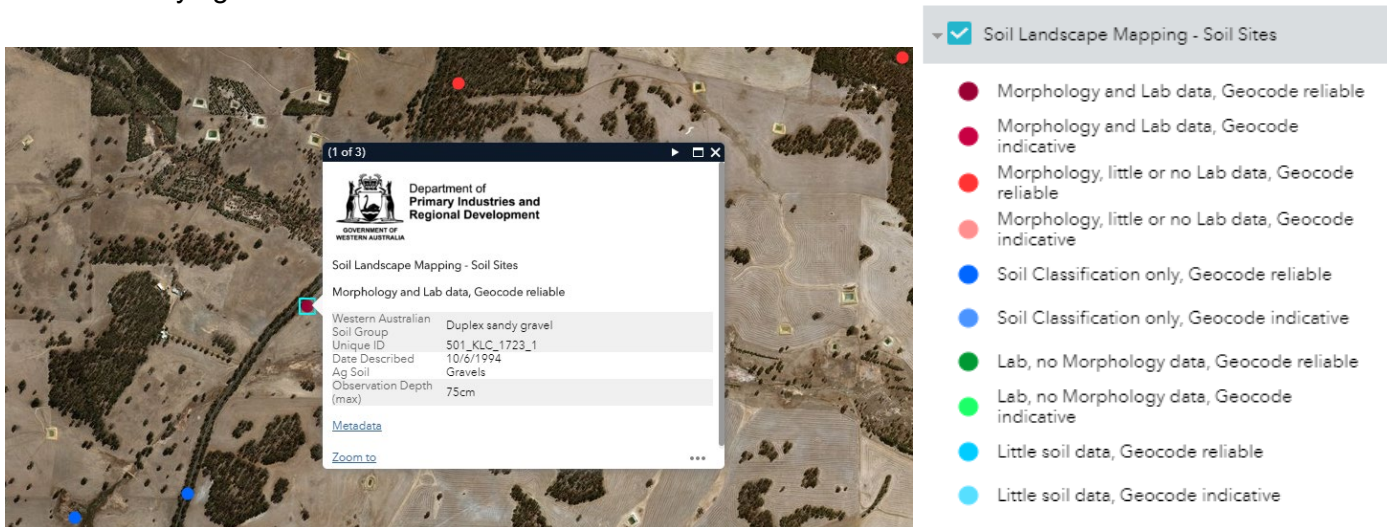


Figure A-4. Examples of information shown on the Queensland Government mapping portal.

Separately there is a soil API available (<https://www.agric.wa.gov.au/soil-api-10>) and requires registration to obtain an API key. The documentation states:

*This API looks at the different percentages of ‘soil super groups’ that are likely to be in a certain area. The API provides an aggregation of the soils, as to what soils are likely to be there.*

In addition, there are several other APIs at [agric.wa.gov.au/web-apis](http://agric.wa.gov.au/web-apis) which seem to be more mature, for example: <https://www.agric.wa.gov.au/science-api-20>

*This API provides the ‘back end’ to the Rainfall to Date, Potential Yield, and Soil Water tools already available on the DPIRD website.*

## NEW SOUTH WALES

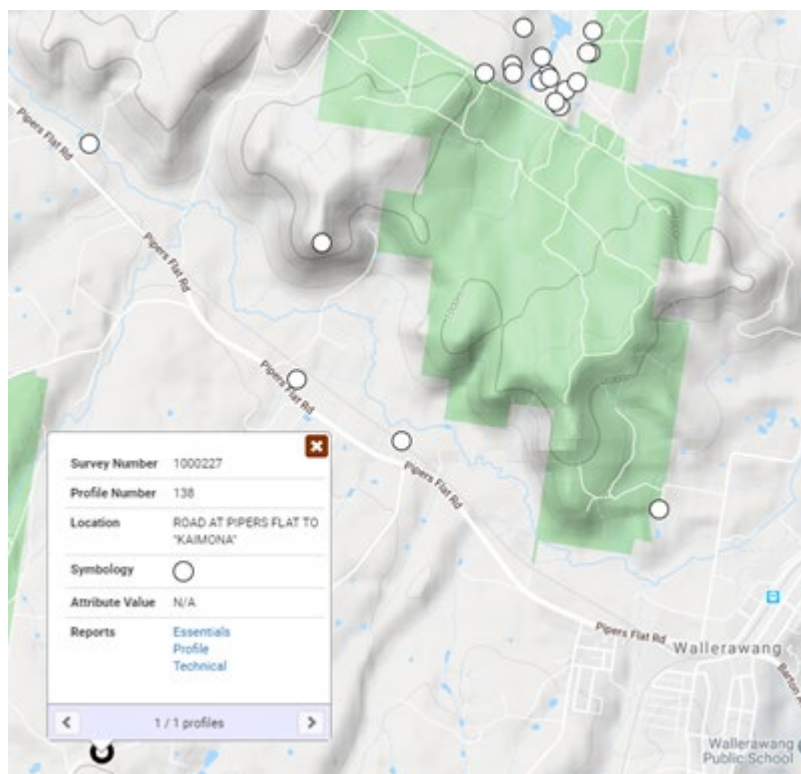


Figure A-5. An example of information shown on the NSW eSPADE portal.

There seems to be no ability to download the data, since no web services that could be found. Everything public requires the user to go through eSPADE web application. The portal is managed by the Office of Environment and Heritage (OEH) rather than through the Department of Primary Industries (DPI).

eSPADE [environment.nsw.gov.au/eSpade2Webapp](https://environment.nsw.gov.au/eSpade2Webapp)

eSPADE provides access to soil profile and soil map information published by the NSW Office of Environment and Heritage, including map data, reports and images, primarily sourced from the NSW Soil and Land Information System (SALIS).

Essentially this constitutes a web map with an internal API that displays the locations of soil profiles once the user zooms in far enough. When an individual site is queried, links to various PDF reports are provided, for example:

- <https://www.environment.nsw.gov.au/espade2webapp/report/essentials/4732>
- <https://www.environment.nsw.gov.au/espade2webapp/report/profile/4732>
- <https://www.environment.nsw.gov.au/espade2webapp/report/technical/4732>

There are two other associated web apps:

- eDIRT (for adding new soil profile data)
- SALIS <https://www.environment.nsw.gov.au/salis5app/> (login access only, requires an account).



## TASMANIA

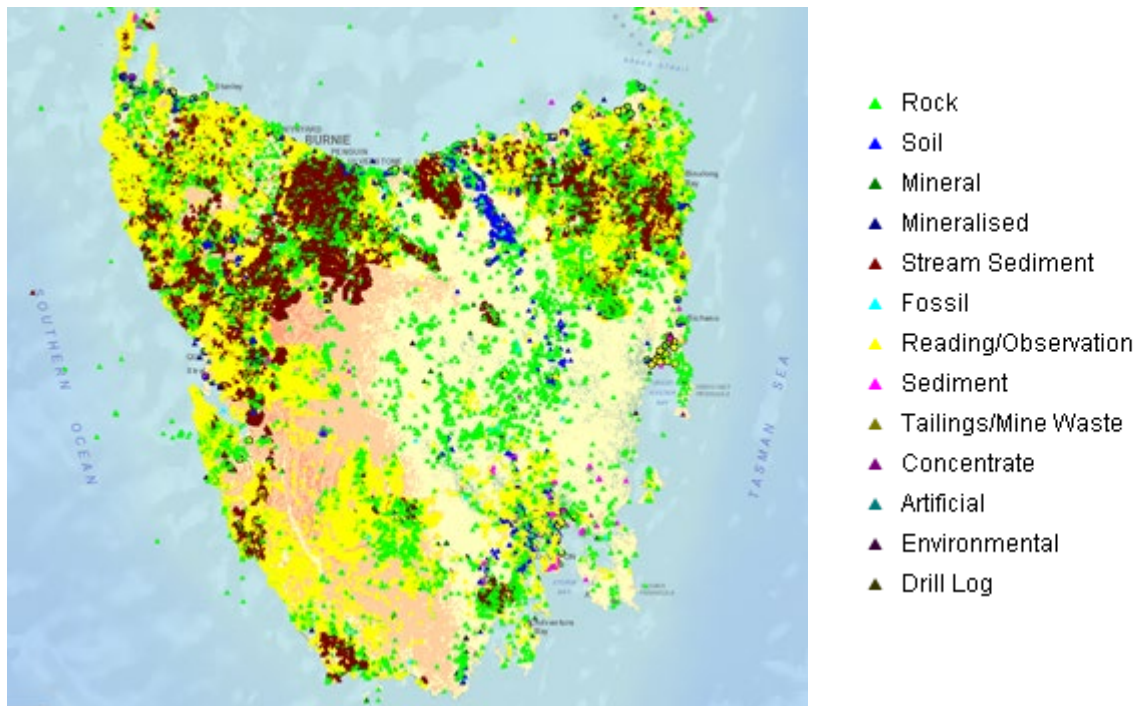


Figure A-6. An example of data on the Tasmanian Government data portal.

There does not appear to be much publicly available soil sites/test information. TheList is the main source of government data:

- <https://www.thelist.tas.gov.au/app/content/data>

A search for 'Soil' returns 73 results. Most are soil classification/series mapping of specific areas and gridded surfaces.

Mineral Resources Tasmania (MRT) appear to hold the most comprehensive dataset 'Soil tests'. TheListMap provides a layer called 'Sample Observation Points' which contains a subset of soil sites, the spatial distribution of Soil Observations can be viewed but only high level metadata is available on each point via query.

There is a search available on the MRT website ([mrt.tas.gov.au/portal/samples-and-geochemistry-search](http://mrt.tas.gov.au/portal/samples-and-geochemistry-search)). While there is not much detail in the search results shown, exporting the results to a CSV file provides a large number of additional fields including lithology & mineral composition.

## VICTORIA

The Victoria Soil Information System (VSIS) has been historically difficult to access. There is a new Agriculture Victoria Soils API (<https://discover.data.vic.gov.au/dataset/agriculture-victoria-soils-api>) but it is fairly limited in its available data and function. Hopefully this will expand over time.

From the [developer.vic.gov.au](https://developer.vic.gov.au) site.

*The Agriculture Victoria Research Soils API provides access to a huge wealth of topsoil and subsoil properties data from across the state of Victoria. API responses contain combined topsoil and subsoil property data from six depths relating to the corresponding area of collated research data.*

The user is required to send a set of coordinates to the API in order to receive soil estimates (from modelled DSM data) at six depths for soil type, soil texture, soil pH, soil carbon %, clay % and electricity conductivity.

## NORTHERN TERRITORY

There is very little soil information so far found on the two available sites:

- <https://data.nt.gov.au/>
- [data.nt.gov.au/dataset/northern-territory-geological-survey-geochemistry-soil](https://data.nt.gov.au/dataset/northern-territory-geological-survey-geochemistry-soil)

## AUSTRALIAN CAPITAL TERRITORY

There is very little soil information so far found on the two available sites:

- ACTMapi <http://www.actmapi.act.gov.au/home.html>
- Soil and hydrogeological landscapes  
<http://app.actmapi.act.gov.au/actmapi/index.html?viewer=shl>

## NATIONAL

The TERN SoilDataFederator developed by CSIRO is very useful:

- <http://esoil.io/TERNLandscapes/SoilDataFederatoR/help/index.html>

It is an aggregator of soil observation data from various state governments departments and other agencies like GA. It has an API to access the observations in a simple standard format.

```
{
  Provider: "NSCC",
  Dataset: "TasGovernment",
  Observation_ID: "601_CRGKH_475_1",
  SampleID: "1",
  SampleDate: "08101986",
  Longitude: 147.4257,
  Latitude: -42.7106,
  UpperDepth: 0.62,
  LowerDepth: 1,
  PropertyType: "LaboratoryMeasurement",
  ObservedProperty: "3A1",
  Value: "0.430000007152557",
  Units: "dS/m",
  QualCollection: "3",
  QualSpatialAgg: "2",
  QualManagement: "5",
  ExtractTime: "2019-11-06T14:42:04"
```

It allows filtering by provider and observed property. There is no spatial filter or grouping by site/sample. Swagger docs are available from:

- [https://esoil.io/TERNLandscapes/SoilDataFederatoR/swagger/#/Soil%20Data%20Federator/get\\_SoilDataAPI\\_SoilData](https://esoil.io/TERNLandscapes/SoilDataFederatoR/swagger/#/Soil%20Data%20Federator/get_SoilDataAPI_SoilData)

All the code including the aggregator is written in R and available from:

- [rdrr.io/github/RossDSearle/SoilDataFederatoR/](https://github.com/RossDSearle/SoilDataFederatoR/)

As an example, this returns soil data for specified provider → Tasmania Government and Observed Property (3A1 - EC):

- <https://esoil.io/TERNLandscapes/SoilDataFederatoR/SoilDataAPI/SoilData?observedProperty=3A1&providers=TasGovernment&key=gkMicmUcKsE8lq&usr=a.macleod@federation.edu.au>

## AUSPLOTS

Yet to be fully reviewed:

- [swarmapi.ausplots.aekos.org.au](https://swarmapi.ausplots.aekos.org.au) (PostgREST endpoint)
- [swarmapi.ausplots.aekos.org.au/om\\_observation?limit=500&"sosa:observedProperty"=eq.ec](https://swarmapi.ausplots.aekos.org.au/om_observation?limit=500&)
- [swarmapi.ausplots.aekos.org.au/om\\_observation\\_collection?\\_id=eq.soil\\_characterisation/58057/WAA053908](https://swarmapi.ausplots.aekos.org.au/om_observation_collection?_id=eq.soil_characterisation/58057/WAA053908)
- [http://swarmapi.ausplots.aekos.org.au/om\\_site\\_visit?\\_id=eq.58057](http://swarmapi.ausplots.aekos.org.au/om_site_visit?_id=eq.58057)
- [swarmapi.ausplots.aekos.org.au/om\\_site?\\_id=eq.60766](https://swarmapi.ausplots.aekos.org.au/om_site?_id=eq.60766)
- [http://swarmapi.ausplots.aekos.org.au/om\\_site\\_point?\\_id=eq.1367542](http://swarmapi.ausplots.aekos.org.au/om_site_point?_id=eq.1367542)
- [http://swarmapi.ausplots.aekos.org.au/om\\_context](http://swarmapi.ausplots.aekos.org.au/om_context)

Looks like about 15,000 points.

## APPENDIX B INFORMATION MODEL

**Note: this section has significant contributions by Bruce Simons and Megan Wong**

The VAS repository for soil data uses an aggregator for use by farmer groups, catchment management authorities, universities and other data custodians who do not currently have the capability to interoperably serve data in the required format. The aggregator takes the form of a set of processes that allow the data custodians to input data through a self-serve system. The self-serve system, when completed, will allow data custodians to upload and set access controls on who gets to view or consume all or parts of their data, and, if required, set an embargo period for that access. The self-serve system also collects the required headers (Appendix C) to auto-create the minimum required metadata to make the data findable and accessible. This includes information about data ownership, licensing, and project information.

The aggregator stores the data in the CeRDI Observations System (CeRDI OS). This system then makes the data available in standard formats, with standardized content, through the VAS portal and APIs (Figure B-1).

CeRDI OS was designed and developed by Bruce Simons and Andrew Macleod in collaboration with the CeRDI research team, VAS participants and other users. From the users' perspective, the system, together with the VAS portal, behaves as their soil data management system that allows the custodians to input data, set access controls, delete data, filter data, graph data, output data in the form of spreadsheets or reports, or serve it directly to other interoperable systems.

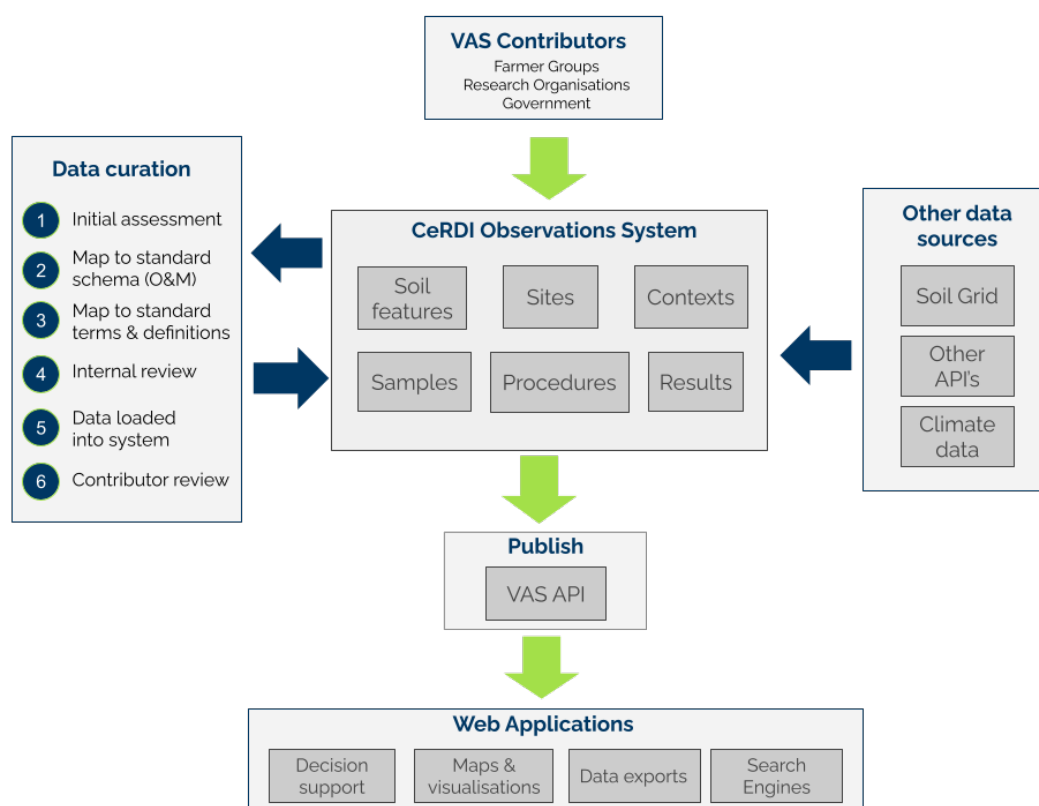


Figure B-1. The conceptual overview of the VAS architecture.

## CERDI OBSERVATIONS SYSTEM (CERDI OS) OVERVIEW

The primary purpose of the observations system is to collect and store observation and measurement data and publish this data for researchers, industry, government and the public. It uses the ISO19156<sup>21</sup> and Open Geospatial Consortium (OGC)<sup>22</sup> Observations and Measurements (O&M) conceptual model to store field and laboratory environmental data in a domain independent structure (Figure B-2). The observation model states:

“An **Observation** is an action whose **result** is an estimate of the value of some **property** of the **feature-of-interest**, obtained using a specified **procedure**.”

The key insights are:

1. to separate
  - the **observation act** from
  - the **procedure**
    - which may be used for other observations and
  - the **feature-of-interest**
    - which has many properties, the values of each of which may be estimated more than once, at different times or using different procedures
2. and to recognise that the outcome of an Observation is a **result**
  - the value of which constitutes an estimate of a value of a property
    - which may be a value or range of values if a measurement, or a term, a term range or a description if an assertion

In addition to standardising the data structure, the system makes use of existing domain-specific controlled vocabularies and ontologies to standardise the semantic content.

---

<sup>21</sup> Cox, Simon Jonathan David (2011). ["ISO 19156:2011 Geographic information – Observations and measurements"](#). International Organization for Standardization. [doi:10.13140/2.1.1142.3042](#)

<sup>22</sup> ["OGC Abstract Specification Topic 20: Observations and measurements"](#). 2010. Retrieved 2010-11-22.

## DATABASE SCHEMA OVERVIEW

### PART 1: OBSERVATIONS, PROCEDURES, SPECIMENS, SAMPLING FEATURES, FEATURES OF INTEREST

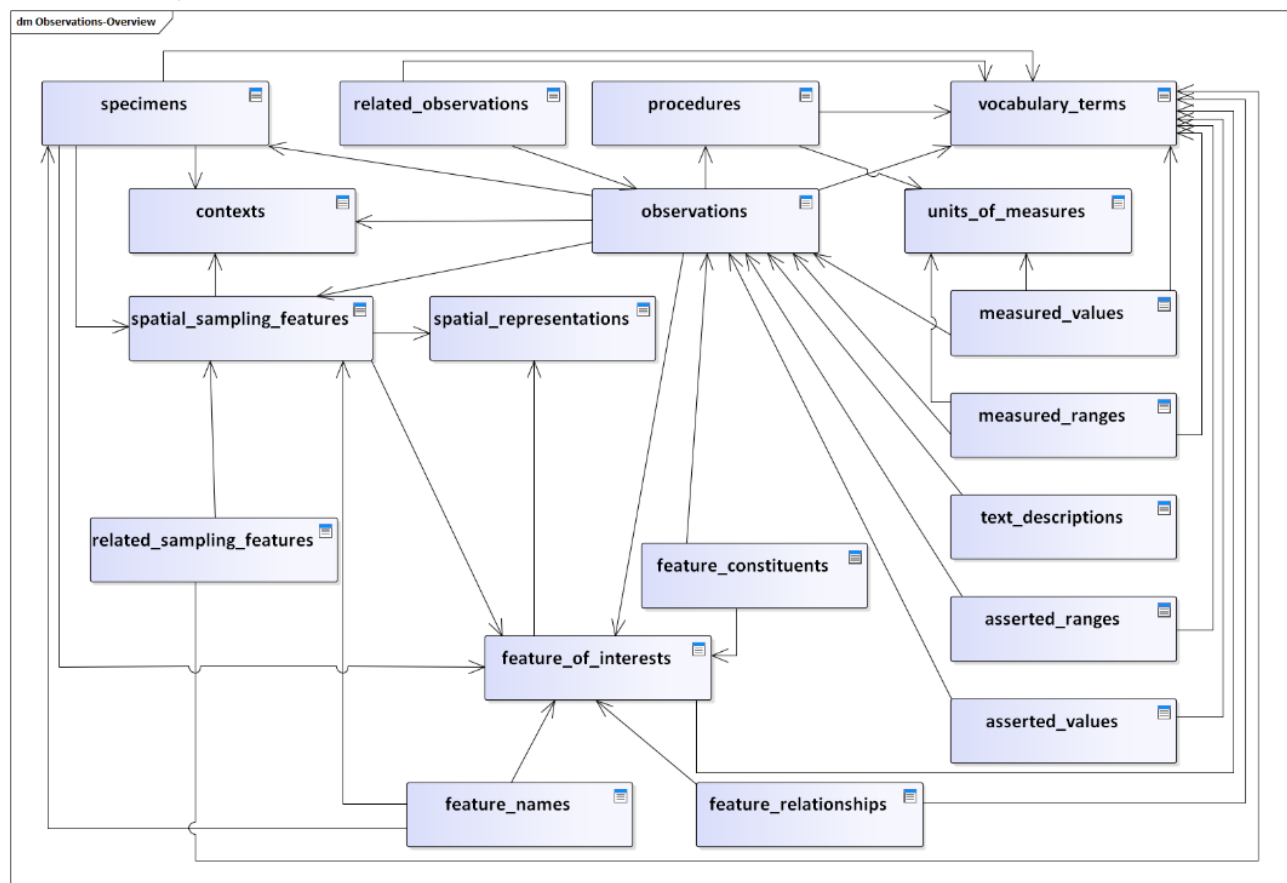


Figure B-2 The CeRDI OS diagrammatic overview: Database Tables relating to Observations.

The 'observations', 'procedures', and 'features\_of\_interests' database tables relate directly to concepts described in O&M. For VAS the 'features of interest' are soil features, such as soil layers (from-to depths), soil horizons, soil profiles and soil bodies. Specimens (for storing sample data) is also taken directly from the O&M model. Additional tables allow capturing relationships and names of these features. Results are stored depending on the type of result: measured values and ranges (with links to associated units of measure), asserted values (terms) and ranges (terms for upper and lower values), and text descriptions.

The 'spatial\_sampling\_feature' table stores information about the sampling feature, that is, the human construct that was used to sample the real world feature whose property is being measured. For VAS the spatial sampling feature is the "Plot", "Site", "Station", "Paddock", etc. along with the sampling method, such as "soil pit", "auger", "grab sample".

The 'vocabulary\_terms' table stores required terms (see "Part 3: Auxiliary Tables (Vocabulary Terms, Units of Measure, Feature Names, Reports)"). These may be terms from traditional look-up tables such as terms associated with 'laboratory methods' (such as "pH using 1.5 HCl", "..."), or with 'drainage' (such as "well-drained", "poorly drained"). They may also be the terms that are traditionally the look-up table names and appear as database columns, such as "laboratory method", "pH", "P concentration" etc. Vocabularies in the vocabulary\_terms

table are also used where terms are required elsewhere in the database, such as specifying the "role" organisations and users play in projects and the type of organisation.

Importantly, in addition to a set of terms the table indicates what schema the term comes from (what the look-up table name would be that the term came from if from a look-up table), a description, and a persistent identifier to an external resource that contains more information (such as images, other relationships, alternative labels or languages etc.).

A separate 'units\_of\_measures' table (see "Part 3: Auxiliary Tables (Vocabulary Terms, Units of Measure, Feature Names, Reports)") is used for the special set of vocabularies with additional properties associated with units of measure, where, in addition to alternative names (e.g. "metre" and "meter") and notation (e.g. "m", "cm"), specifying what quantity kind the unit is used for (e.g. "length") is required. The unit of measure identifier provides a link to an external ontology maintained by an appropriate authority.



## Observations

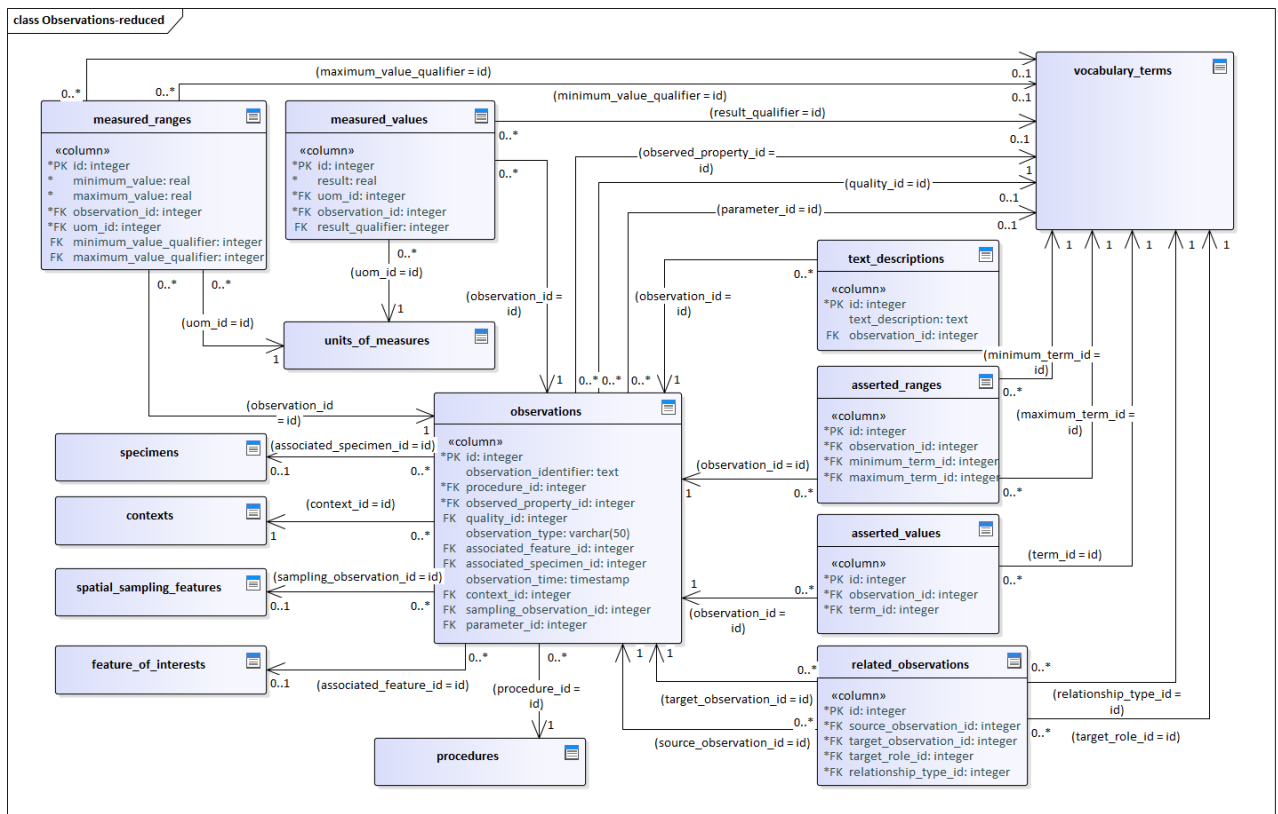


Figure B-3. Diagram: "observations" and the results tables. Columns of related tables and associated relationships are not shown.

### 'observations' table description

The **observations** table records information about individual observations (measurement or assertions) made on the soil feature.

The type of Observation (e.g. "Laboratory Measurement", "Field Observation"). This could potentially be expanded to include "Application Rate", "Crop Yield", etc.

The result of the Observation could be a numerical range, single numerical result, a range of terms or a single term. This could be expanded to include raster, time series, image results.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
observation_identifier	text	False	The persistent identifier of the individual Observation (URI)
procedure_id	integer	True	The foreign key to the procedures table identifying the procedure used to generate the observation
observed_property_id	integer	True	Foreign key to the vocabulary_terms table identifying the observed property being measured (e.g. "K concentration", "pH").

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
quality_id	integer	False	Foreign key to the vocabulary_terms table to describe the quality of the observation.
observation_type	varchar(50)	False	The type of Observation (e.g. "Laboratory Measurement", "Field Observation"). This could be expanded to include "Application Rate", "Crop Yield", etc. Future upgrade to a Foreign key to the vocabulary_terms table
associated_feature_id	integer	False	Foreign key to the soil feature the observations were made on (the real world FeatureOfInterest such as a Soil Body, Soil Horizon, Soil Profile, Soil Layer).
associated_specimen_id	integer	False	The foreign key to the (Soil) Specimen that the observation was made on.
observation_time	timestamp	False	The time and date of the observation. The O&M model allows for additional time properties, such as the time the observation was applicable, the time the observation was made, the time the observation is valid for. These variations have not been included here.
context_id	integer	False	A foreign key to the contexts table specifying the context in which the observation was made. Usually a job or set of tasks undertaken by a single observer/machine as part of a broader project. Links Job, Project, Client information
sampling_observation_id	integer	False	A foreign key to the feature that was used to sample the real world environmental feature such as SoilProfile, or Specimen that the observation was made on. SpatialSamplingFeatures may be "Plot", "Site", "Station", "Paddock", etc.
parameter_id	integer	False	Describes an arbitrary event-specific parameter. This might be an environmental parameter, an instrument setting or input, or an event-specific sampling parameter that is not tightly bound to either the feature-of-interest or to the observation procedure. In some contexts the <b>Observation::procedure</b> is a generic or standard procedure, rather than an event-specific process. In this context, parameters bound to the observation act, such as instrument settings, calibrations or inputs, local position, detection limits, asset identifier, operator, may augment the description of a standard procedure. EXAMPLE The fraction of the soil ('coarse fraction', 'fine fraction', 'whole soil') from which the particle size proportion was taken.

### 'related\_observations' table description

The **related\_observations** table records information about relationships between individual observations.

In addition to identifying observations that were made as part of a project for example, this table could be used to relate previous application or yield results to current results.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
source_observation_id	integer	True	The foreign key to the observation that is the source in the relationship
target_observation_id	integer	True	The foreign key to the Observation that is the target in the relationship
target_role_id	integer	True	The foreign key to the vocabulary_terms table that specifies the role that the target Observation plays in the relationship (e.g. "previous application rate", "related observation", "alternative result").
relationship_type_id	integer	True	The foreign key to the vocabulary_terms table that specifies the type of relationship between the two Observations

The results of an observation may be one of five types:

1. A numeric value resulting from a measurement (**measured\_values**)
2. A numeric range resulting from a set of measurements (**measured\_ranges**)
3. A term value resulting from an asserted value (**asserted\_values**)
4. A range of term values resulting from a set of asserted values (**asserted\_ranges**)
5. A description (**text\_descriptions**).

Depending on the type of result, these are stored in separate tables.

#### 1. 'measured\_values' table description

Table to store empirical results from laboratories, field measuring devices etc. where the result is a single numeric value.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
result	real	True	The numerical result of the Observation
uom_id	integer	True	The Unit Of Measure for the result. This is a foreign key to the units_of_measures Table.
observation_id	integer	True	The Observation that the result relates to.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
result_qualifier	integer	False	Foreign key to the vocabulary_terms table. Qualifies the result. Examples are '<', '<=', '>', '>=', '~', 'below detection limit', 'above detection limit'. Also to capture the gml:nilReasonTypes ('unknown', 'withheld', 'inapplicable', 'missing', 'template').

## 2. 'measured\_ranges' table description

Table to store empirical results from laboratories, field measuring devices etc. where the result is a single numeric value.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id for the measured range
minimum_value	real	True	The smallest numerical value in the range
maximum_value	real	True	The largest numerical value in the range
observation_id	integer	True	The foreign key to the Observation which created this result
uom_id	integer	True	A foreign key to the units_of_measures table for the Unit Of Measure
minimum_value_qualifier	integer	False	Foreign key to the vocabulary_terms table. Qualifies the minimum value result. Examples are '<', '<=', '>', '>=', '~', 'below detection limit', 'above detection limit'. Also to capture the gml:nilReasonTypes ('unknown', 'withheld', 'inapplicable', 'missing', 'template').
maximum_value_qualifier	integer	False	Foreign key to the vocabulary_terms table. Qualifies the maximum value result. Examples are '<', '<=', '>', '>=', '~', 'below detection limit', 'above detection limit'. Also to capture the gml:nilReasonTypes ('unknown', 'withheld', 'inapplicable', 'missing', 'template').

## 3. 'asserted\_values' table description

Table to store the result of an observation that is a term from a vocabulary.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
observation_id	integer	True	The foreign key to the Observation that generated the term result
term_id	integer	True	The foreign key to the vocabulary_terms table

#### 4. 'asserted\_ranges' table description

Table to store the result of an observation where the results are a lower term and an upper term.

COLUMN NAME	DATATYPE	NOT NULL	DESCRIPTION
id	integer	True	The database id
observation_id	integer	True	The foreign key to the Observation that generated the term ranges
minimum_term_id	integer	True	The foreign key to the vocabulary_terms table for the minimum term (e.g. "poorly drained").
maximum_term_id	integer	True	The foreign key to the vocabulary_terms table for the maximum term (e.g. "well-drained").

#### 5. 'text\_descriptions' table description

Table to store the result of an observation where the result is a free text description.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	database id
text_description	text	False	The free text description result
observation_id	integer	False	The foreign key to the Observation that generated the free text description.

## Procedures

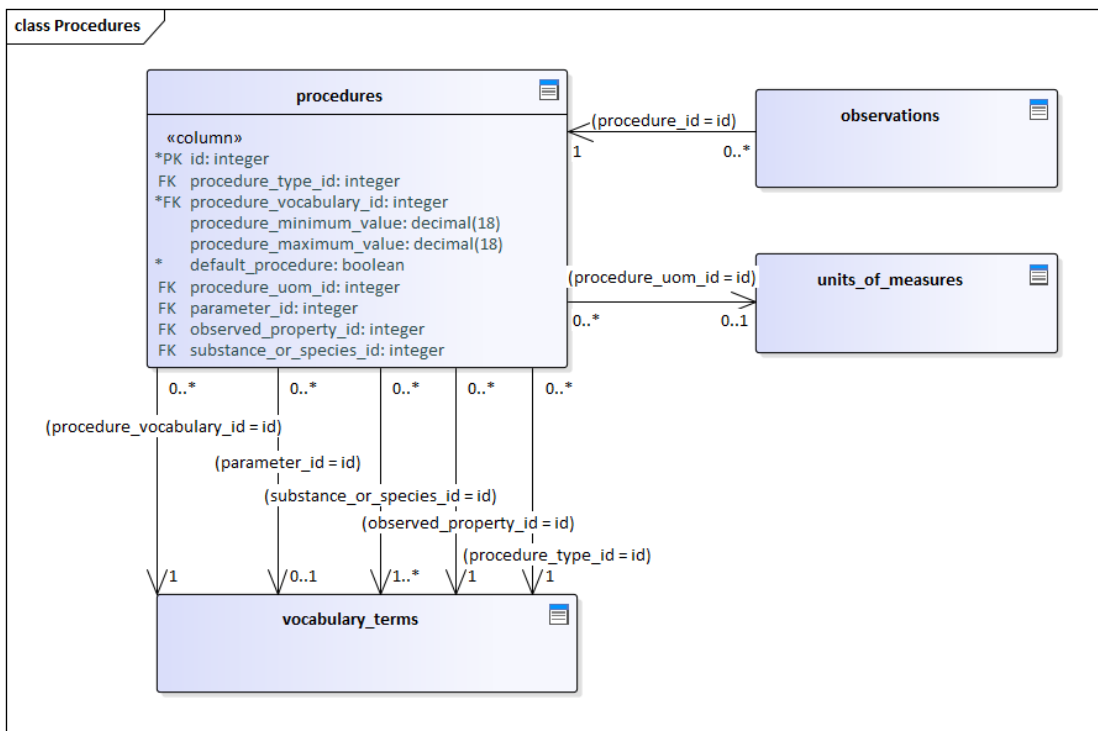


Figure B-4. Diagram: “procedures” table. Columns of related tables and associated relationships are not shown.

### ‘procedures’ table description

This table records information about the procedures or methods used to record the observation. A challenge with recording procedures is that they sometimes need to be treated as a series of activities or steps, sometimes as a term with additional procedure specific information, and sometimes just as a term from a vocabulary. The former can cater for these variations, but comes with added complexity, particularly during data entry. The current database design caters for the second approach by accommodating additional information, such as whether the procedure is the default method, what result unit of measure the procedure uses, and what the data ranges.

The purpose of the 'procedures' table is to group the properties of units of measure, observed property, minimum and maximum values and substance or species that apply to a certain procedure. From an observation perspective the 'procedures' table only provides a link via the procedure\_id to the vocabulary term that identifies the procedure. The Units of Measure for the actual observation are stored against the result. The observed property for the observation is stored against the observation.

The downside of this is that there is no guarantee that the Units of Measure and observed property associated with the selected procedure will match the Units of Measure stored against the result or the observed property stored against the observation.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
procedure_type_id	integer	False	A foreign key to the vocabulary_terms table specifying the type of procedure ("measurement", "interpretation", "assertion", "calculation").
procedure_vocabulary_id	integer	True	A foreign key to the vocabulary_terms table that identifies the procedure used (e.g. "pH using 1:5 HCl")
procedure_minimum_value	decimal(18)	False	The minimum allowable value for the procedure if the result is expected to be a measurement.
procedure_maximum_value	decimal(18)	False	The maximum allowable value for the procedure if the result is expected to be a measurement.
default_procedure	boolean	True	Whether the Procedure is the default procedure for this Observed Property
procedure_uom_id	integer	False	A foreign key to the units_of_measures table corresponding to the unit of measure for the procedure
parameter_id	integer	False	A foreign key to the vocabulary_terms table to specify a procedure specific parameter term
observed_property_id	integer	False	A foreign key to the vocabulary_terms table that identifies the observed property being measured by the procedure (e.g. "soil texture")
substance_or_species_id	integer	False	A foreign key to the vocabulary_terms table of the term that identifies the object, substance or species that the property is related to (e.g. "Potassium", "Ammonia").

## Specimens

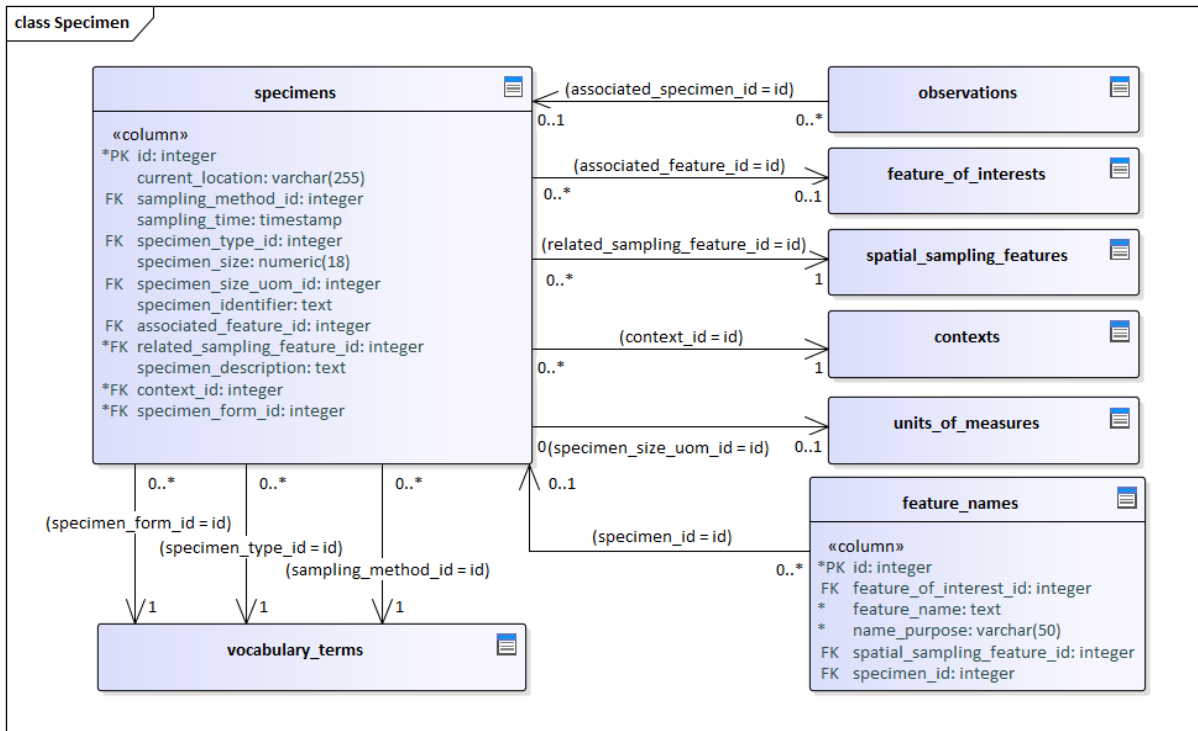


Figure B-5. Diagram: 'specimens' table. Columns of related tables and associated relationships are not shown

### 'specimens' table description

The Specimen table captures information relating to Specimens (= Soil Samples). These may be temporary field specimens that have been discarded or ones that have been archived. It is analogous to O&M SF\_Specimen.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id of the Specimen
current_location	varchar(255)	False	The current storage facility and/or shelf and/or container of the Specimen
sampling_method_id	integer	False	A foreign key to the term in the vocabulary_terms table that identifies the sampling method. In O&M this is an SF_Process, i.e. the sampling procedure consists of a series of steps (to describe for instance combining samples from 10 sites along a transect), not just a single term method. This may be required in future developments.
sampling_time	timestamp	False	The date/time that the specimen was obtained in the field.



COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
specimen_type_id	integer	False	A foreign key to the term in the vocabulary_terms table that identifies the specimen type.
specimen_size	numeric(18)	False	The size of the actual specimen.
specimen_size_uom_id	integer	False	A link to the term in the units_of_measures table that identifies the Unit of Measure of the specimen size.
specimen_identifier	text	False	The persistent identifier (URI) of the Specimen
associated_feature_id	integer	False	A foreign key to the associated features_of_interests table (i.e. the Soil Feature)
related_sampling_feature_id	integer	True	A foreign key to the spatial_sampling_features table to the feature that sampled the Specimen.
specimen_description	text	False	A text description or comment associated with the Specimen.
context_id	integer	True	A foreign key to the context in which the specimen was taken. Usually a job or set of tasks undertaken by a single observer/machine as part of a broader project. Links Job, Project, Client information
specimen_form_id	integer	True	A foreign key to the vocabulary_terms table specifying the basic form of the specimen. e.g. "polished section", "core", "pulp", "solution" Corresponds to SF_Specimen/specimenType

## Sampling Features

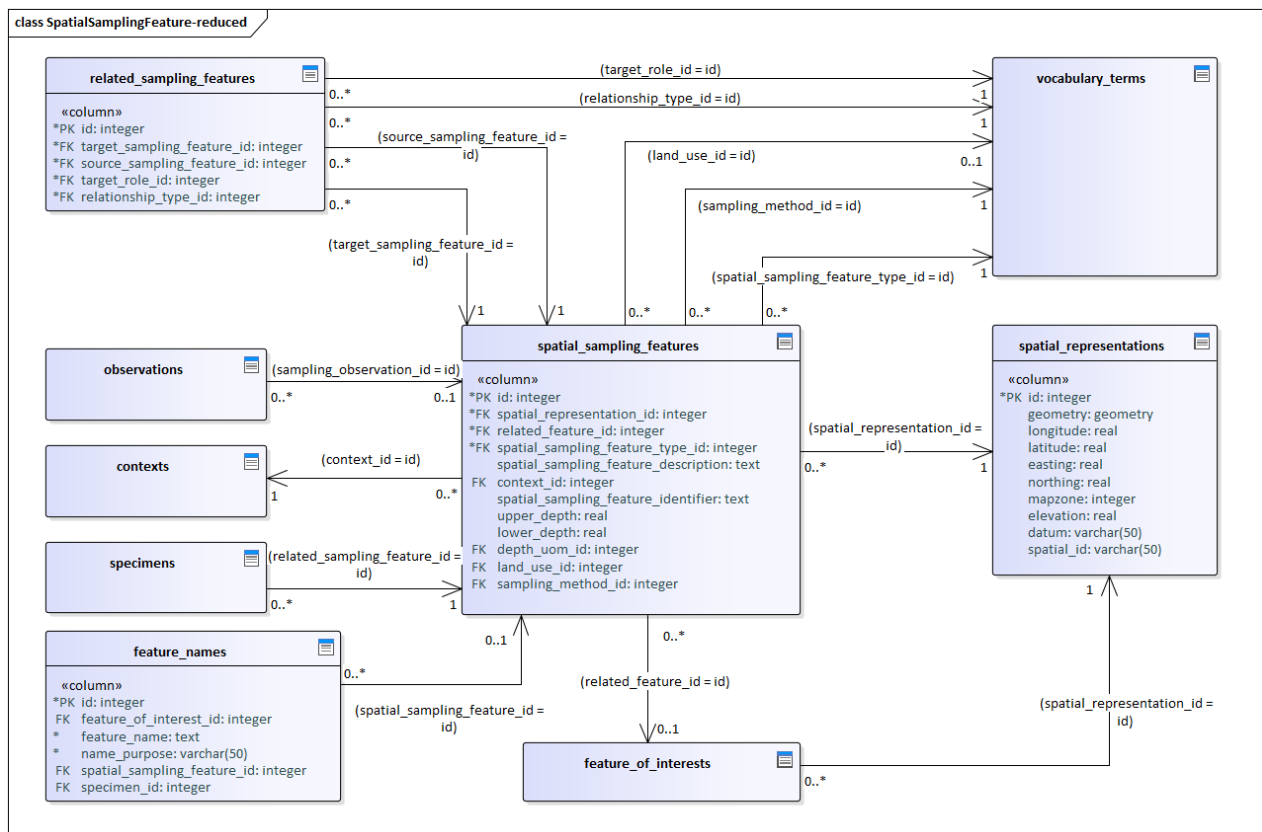


Figure B-6. Diagram: “spatial\_sampling\_features”, “related\_sampling\_features” and “spatial\_representations” tables. Columns of related tables and associated relationships are not shown. Together the “spatial\_sampling\_features” and “spatial\_representations” tables describe the 3 dimensional geometry

### ‘spatial\_sampling\_features’ table description

The spatial\_sampling\_features and spatial\_representations tables store properties analogous to those of the O&M SF\_SpatialSamplingFeature. A spatial sampling feature is used when property observations are made on a geospatial feature. This sampling feature may be in one, two or three spatial dimensions. Properties observed on sampling features may be time-dependent, but the temporal axis does not generally contribute to the classification of sampling feature classes. Sampling feature identity is usually less time-dependent than is the property value.

The SF\_SpatialSamplingFeature is a type of SF\_SamplingFeature. A SF\_SamplingFeature is intended to sample some feature of interest in an application domain. They are artefacts of an observational strategy and have no significant function outside of their role in the observation process. The physical characteristics of the features themselves are of little interest, except perhaps to the manager of a sampling campaign.

SpatialSamplingFeatures may be ‘Plot’, ‘Site’, ‘Station’, ‘Paddock’, etc. These types are managed through the vocabulary\_terms table. A ‘station’ is essentially an identifiable locality where a sensor system or procedure may be deployed and an observation made. In the context of the observation model, it connotes the ‘world in the vicinity of the station’, so the observed properties relate to the physical medium at the station, and not to any physical

artefact such as a peg, stake, fencepost, monument, well, etc. A transient sampling feature, such as a quadbike-track or drone flight-line, may be identified and described, but is unlikely to be revisited exactly. The sampling method, such as “soil pit”. “hand auger”, “shovel”, is included as part of the spatial sampling feature.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
spatial_representation_id	integer	True	A link to the SpatialRepresentation of the SpatialSamplingFeature
related_feature_id	integer	True	The real-world feature (e.g. for a SoilFeature the Soil Body, Soil Horizon, Soil Layer, Soil Profile) that the SpatialSamplingFeature samples.
spatial_sampling_feature_type_id	integer	True	Identifies the type of SpatialSamplingfeature (e.g. Pit, Plot, Site, Paddock, Borehole, etc). This is a link to a term from a vocabulary in the VocabularyTerm table.
spatial_sampling_feature_description	text	False	A text description or comment related to the SpatialSamplingFeature
context_id	integer	False	A foreign key to the context in which the specimen was taken. Usually a job or set of tasks undertaken by a single observer/machine as part of a broader project. Links Job, Project, Client information
spatial_sampling_feature_identifier	text	False	The persistent identifier (URI) of the SpatialSamplingFeature
upper_depth	real	False	The distance from the surface to the top of the sampling feature
lower_depth	real	False	The distance from the surface to the bottom of the sampling feature
depth_uom_id	integer	False	Identifies the units_of_measures unit that contains the Unit of Measure for the depth values.
land_use_id	integer	False	Identifies the vocabulary term that corresponds to the land use of the spatial sampling feature.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
sampling_method_id	integer	False	A link to the term in the VocabularyTerm table that identifies the sampling method. This should be a SF_Process, i.e. the sampling procedure consists of a series of steps (to describe for instance combining samples from 10 sites along a transect), not just a single term method. This may be required in future developments.

### **‘spatial\_representations’ table description**

Spatial representation (geometry) of the feature of interest. The depth range and associated unit of measure for true 3D information is stored with the associated spatial\_sampling\_feature due to database geometry handling considerations.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
geometry	geometry	False	FK to the Spatial geometry of the feature
longitude	real	False	
latitude	real	False	
easting	real	False	
northing	real	False	
mapzone	integer	False	
elevation	real	False	
datum	varchar(50)	False	
spatial_id	varchar(50)	False	A non database, non-unique historical id for the GIS geometry.

### **‘related\_sampling\_features’ table description**

A table that identifies the relationships between the Spatial Sampling Features and the roles each feature plays in the relationship.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id of the relationship
target_sampling_feature_id	integer	True	The target SpatialSamplingFeature

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
source_sampling_feature_id	integer	True	The source SpatialSamplingFeature
target_role_id	integer	True	A link to a term in the vocabulary_terms table that identifies the role the target SpatialSamplingFeature plays in the relationship.
relationship_type_id	integer	True	A link to a term in the vocabulary_terms table that identifies the type of relationship between the SpatialSamplingFeatures.

## Features of Interest

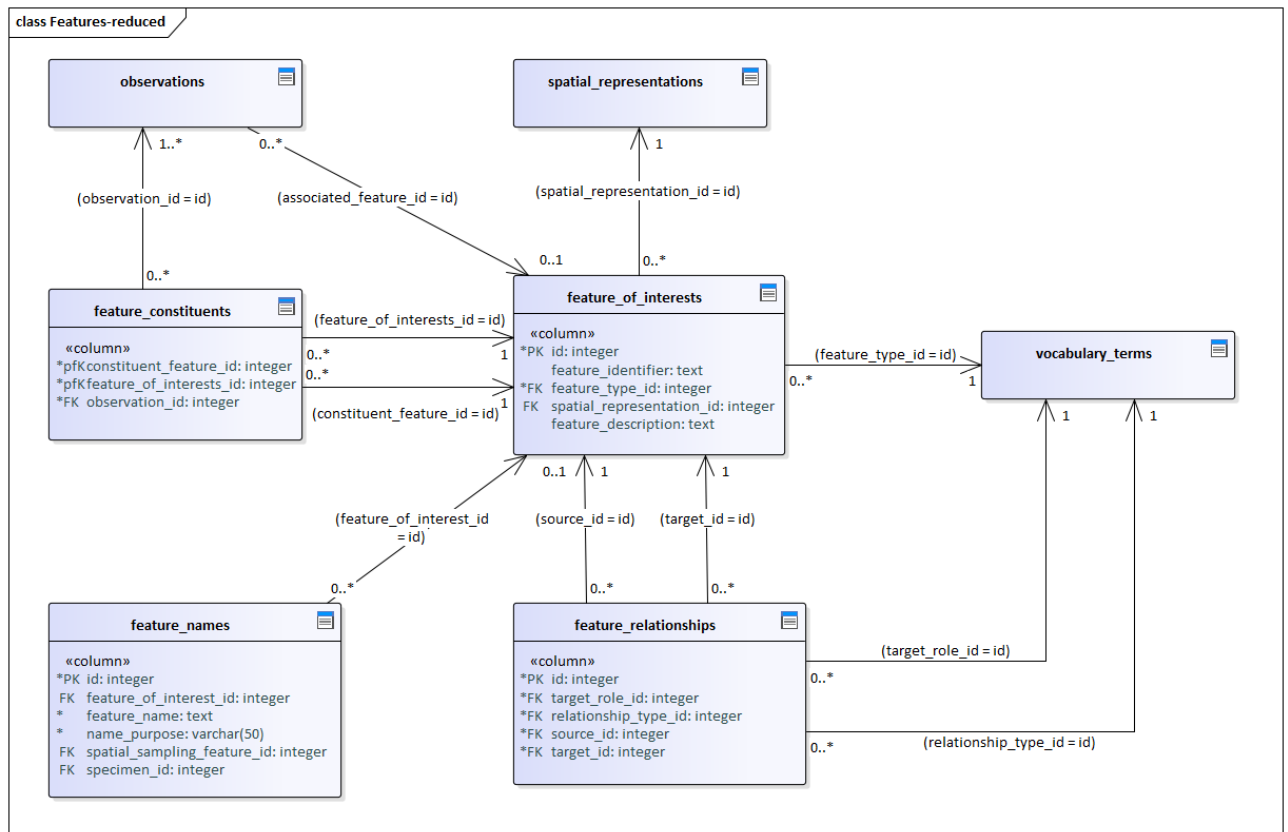


Figure B-7. Diagram: “features\_of\_interests”, “feature\_relationships” and “feature\_constituents” tables. Columns of related tables and associated relationships are not shown.

### ‘feature\_of\_interests’ table description

A generic class to handle the real-world features being described by the observations. SoilFeature:featureTypes are "SoilBody", "SoilHorizon", "SoilLayer", "SoilProfile". GroundwaterFeature:FeatureTypes are Aquifer, Borehole, FluidBody

The featureTypes are managed through the vocabulary\_terms table with values populated from a feature type catalogue.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id of the Feature of interest
feature_identifier	text	False	Unique persistent identifier (URI) of the feature
feature_type_id	integer	True	Whether a "SoilBody", "SoilHorizon", "SoilProfile", "SoilLayer", "Aquifer", "EarthMaterial", "SoilConstituent", etc. Values from the vocabulary_terms table.
spatial_representation_id	integer	False	The link to the spatial representation, including depth range, of the Feature.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
feature_description	text	False	A free text description of the feature.

#### **'feature\_relationships' table description:**

A table that identifies the relationships between the Features of Interests and the roles each feature plays in the relationship. Captures relationships between features such as Soil Horizon to Soil Horizon, Soil Horizon to Soil Body, Soil Horizon to Soil Profile and Soil Layer to Soil Body relationships.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id of the relationship
target_role_id	integer	True	Identifies the relationship role the target plays from VocabularyTerm table.
relationship_type_id	integer	True	Identifies the relationship type in VocabularyTerm table
source_id	integer	True	Identifies the source soil feature in SoilFeature table
target_id	integer	True	Identifies the target soil feature in SoilFeature table

#### **'feature\_constituents' table description**

A table that identifies the constituent parts of a feature and the proportion that each part comprises.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
constituent_feature_id	integer	True	Foreign Key to the constituent feature of interest. Is part of a composite primary key with the feature_of_interest_id - allows the PK index to group record for the same feature together.
feature_of_interests_id	integer	True	Foreign Key to the undivided feature of interest. Is part of a composite primary key with the constituent_interest_id - allows the PK index to group record for the same feature together.
observation_id	integer	True	The observation that provides the proportion ('asserted values', 'measured ranges', 'measured values')

## PART 2: CONTEXTS (CONTEXTS, PROJECTS, ORGANISATIONS, USERS)

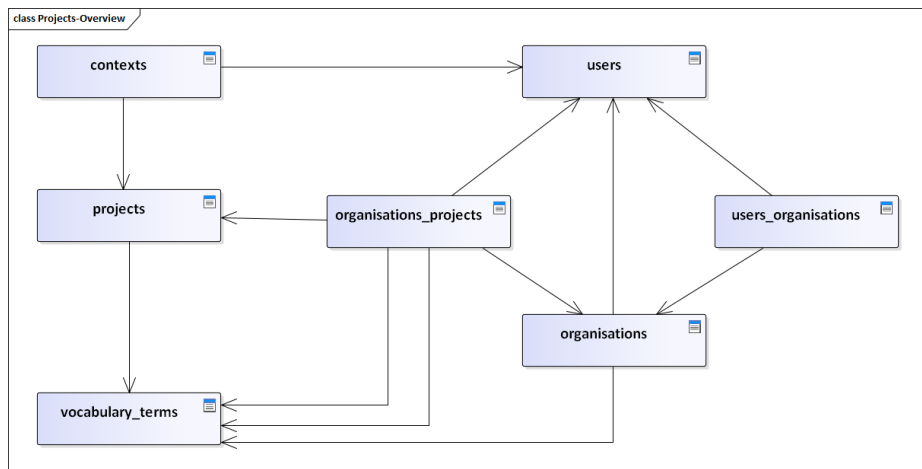


Figure B-8. Diagram: Database tables that relate to data contexts

The 'Contexts' component stores information about the context of the observational data, its 'metadata'. This includes job information, project names, organisations and their roles, and users and their roles.

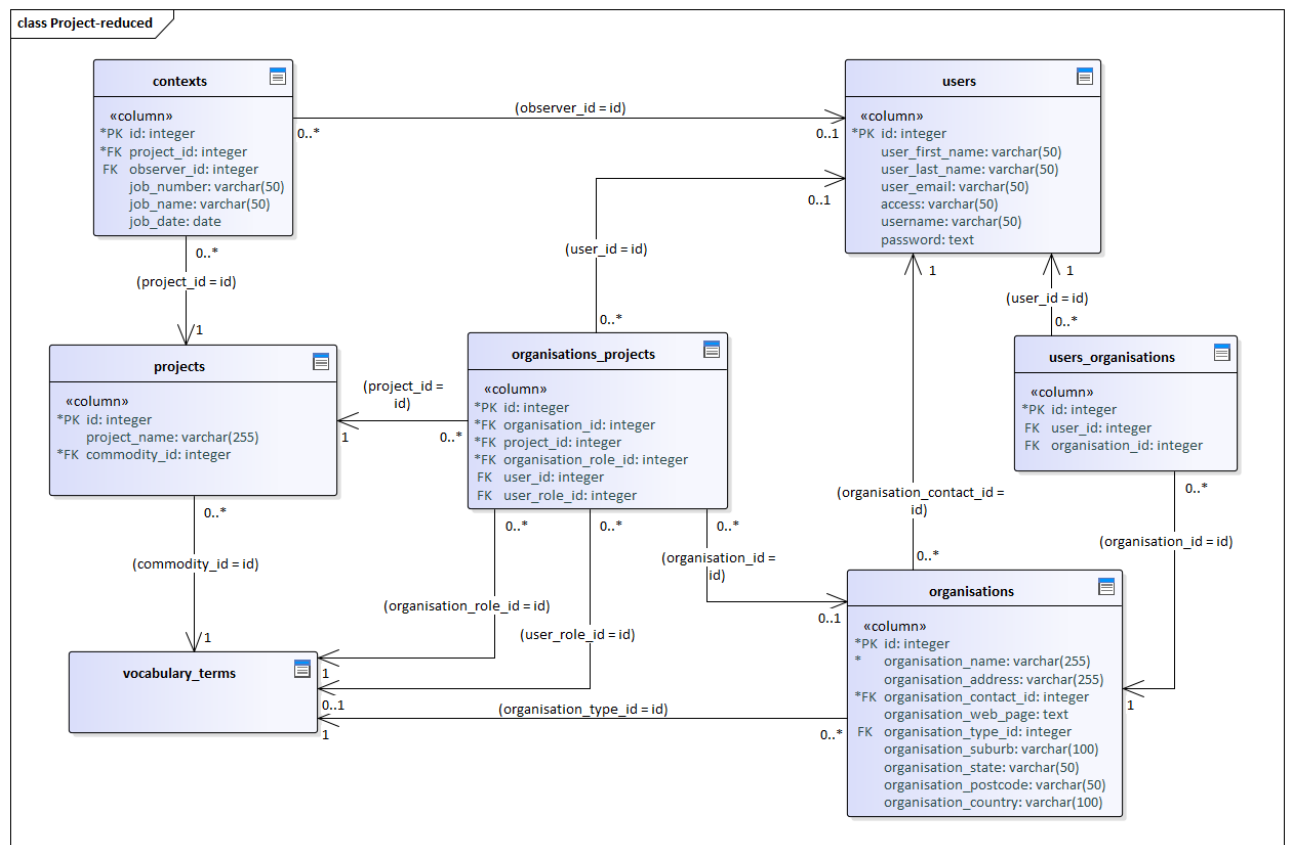


Figure B-9. Diagram: Database tables that relate to data contexts: "contexts", "projects", "users", "organisations", "organisations\_projects" and "users\_organisations" tables. Columns of related tables and associated relationships are not shown.



### **‘contexts’ table description**

The context in which the observation, specimen, sampling feature was made. Usually a job or set of tasks undertaken by a single observer/machine as part of a broader project. Links Job, Project, Client information

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
project_id	integer	True	id of the Project associated with the sampling/observing activity
observer_id	integer	False	id of the user who made the observation or took the sample
job_number	varchar(50)	False	The job number.
job_name	varchar(50)	False	The job name
job_date	date	False	Date the job was initiated. Necessary?

### **‘projects’ table description**

The Project that the observations and/or sampling were part of. Projects are a collection of activities, jobs or tasks that together have a common goal. Although multiple organisations may be associated with a Project, usually only a single organisation has the role “client”.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id
project_name	varchar(255)	False	The name of the Project
commodity_id	integer	True	The FK to the id in the vocabulary_terms table that corresponds to the commodity that the project is primarily interested in, such as “soil”, “groundwater”.

### **‘users’ table description**

An individual user’s contact details. Multiple users may have multiple roles in a Project.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database id of the user
user_first_name	varchar(50)	False	First name of the contact
user_last_name	varchar(50)	False	The user's last name

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
user_email	varchar(50)	False	email contact of the user
access	varchar(50)	False	
username	varchar(50)	False	
password	text	False	

### **‘organisations’ table description**

Individual organisation details.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	Database id of the organisation
organisation_name	varchar(255)	True	Name of the organisation
organisation_address	varchar(255)	False	First line of Address of the Organisation
organisation_contact_id	integer	True	Link to the individuals information
organisation_web_page	text	False	The Web home page of the organisation
organisation_type_id	integer	True	A FK to the vocabulary_terms table that corresponds to the term that describes the type of organisation, such as “government”, “research”, “grower group”.
organisation_suburb	varchar(100)	False	Town or suburb address of the organisation
organisation_state	varchar(50)	False	State or Territory of the address of the organisation
organisation_postcode	varchar(50)	False	Postcode of the address of the organisation. Needs to be able to handle international postcodes
organisation_country	varchar(100)	False	Country of the address of the organisation

### **‘organisations\_projects’ table description**

Multiple organisations may be associated with a project and multiple projects per organisation.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	Database id
organisation_id	integer	True	FK to the Organisation Table
project_id	integer	True	FK to the Project table

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
organisation_role_id	integer	True	FK to the vocabulary_terms table for the term that corresponds to the role the organisation plays in the Project. A project, such as 'Precision Agriculture soil tests' may have multiple organisations such as 'Farm A', 'Farm B' playing the role 'contributor' with the organisation 'Precision Agriculture' playing the role 'client'. In different projects the same organisations may play different roles.
user_id	integer	False	FK to the users table for the user who represents the organisation in the Project
user_role_id	integer	False	FK to the vocabulary_terms table for the term identifying the role the user has in the Project

### PART 3: AUXILIARY TABLES (VOCABULARY TERMS, UNITS OF MEASURE, FEATURE NAMES, REPORTS)

The 'vocabulary\_terms' and 'units\_of\_measures' tables are central to providing controlled terms to the concepts stored in the database. These two tables play the role of 'look-up' tables, with additional associated information and links to external references and ontologies for the concepts. The intention is that external parties manage these terms using external applications that then make the vocabularies available via web services. The CeRDI Observations Database harvests these concepts and stores them as a local cache.

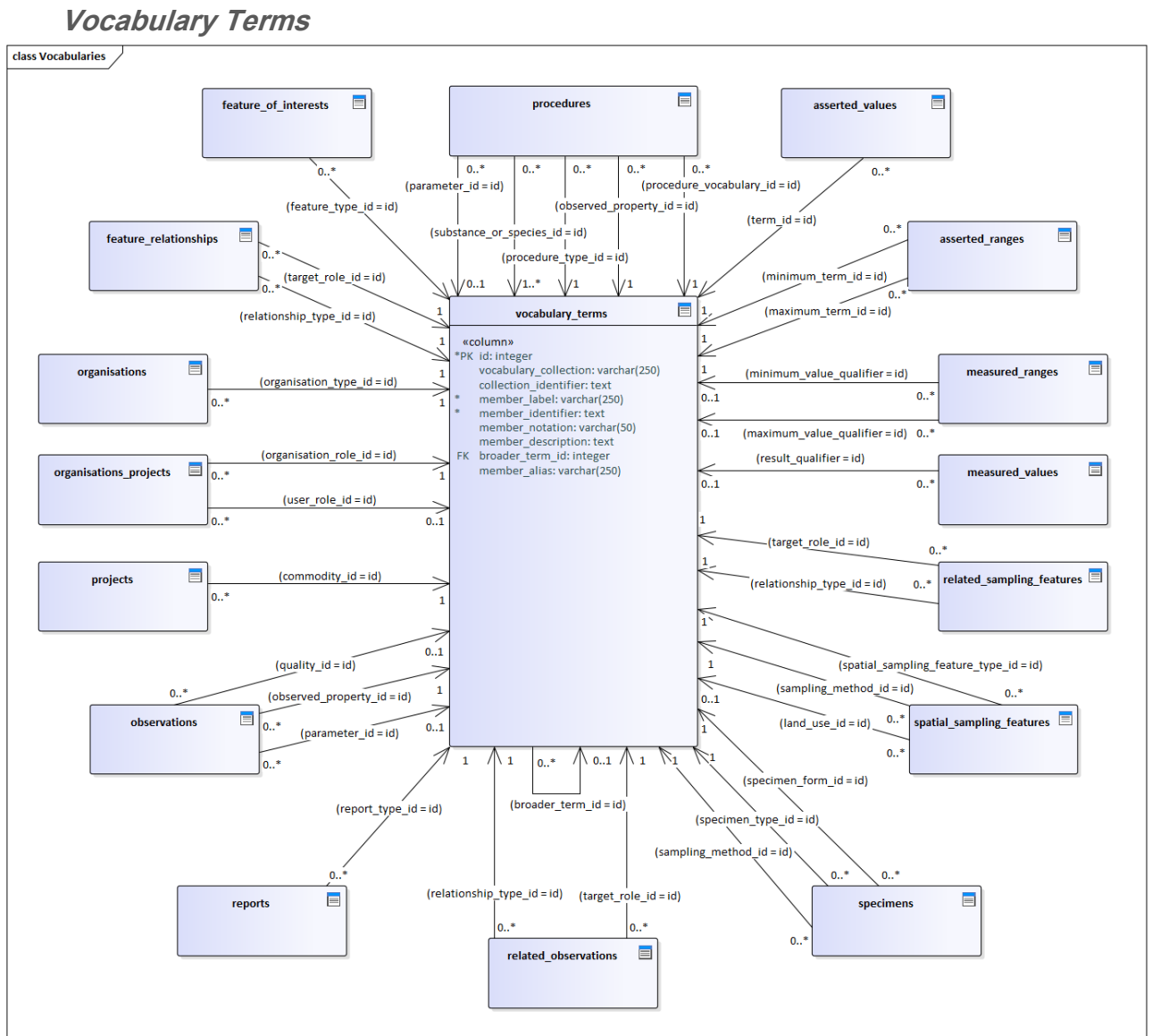


Figure B-10. Diagram: The 'vocabulary\_terms' table and the associated tables that use it. Columns of related tables and associated relationships are not shown.

### ‘vocabulary\_terms’ table description

A table to store required terms. These may be terms from traditional look-up tables such as terms associated with 'laboratory methods' (such as "pH using 1.5 HCl"), or with 'drainage' (such as "well-drained", "poorly drained"). They may also be the terms that are traditionally the look-up table names and appear as database columns, such as "laboratory method", "pH", "P concentration" etc.

Importantly, in addition to a set of terms the table indicates what schema the term comes from (what the look-up table name would be that the term came from if from a look-up table), a description, and a persistent identifier to an external resource that contains more information (such as images, other relationships, conversion values, alternative labels or languages etc.).

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
vocabulary_collection	varchar(250)	False	The label of the collection that the vocabulary member belongs to (e.g. "drainage", "laboratory method", "observed property")
collection_identifier	text	False	Unique persistent id (URI) of the collection the vocabulary term belongs to
member_label	varchar(250)	True	Preferred label of the vocabulary member (e.g. "poorly drained"). This is what is normally stored in the look-up table.
member_identifier	text	True	Unique external persistent identifier (URI) of the vocabulary term
member_notation	varchar(50)	False	Code/abbreviation for the vocabulary member
member_description	text	False	Description of the vocabulary term
broader_term_id	integer	False	The id of the broader term in a vocabulary hierarchy, if present
member_alias	varchar(250)	False	An alternative member label

## Units of Measure

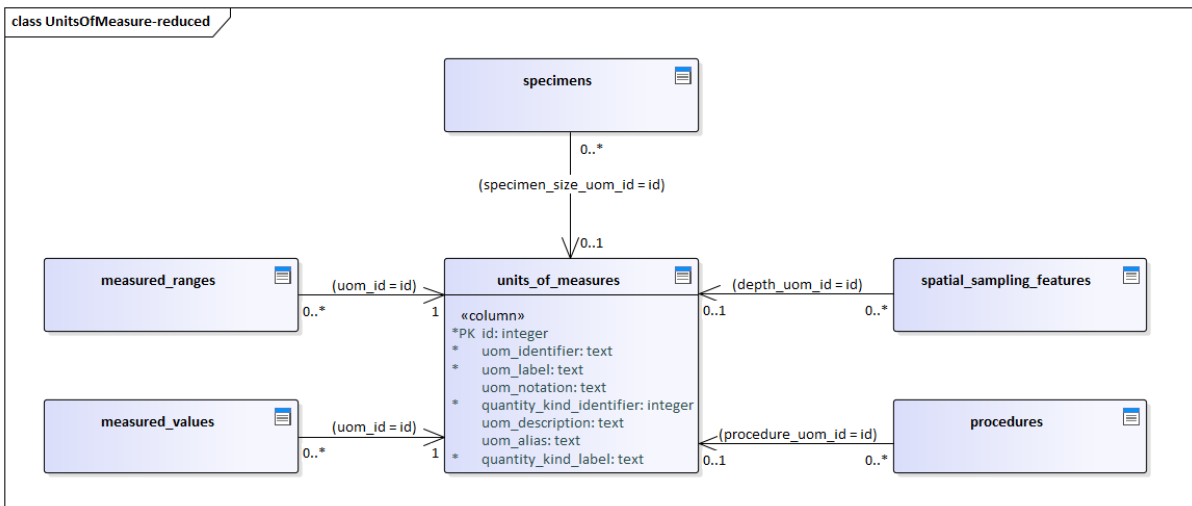


Figure B-11. Diagram: The 'units\_of\_measures' table and the associated tables that use it. Columns of related tables and associated relationships are not shown.

### 'units\_of\_measures' table description

The units of measure is based on the Quantities, Units, Dimensions and Types Ontology (QUDT). It allows specifying the label (e.g. "metre"), its abbreviation (e.g. "m"), alternative labels (e.g. "meter"), and the kind of measure the unit of measure relates to (e.g. "Length"). The table also caters for the identifiers (URIs) for the unit of measure and its quantity kind.

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	Database column id
uom_identifier	text	True	URI of the unit of measure term. From Linked Data Registry = rdf:about
uom_label	text	True	The text label of the UoM term. From Linked Data Registry = rdfs:label
uom_notation	text	False	The abbreviation for the UoM term. From Linked Data Registry = qudt:abbreviation or qudt:symbol or skos:notation

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
quantity_kind_identifier	integer	True	<p>The URI for the type of quantity that the UoM is (e.g. 'metre' is a 'Length' 'QuantityKind')  From Linked Data Registry = <a href="http://registry.it.csiro.au/def/qudt/1.1/qudt-unit#qudt:quantityKind">http://registry.it.csiro.au/def/qudt/1.1/qudt-unit#qudt:quantityKind</a>  <a href="http://registry.it.csiro.au/def/environment/unit#qudt:quantityKind">http://registry.it.csiro.au/def/environment/unit#qudt:quantityKind</a></p> <p>Because QuantityKinds are hierarchical a Unit of Measure may have multiple QuantityKinds: e.g. '<b>Atoms per Litre</b>' is a unit of measure for QuantityKinds of '<b>Concentration</b>' as well as its specialization '<b>Amount of Substance Per Unit Volume</b>'.</p> <p>The database design only allows for a Unit of Measure to be assigned to a single QuantityKind, preferably the most specialized.</p>
uom_description	text	False	<p>Text to describe the UoM.  From Linked Data Registry = qudt:description or dct:description or skos:definition</p>
uom_alias	text	False	<p>An alternative label or spelling of the UoM term (e.g. 'metre' rather than 'meter')  From Linked Data Registry = skos:altLabel</p>
quantity_kind_label	text	True	<p>The label of the QuantityKind (e.g. "Length").  From Linked Data Registry = <a href="http://registry.it.csiro.au/def/qudt/1.1/qudt-quantity#rdfs:label">http://registry.it.csiro.au/def/qudt/1.1/qudt-quantity#rdfs:label</a></p>

## Feature Names

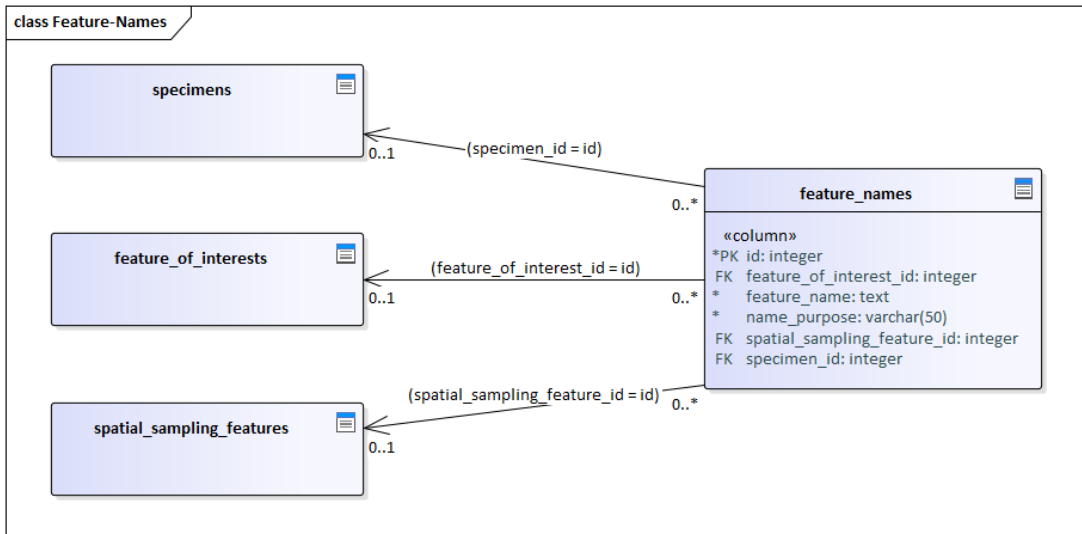


Figure B-12. Diagram: The feature\_names' table allows specifying multiple names (identity) to individual features, whether they are features of interest, specimens or spatial sampling features. Columns of related tables and associated relationships are not shown.

### 'feature\_names' table description

Allows for distinguishing multiple names of any single feature

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
feature_of_interest_id	integer	False	The identity of the feature of interest that the name applies to
feature_name	text	True	The name of the feature (e.g. "Lismore Grey Clays" and/or "33", or "Burt's Paddock", or "Specimen Barcode 222222", "Borehole M124"etc.)
name_purpose	varchar(50)	True	Allows distinguishing the reason for multiple names
spatial_sampling_feature_id	integer	False	The identity of the spatial sampling feature that the name applies to
specimen_id	integer	False	The identity of the specimen that the name applies to



## Reports

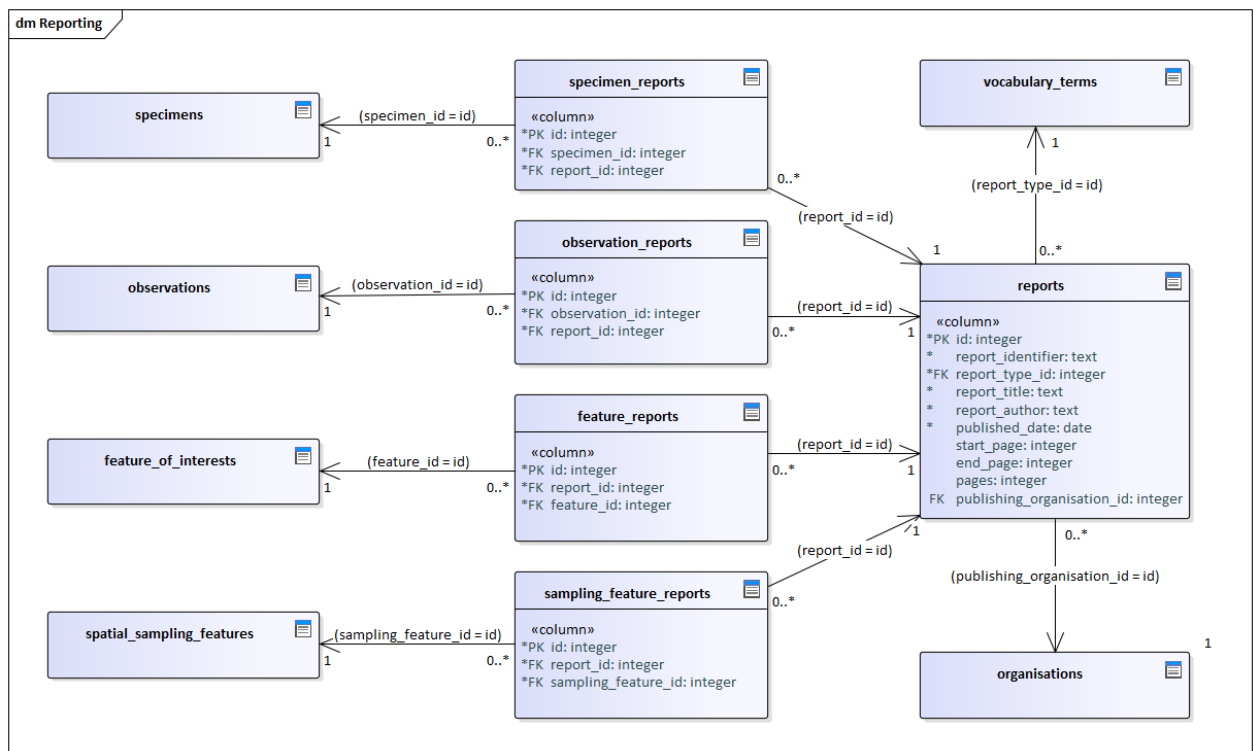


Figure B-13. Diagram: The reporting tables used for reporting the data for the specimens, observations, features of interest and spatial sampling features. Columns of related tables and associated relationships are not shown.

### ‘reports’ table description

Stores information about individual reports

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
report_identifier	text	True	The persistent identifier of the report. Will be a URI (doi, handle, URL, etc).
report_type_id	integer	True	A link to a vocabulary term describing the type of report (annual, technical, unpublished, project, thesis, etc.).
report_title	text	True	The title of the report
report_author	text	True	The author of the report
published_date	date	True	The date or year that the report was published, printed or finalised
start_page	integer	False	The first page of the report or sub-section
end_page	integer	False	The last page number of the report or subsection

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
pages	integer	False	The number of pages in the report or subsection (= end_page-start_page?)
publishing_organisation_id	integer	False	A link to the organisation that published the report

### **‘specimen\_reports’ table description**

Allows multiple reports to be associated with multiple specimens

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
specimen_id	integer	True	The specimen associated with the report
report_id	integer	True	A link to the report

### **‘observation\_reports’ table description**

Allows multiple reports to be associated with multiple observations

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
observation_id	integer	True	The observation associated with the report
report_id	integer	True	A link to the report

### **‘feature\_reports’ table description**

Allows multiple reports to be associated with the multiple features of interest

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
report_id	integer	True	A link to the report
feature_id	integer	True	The feature of interest associated with the report

### **‘sampling\_feature\_reports’ table description**

Allows multiple reports to be associated with multiple spatial sampling features

COLUMN NAME	DATATYPE	NOT NULL	COMMENTS
id	integer	True	The database column id
report_id	integer	True	A link to the report
sampling_feature_id	integer	True	The spatial sampling feature associated with the report

## APPENDIX C DATA ENTRY HEADERS

***Note: this section has significant contributions by Bruce Simons and Angela Neyland***

Each data set loaded into the CeRDI Observations System (the aggregator) requires a certain level of detail associated with it so that metadata can be auto-generated and attached to the datasets. The metadata is the key component in making the data findable and sets the access control rules for data accessibility. Metadata also helps considerably with interoperability and reusability.

At present there are two methods being considered in the VAS project to auto-harvest the required information for metadata when the data custodian self-loads soil data into the observation system. The question is whether the self-serve user interface should comprise a series of questions, or a form. While the questionnaire can be more comprehensible and unambiguous, some users prefer a template, since answering a series of questions may be regarded as tedious. The forms take a minimalist approach but can be easily duplicated for multiple dataset loading. While lots of metadata would be good, everyone finds it tedious to complete. Achieving the balance between what researchers want, and what data custodians are prepared to do, is delicate.

The current system in place requires the custodian to answer a series of questions (see screen shots on the following pages).

This may be supplemented with one or more downloadable templates which would be completed by the data custodian and loaded with their dataset. The templates would have details of licensing and access as well as details of the actual dataset and data being loaded.

These templates are untested at present but may have the advantage of being more suited to self-serve data mapping, since the aim is for the data provisioner to do the mapping as it gets semi-auto loaded. The questionnaire approach may not achieve that.

Although still an experimental work-in-progress, the following pages outline the information currently being harvested.

## SELF-SERVE SYSTEM QUESTIONNAIRE



VISUALISING AUSTRALASIA'S SOILS

### HOW TO SUBMIT A DATASET

You need to register and have an VAS user account to upload data. You do not need any specialised software, although data formatted to fit one of the templates makes it much easier to integrate in to the system.

You can upload data from a desktop or mobile device, such as a smart phone or iPad.

**Before you load data please make sure that:**

1. you have read the information about defining a dataset (below)
2. the data is formatted into one or more structured datasets (e.g. a spreadsheet or table)
3. you know the ownership and privacy settings for each dataset
4. you know the contact details, project name, funding sources, etc. (the metadata) for each dataset

**Step 1** Defining what you are submitting

Before we can begin, the first and most important consideration is to define what it is that you are about to provide. This submission process is designed for collections of soil data that can be considered as a data set. You may decide to combine or split data based on this self-assessment.

**Important:** Datasets should be provided one at a time. Do not submit a 'data dump' of many files from many different projects or sources without reading the following information.

**?** How to decide if your data is one or many datasets

Step through each screen filling out as much detail about the dataset you are submitting as you are able. The final step involves selecting the file containing the dataset to be submitted.

\* Indicates Mandatory fields

^

▶ Begin

Figure C-1. The opening screen of the self-serve data loading system.

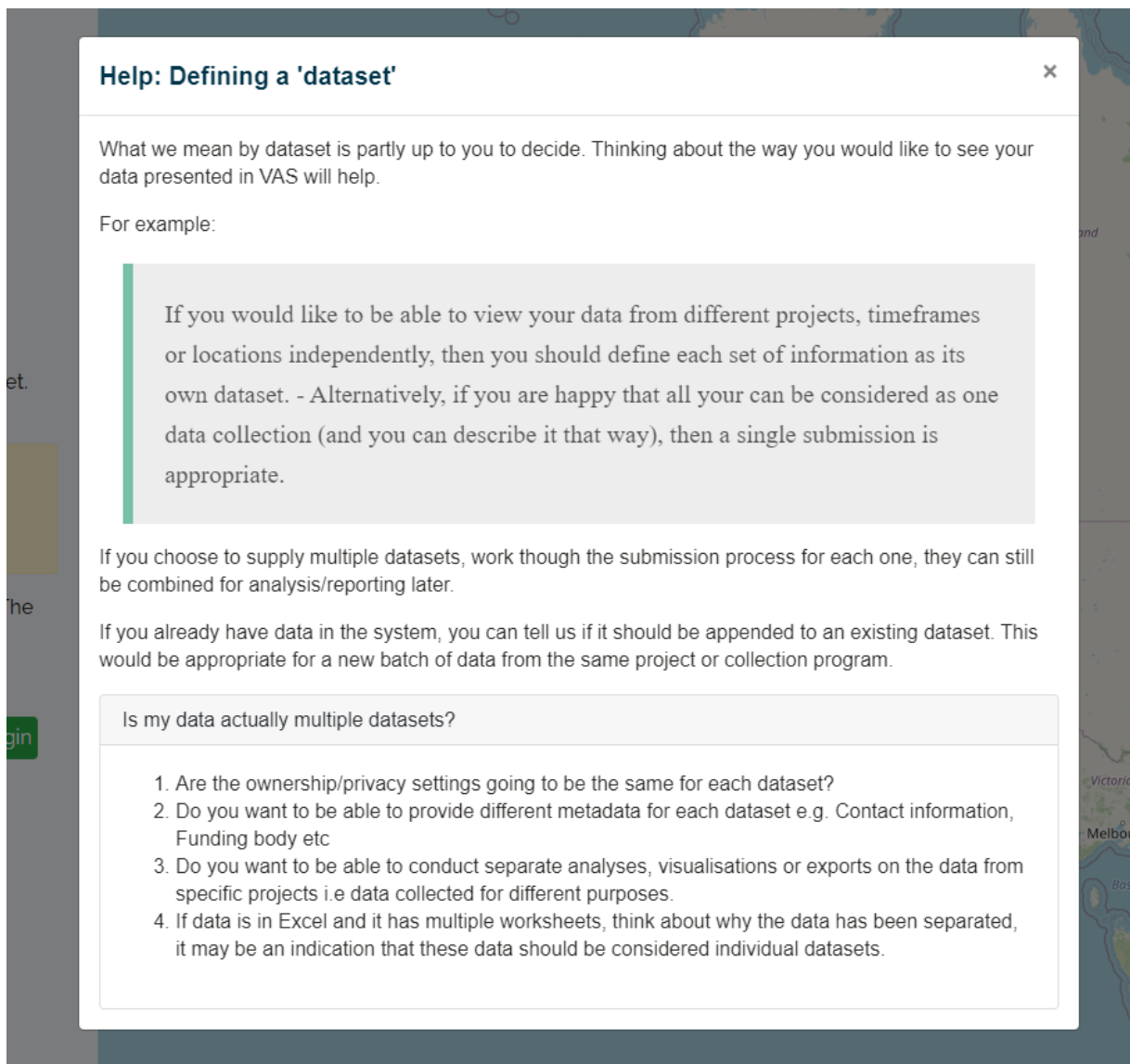


Figure C-2. The help for the question on the opening screen "How to decide if your data is one of many datasets".

🗨 ✕

### Step 2 Data file checklist

The next step in submitting a data set is to produce a file suitable for loading into VAS. If you manage your data in a spreadsheet, you may need to review how it is structured (column headings etc) and ensure the minimum content requirements are met using the checklist below. Alternatively you can copy your data into one of the templates provided.

- Is the data well-structured into consistent columns (with clear names) and rows of data.
- Have you included the sample dates/depths for each row of data
- If Laboratory analysis is included, have you included details of the Lab.
- If there is data from more than one Laboratory, has this been indicated clearly within the data
- Does the data include GPS/coordinates, or other locational data (KML/Shapefile).  
▲ Data without spatial (location) information cannot be processed.
- Is there a clear link between the results data and its spatial representation?
- Are the analyses/results columns well described and unambiguous?  
It is important that we are able to record the thing being measured e.g. Phosphorus Concentration, the procedure by which it was determined e.g. Bicarbonate extractable P (Colwell) and the units of measure e.g. mg/kg
- If lab analysis data, has it been modified from it's original format?  
If you have it in digital form (and depending on the format) sometimes it is easier for us to process data as supplied by the lab

Laboratory name(s)  
📌 Specifying the lab(s) assists us in confirming exact methods/procedures e.g. APAL, Nutrient Advantage

Sampling method? ? Land use? ?

Is the spatial data in a standard projection?  
📌 e.g. lat/lon; easting/northing; WGS84

Yes  No

Please provide the projection (if known) ?



Is this a complete dataset (or a subset / sample data)? ?

Complete  Sample/Subset

⬆

⏪ Back
⏩ Continue

Figure C-3. The second step screen of the self-serve data loading system.

### Step 3 Project/Collection Information

Is this data closely related to a previously supplied dataset? Only answer YES if you would like this data to be attached to that collection (or project)

**i** e.g. Under the same project, additional data collection, resampling of same sites

Yes
  No

**\*** Name of dataset **?**

Short name

**i** Abbreviation or Acronym

**\*** Description **?**

**Temporal coverage (the time period that this dataset covers)**

**\*** From start date **?**

To end date **?**

Is there a project report you could attach or link to?

**i** Provide link in comments below, include attachments in file upload (final stage)

Yes
  No

What was the nature of the project (if applicable)? **?**

Funding body (if applicable)

Any other comments or notes **?**

Figure C-4. The third step screen of the self-serve data loading system.



**Step 4 Contact Information**

Please provide contact details for the person most closely associated with this dataset. This should be the person who is best placed to assist with questions about the data provided.

\* Contact person ⓘ

Role ⓘ

\* Email ⓘ

Phone ⓘ

⬆

⏪ Back ⏩ Continue

Figure C-5. The fourth step screen of the self-serve data loading system.

**Step 5 Data file**

Finally. This is the last step where you select the relevant file or files from your desktop or device to supply to us. Please ensure you have reviewed the data checklist in step 2 and the file(s) reflect those requirements.

\* Choose File ⓘ No file chosen

⬆

⏪ Back ⏩ Submit

Figure C-6. The final step screen of the self-serve data loading system.

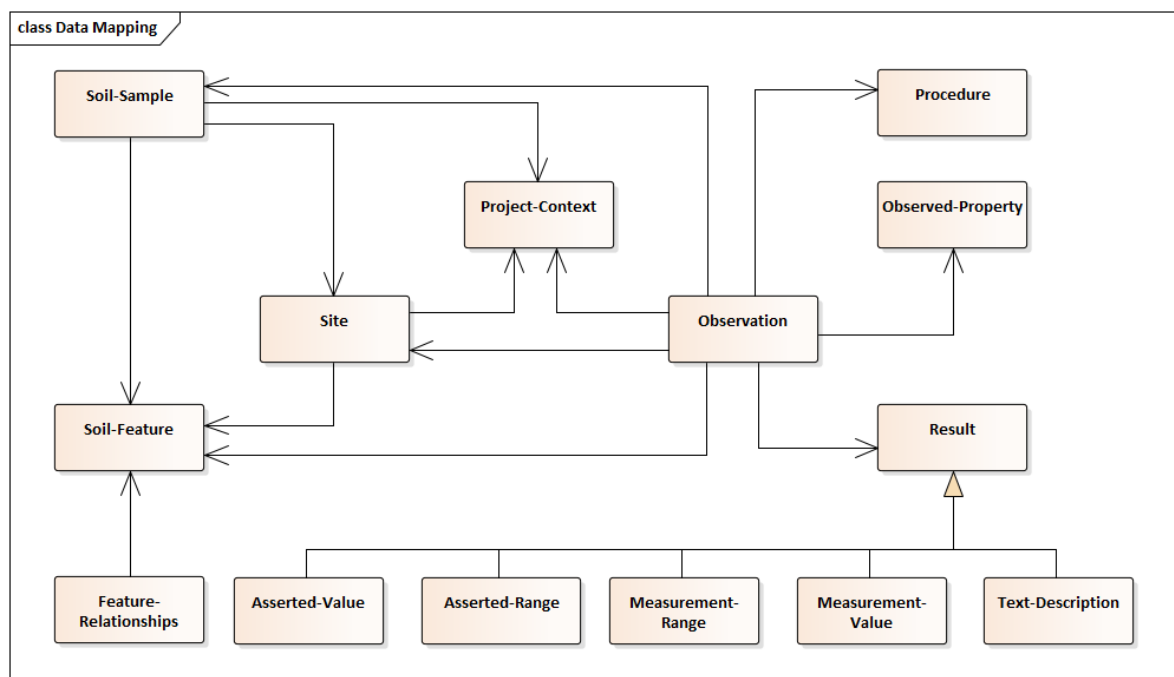
# APPENDIX D DATA AND VOCABULARY MAPPING

*Note: this section has significant contributions by Megan Wong, Bruce Simons and Angela Neyland*

## DATA STRUCTURE MAPPING

The challenge faced by VAS was to take varying data content and formats from different data providers and make it available to potential users in a standard format, with standard content, via a standard mechanism. That is, to make it more FAIR. In order to simplify and standardise the data discovery and delivery process, the database used to store the VAS data is based on the Observations & Measurements (O&M) standard. However, a consequence is that the data import process, the data structure and content mapping, becomes more complicated.

A summary of the database concepts that the data is mapped to is shown below. While the classes (boxes) in the diagram represent the concepts captured in VAS, the actual table names used in the VAS database are more closely aligned with O&M features. Each Observation uses a Procedure to generate a Result of some Observable Property. The observation results may be a measurement value (a numeric), a range of measurement values (a minimum and maximum numeric value), an asserted vocabulary term value, an asserted range of vocabulary terms, or a non-vocabulary controlled text description. The Observation occurs at a Site/Plot/Pit/Paddock (the Spatial Sampling Feature), where the Sample is taken from and where the Soil Feature exists in the real world. The Observation may be on a Soil Sample (specimen), or a Soil Feature, such as a Soil Layer, Soil Horizon, Soil Profile or Soil Body. These Soil Features may be related, such as Soil Horizons being part of a Soil Profile, which in turn is representative of a Soil Body. The human constructs related to the data, the Observation, Sample and Site, happen within some Project Context: the project and the organisations and people involved.



*Insert caption*

As part of the standardisation requirements, the values of the result asserted terms, units of measure, observable properties and procedures are all from controlled vocabularies.

The data structure mapping activity requires mapping the components of the data provided in flat spreadsheets to the equivalent tables in the relational database structure.

## DATA CONTENT REQUIREMENTS

There is a minimum data content standard required to include data contributors' soil datasets in VAS. More information about the data, the gold standard, allows potential data users a higher resolution of search and discovery, and helps determine whether the results are fit for purpose.

Minimum data content standards are:

- name of organisation and/or project supplying and/or responsible for the data
- data to be provided in an accessible format, such as an excel worksheet with a single header row without logos or proprietary formatting
- spatial coverage of dataset (provided as GPS coordinates or standard spatial data format, such as a \*.shp file)
- temporal information of each observation
- soil data values contain an identifiable observable property, procedure, and result, and a unit of measure if the result is a measurement
- unique identification (naming) of the site of the observations
- unique identification (naming) of the samples if there are multiple samples taken at the same location
- which results are associated with which samples?

Gold data content standards also includes the following:

- the name of the laboratory/laboratories used to collect the data (and where more than one is used clearly associating this with specific parts of the dataset)
- identifying the standard laboratory specific procedures (such as by "Green Book" name or code) used for each observation
- providing information about other organisations and their roles in the project associated with the dataset
- title of the collection and/or name of the project for which the data was originally collected
- who can access the data and under what conditions?
- what are the licencing conditions for the data (such as CC-BY)
- what was the method used to sample the soil?
- what was the land at the soil site being used for at the time of sampling?
- laboratory identifier (ID) numbers/bar codes of samples
- where the soil sample is stored if preserved.

Where available, additional data (metadata) about the project that generated the soil dataset improved the useability of the data, such as:

- dataset citations
- funding information
- keywords (such as ANZSRC-FOR codes).

## DATA MAPPING CHALLENGES

### VARIETY OF THE DATA

The variety of data, the different structures and formats in which it was provided, and the varying degrees of documentation made understanding the datasets a challenge for VAS. For example, different types of results were provided from different soil science domains (chemical, physical, biological), using different sampling regimes. Data was provided in a range of formats (such as .pdf and excel) and structures (different column layouts), and the results described in different ways (different column headers).

**Formats:** In most cases data was provided in Microsoft Excel spreadsheets, either from the grower groups or the soil testing laboratories. Occasionally, a combination of datasets was merged together into the one Excel spreadsheet. Rarely was the soil data extracted from a data provider's organisation database or from text based (such as \*.pdf) reports. At times, the data was provided in Excel formats with multiple layers of horizontal and vertical headers or with multiple separate tables one under the other in the same Excel spreadsheet with separate headers. Reformatting of the provided data was usually necessary.

**Data Domains:** Different types of information from different soil domains that needed to be mapped into the VAS database included:

- Measurements made either in the field or the laboratory

- The majority of datasets included:
  - soil physicochemical properties (such as pH, cation exchange capacity), and dry chemistry (such as mid infrared spectroscopy (MIR))
  - soil physical observations such as texture and bulk density
- A few datasets included:
  - Soil biology observations (such as microbial biomass)
  - Soil profile observations (such as horizon classifications and soil profile descriptions)
- Sometimes treatments were provided, such as stubble retention and manure application rates, but there is currently no strategy in place to store farm treatment data in VAS.

**Procedures:** Identifying and mapping the procedures used to generate the observation results was challenging. For soil physicochemical properties, in most cases the result column headers could be identified as a standard citable published method, such as from the "Green Book" (Soil Chemical Methods, Australasia, scma) or specific laboratory adaptations. In some cases, field observations could be associated with a specific procedure, such as in-field pH and 'fizz test'. However, there were challenges mapping to standard procedure vocabularies when the procedures were 'in-house' laboratory specific or unspecified. Additionally, the same property (such as 'Total Carbon concentration') could be provided multiple times in the one dataset, using different, but unspecified, procedures.

In some cases the procedures used were a calculations based on the results of some other procedure. The most common were Cation Exchange Capacity and Effective Cation Exchange Capacity calculations, ratios (mostly of Nitrogen, Phosphorus, Carbon and Potassium). Sometimes, conversion factors were applied (such as Organic Matter to Total Organic Matter conversion factor).

**Sampling Regimes:** Different experimental design, factors or treatments, and sampling regimes needed to be identified and mapped to the database structure. Measurements may have been made at the farm paddock (mostly), block, plot and or subplot levels, with replication of observations also made at these different levels. Repeated measurements may be made over time, at the same or a nearby location. Most measurements have been made at specified depths (soil layers), such as 0-10 cm, 10-20 cm, etc. However, not all results used these depths and no standard set of depth ranges was used. Some soil profile observations were made on pedogenic-based depths (soil horizons). Details of the sampling method, such as disturbed, undisturbed, core diameter, were not provided.

Location was usually provided as point locations, and occasionally in a spatial format (such as via a shape file), and rarely as both.

Further details of the sampling regime were most often not specified, for example if sampling was randomised or taken along a transect, and rarely was it georeferenced.

**Samples:** Little information about the soil samples/specimens were provided. Laboratory accession codes were sometimes provided, and if so they were included in the specimen database data. Details of any sub-sampling in the laboratory, including any pre-treatment and storage of samples prior to measurement (such as subsample air drying, oven drying or freezing), were not provided within the datasets.

**Data Vintages:** The age of the data provided included legacy data, recent data and ongoing sampling campaigns.

**Data Types:** Observation results were provided in a number of different formats, depending on the procedures used and observed property. These included:

- Single value measurements (numeric).
- Range of measurements (numeric).
- Terms, most often from a National Standard or Laboratory schema (for example result terms from the Australian Land and Soil Field Handbook (the "Yellow book")).
- Ranges of terms.
- Multiple results reported as a single result.
- Suggested range/interpretation values provided by labs and/or grower groups (e.g. look-up codes).
- Free-text descriptions.
- Mixed result types for the one measured variable (range of number and terms).
- Missing data.
- Above and below detection limits (>, ≥, <, ≤, adt, bdt).
- Results as lab specific indices, such as indices for soil biological health and stress.

**Units of Measure:** In most cases where the result was a measurement or range of measurements, the units of measure were provided, and were mapped to a standard units of measure vocabulary (QUDT; <http://www.qudt.org/>). The same measurements may be reported with different units of measure (for example reporting as both % and 'parts per million' - ppm), in which case the custodian specified unit of measure was associated with the result. As most methods can use multiple units of measure, if the units of measure are not provided for each procedure, the data cannot be imported.

## RELATIONSHIPS

Mapping the relationships between observations within and between datasets is a key component of the data mapping process to help ensure that data can be compared over time and space, both within and across datasets. Most often datasets have complex relationships that include both spatial and temporal relationships, which need to be mapped. These may be observations made at multiple depths in one location and observed again at those depths at a later time (related observations). Or they may be different sample and sample types being related to each other in space, such as replicates, or soil samples at depth (0 - 10 cm) taken on nearby soil profiles (related features). Or relationships between procedures that measure the same Observable Property.

The following relationships have been mapped:

**Temporal:** Multiple observations that have been taken on the same feature at the same place, or nearby places, at different times. Sometimes this data is provided within a single spreadsheet and the relationship specified (such as via the site IDs). In this case the relationship has been captured in the database. At other times it may be provided separately and at different times, and there is often not a data custodian provided relationship between the two sets of data. In this case, database analysis of depth, spatial and context data was used to generate the temporal relationships.

An aspect of temporal relationships not dealt with by VAS is capturing the information about what happened at the site, such as farm treatment data, between the two sets of observations.

**Spatial:** The relationships the different depths (soil layers) at the same location have to each other and to the overall soil body at the site, the spatial relationships between replicates, and the relationships between different spatial sampling features (for example soil layers 1 - 10 cm that were sampled from nearby soils). Spatial and Temporal relationships were mapped to standard relationship vocabulary terms.

**Categorical:** Observations taken on the same named or classified soil body/unit.

**Vocabulary:** Vocabulary terms may have hierarchical relationships (broader terms). In addition, they are contained within Collections.

## QUALITY AND RESOLUTION

No attempt to assess the accuracy of the custodians' data was made prior to uploading into VAS. However, user evaluations of the usefulness of the data can be made based on the quality and resolution of the information provided.

Where sufficient information was provided, the data could be mapped at a higher resolution. Although obvious for spatial data, this also applied to other data. For example, when the exact procedure used for to obtain the soil observations was known it allowed the results to be fully understood and compared to results from different datasets. When the procedure details were not provided the data had to be mapped to generic procedure standards, which lowers the resolution and usability of the data for comparison and analysis. Although no formal ranking was applied to each dataset, a sense of the usefulness of the datasets can be found according to the level of standardisation they provided. This is particularly the case for the laboratory procedures used. For example, with procedures used, from most useful to least useful:

1. Data from a soil laboratory or field results that use nationally and/or internationally recognized soil observation procedure standards (e.g. such as the Australian soil domain standard "Green" and "Yellow" book codes).
2. Data that uses a citable, publicly available, published on-line procedure.
3. Data that uses laboratory specific procedures, and these procedures are published on-line (e.g. a laboratory manual or website).
4. Data with undocumented laboratory procedures.

Challenges for maximising the use of the participants data in VAS have included:

- The laboratory used and/or the procedure information used was unknown, usually provided as a column header, was insufficient or not included with the results.
- Results were provided from multiple laboratories without specifying which results come from which laboratories.
- Laboratories change the procedures they use over time and previous procedure documentation may not be available. Assigning the laboratories current procedures to legacy data may not be appropriate.
- Soil laboratories may have multiple procedures available for the measurement of the same soil observed property, and if the exact procedure was not provided, knowing the name of the laboratory is insufficient to identify which procedure was used.
- Where the same observed property (such as soil carbon) was measured more than once and the different procedures not identified, multiple sets of results could not be imported.
- Inconsistent or *ad hoc* calculations make it difficult to map against a standard method. It is not always clear whether calculations come from the laboratory or from the soil data

custodian, again making it difficult to map results against a standard method. For example, different datasets contain different in ways that CEC and ECEC are calculated.

## CONTROLLED-VOCABULARIES

Shared terminology is key to accurate communication and enabler for data integration ([Cox et al. 2020](#)). Many domains (including Soil Science/Research) and organisations use curated lists of terms to describe their information including data in databases.

When there is a process for managing them, they are termed 'controlled-vocabularies'. The use of controlled-vocabularies increases the re-use potential by giving the data meaning. Controlled-vocabularies can include lists, synonyms, taxonomies and thesauri. Traditionally managed as, for example, books or lists in spreadsheets stored on a computer hard drive, controlled-vocabularies provide semantic meaning using web resources that are persistent, readable and understandable by machines, via Uniform Resource Identifiers (URIs). These web resources are made more understandable by machines through the application of standard knowledge representation languages (such as Simple Knowledge Organisation System - SKOS) (For an introduction see [Gudivada et al. 2018](#) and [Cox et al. 2020](#)).

In addition to being readily accessible, persistent and machine readable, controlled-vocabularies should be well-governed by an appropriate expert authority and licenced for re-use. This information should be contained in the metadata of the controlled-vocabulary. The vocabulary terms should also be easily understandable by people (as well as machines), for example through providing clear definitions to help the end-user interpret and potentially re-use the data. That is, controlled-vocabularies themselves should be FAIR (For guidelines see [\\_2020](#))

Controlled-vocabularies are used to describe the data in VAS. This helps to makes VAS data:

- More easily shared and understood among end-users, including scientists, as ambiguity is reduced.
- More easily integrated with other datasets (e.g. for benchmarking).
- Able to be harmonised.
- Searchable in an efficient way, using different views and contexts.
- Of better quality, as quality checks can be run easily.

Some examples of the types of controlled vocabulary terms being used to describe data in VAS:

- The procedure used to get the result (e.g. 'Green Book '[10C1](#)').
- The unit of measure (['mg/kg'](#)).
- The property that the procedure is measuring (the Observable Property) (e.g. '[Sulfur Concentration](#)').
- The substance that the procedure is measuring (e.g. ammonium, '[Sulfur](#)').
- The spatial features and features that the measurement is being made on, and the relationships between them (e.g. '[Soil Horizon](#)' is '[PartOf](#)' a '[Soil Profile](#)').
- Sampling method and sample type.
- Role codes for data contributors (e.g. '[Owner](#)').

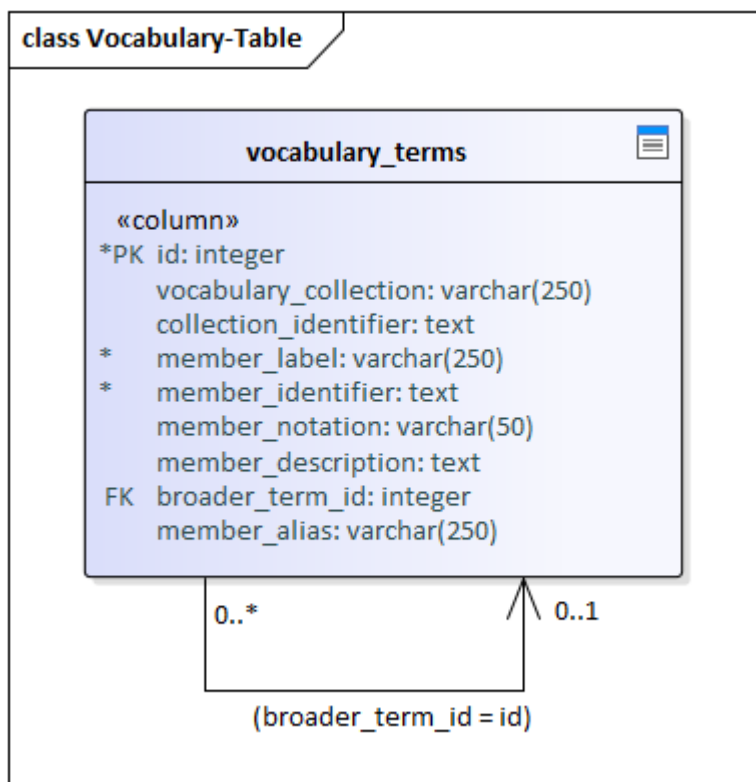


Some of the challenges in applying controlled vocabularies to VAS data have been:

- There are limited available national or international standardised vocabulary for terms used in the data mapping, particularly for the soil observed properties and procedures. Where possible VAS has worked with other appropriate agencies to establish accessible and well-governed vocabularies to fill the gaps.
- Many of the concepts used in the data, particularly many of the procedures, were not 'standard' and not documented. Where possible vocabularies were created for these non-standard concepts, but on-going governance of these is problematic. Ongoing management and governance of these terms should follow FAIR vocabulary guidelines.
- Discovering the most appropriate vocabulary and the ease of using these soil related vocabularies can be difficult. This is a challenge recognised at the international level and across multiple domains.
- Some generic vocabularies required were not available, such as soil data specific units of measure.
- Determining what is a 'standard' citable method or a laboratory 'in-house' method, and what is an extension of a method, or calculation was difficult. A follow-on challenge was deciding when it was appropriate to create and apply a new controlled vocabulary term and when to use a pre-existing one.
- Ongoing access and maintenance of controlled vocabularies requires governance of the vocabularies. A problem not unique to the soil domain is that most vocabularies have been established without determining the roles, rights and responsibilities to publish, maintain and govern the vocabularies. The various soil-related controlled vocabulary collections established to describe VAS data also require appropriate governance to be established.

In the VAS database the vocabulary terms are all contained within a single table, the `vocabulary_terms` table (see figure), rather than as separate look-up tables for each property requiring a vocabulary. This therefore requires identifying what property the term can be applied to (the `'vocabulary_collection'`) along with its external web-resource (the `'collection-identifier'`). Each term has a label (`'member_label'`), a reference to the term in an external vocabulary (`'member_identifier'`), an abbreviation or code for the term (`'member_notation'`), and an alternative name for the term (`'member_alias'`). Each term may be a member of a single hierarchy, identified by a link to the broader term (`'broader_term_id'`).

For performance and access reasons, the vocabulary table is a regularly updated 'cache' for terms that are actually managed in external, web-based controlled vocabularies. As such the properties in the table mimic the Simple Knowledge Organisation System (SKOS) specifications that many vocabularies are published as.



*Insert figure caption*

## IDENTITY

For each observation, sampling feature, soil sample, and soil feature, its identity and relationships (see 'Relationships' above) to other observations and features is crucial to using, integrating and understanding the results associated with it. Each feature may have multiple names (such as sample name and laboratory barcode, site name and paddock name, horizon depth and classifier, etc.), so in addition to storing the multiple names that may be associated with each feature, a persistent and unique identifier for each feature is required. When resolved (for example via a web service request) each identifier should return information about that resource.

Common identity challenges experienced have included:

- Inconsistent naming of sites/samples. Sometimes these are minor typographic errors, and sometimes similar locations are named in entirely different ways. This is particularly an issue if the spatial information is provided separately from the soil observation results and there is no sure way to associate the two. In this case, the dataset does not meet the minimum requirements for mapping. It was also a problem when temporally related datasets had inconsistent site names.
- The relationships were not consistent throughout a whole dataset, preventing automatic data upload.
- When two related (e.g. temporally) datasets were provided together the relationships outlined within the data were added during the data upload. When they were not provided together relating the separate datasets required spatial and vertical co-incidence (that is, identical 3D information associated with each site) and/or standardised site names across

the datasets. Where this was not available the relationships between the datasets were lost.

- Rarely was the identity of the soil sample (such as a barcode) provided with the set of its related observations. Laboratory accession or processing numbers were sometimes provided, which were assumed to belong to the physical sample on which the observation measurement was made.
- Information about the soil sample, such as the type of sample and whether it was pre-treated or retained, was not provided.
- Each dataset considered its features' identities in isolation and was usually only concerned with the set of observations made at a particular depth, at a particular site, at a particular time. However, when combined with datasets from other sources and times, determining unique and resolvable identities was required, often without any guidance provided by the dataset entities.

## **METADATA**

Custodian data was usually provided to VAS via a file (most often a data spreadsheet) and a verbal indication of its source. Where possible, the metadata ('data about the data') was inferred from the dataset content, file names and verbal communication. Although the absence of this information has not precluded data from being imported (it is a 'gold standard' requirement), this metadata is an important part of the FAIR specifications in order to have appropriate VAS data discovery and access, attribution, and maximise re-use potential ( by, for example, evaluating data quality).

Further information required but not usually provided, includes:

- Name of organisation(s) and their role(s) associated with and/or responsible for the data.
- Title of the dataset and/or name of the project that generated it.
- Access rights for the data.
- Licensing for the data.

# APPENDIX E VAS API

**Note: this section has significant contributions by Andrew MacLeod and Bruce Simons**

The VAS Application Programming Interface (API) provides interoperable access to Soil Data from VAS Partners and affiliates that has been loaded into the CeRDI Observations System.

This RESTfull API is documented using the OpenAPI Specification (OAS3). Previous versions of this specification were known by the swagger moniker and this terminology are still somewhat interchangeable.

*The OpenAPI Specification (OAS) defines a standard, language-agnostic interface to RESTful APIs which allows both humans and computers to discover and understand the capabilities of the service without access to source code, documentation, or through network traffic inspection. (Source: <https://swagger.io/specification/>)*

The most current (and machine readable) version of the VAS API docs is available at the following URL <https://app.swaggerhub.com/apis-docs/FedUniCeRDI/vas-soils-api/1.0#/>

## AUTHORISATION AND ACCESS

The majority of VAS datasets are non-public at the time of writing, authorised users are given access to specific datasets via an API key (or token) which is validated against the VAS permissions system. A guest token for accessing publicly available datasets is available.

An authentication header containing a valid API Key is required for many of the REST endpoints. The Authorization field in the HTTP header is used to pass the API key (or token).

## STRUCTURE AND SEMANTICS

The API Endpoints described here have been designed based on the patterns of the Sensor, Observation, Sample, and Actuator (SOSA) ontology <https://www.w3.org/TR/vocab-ssn/> but should not be considered compliant with that specification. Most payloads (content) are provided in JSON-LD syntax (by default) but this should be considered experimental from a linked data perspective. Some content is still semantically invalid or without a valid context/ontology.

## END POINT - DATASETS

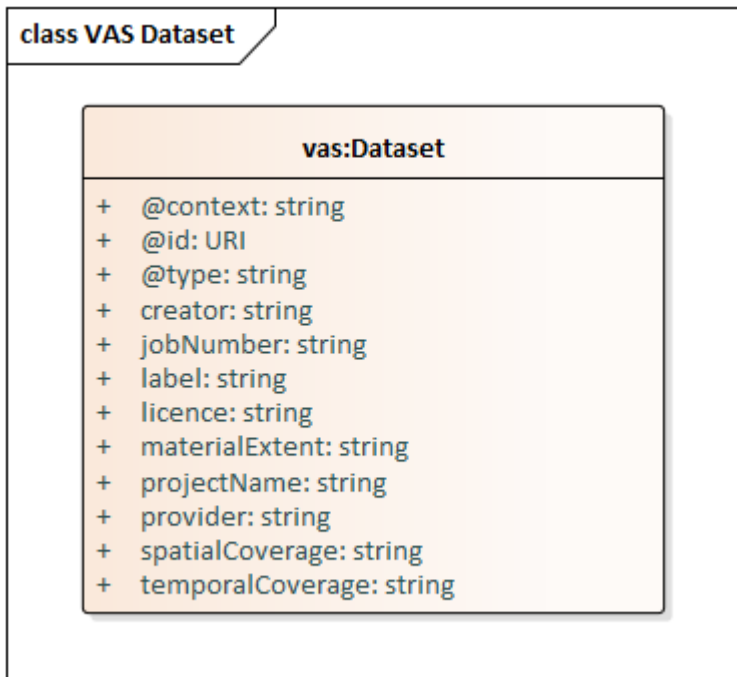
Provides metadata information for each dataset for the specified provider.

### Path:

/[provider]/data/schema/dataset/contextJobNumber

### Example request URL

<https://id.cerdi.edu.au/ccma/data/schema/dataset/000001>



## END POINT – OBSERVED PROPERTIES

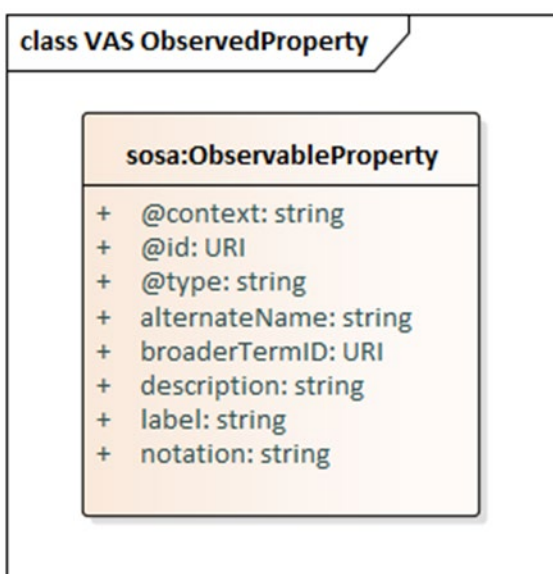
Provides a listing of Observed Properties (things that have been measured/observed) by this provider.

### Path:

/query/observedproperties/[provider]

### Example request URL

<https://id.cerdi.edu.au/query/observedproperties/ccma>



## END POINT – USED PROCEDURE

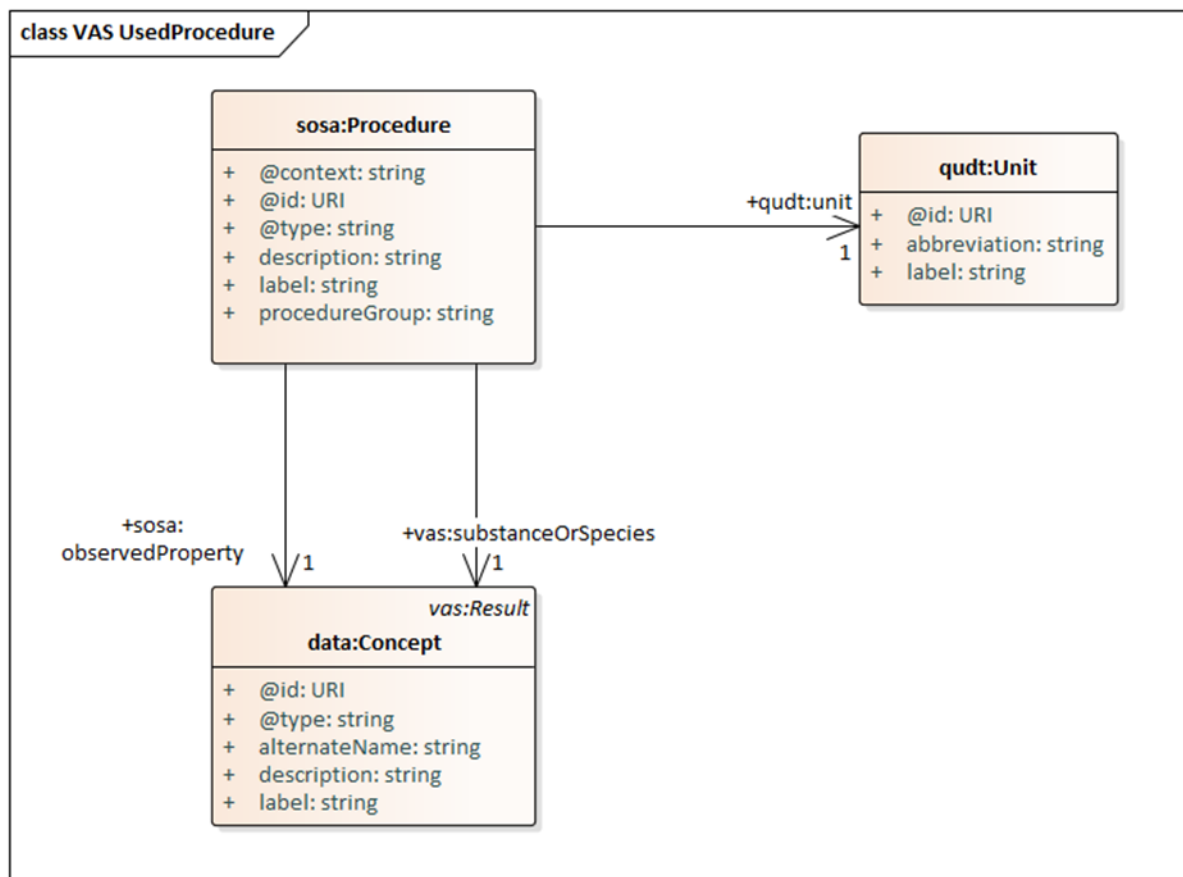
Provides a listing of Used Procedures (the way in which things have been measured/observed) for this provider.

### Path:

/query/procedures/[provider]

### Example request URL

<https://id.cerdi.edu.au/query/procedures/ccma?limit=20>



## END POINT – FEATURE

Provides a listing of real-world features (things against which observations have been made) for this provider. e.g Soil Body, Soil Layer, Soil pit.

### Paths:

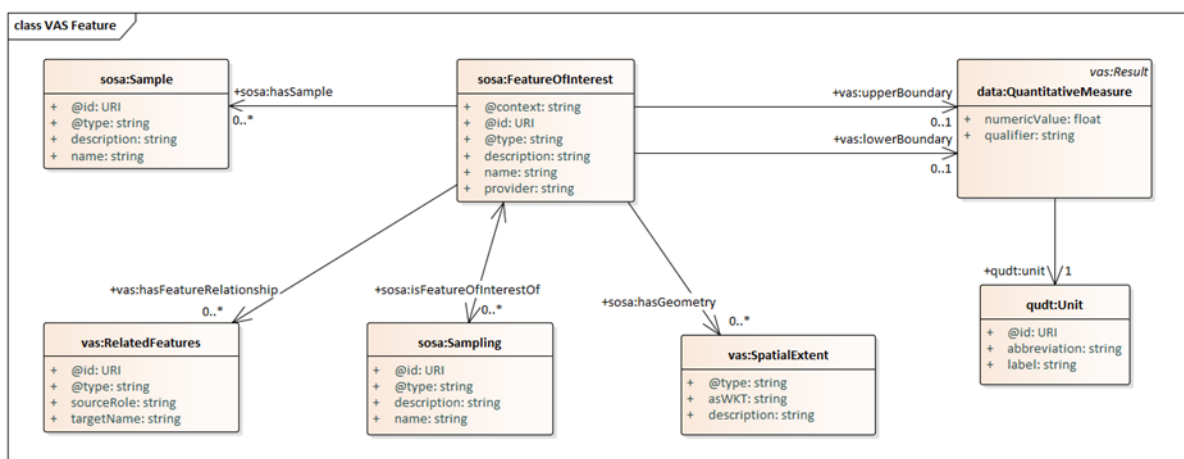
/[provider]/data/sosa/feature/

/[provider]/data/sosa/feature/[featured]

### Example request URLs

<https://id.cerdi.edu.au/ccma/data/sosa/feature/?limit=20>

<https://id.cerdi.edu.au/ccma/data/sosa/feature/ccma.soil.feature.1>



## END POINT – OBSERVATION

The core of the system returns matching observations including the result, time, property, procedure. feature of interest etc.

### Path:

/[[provider]]/data/sosa/observation

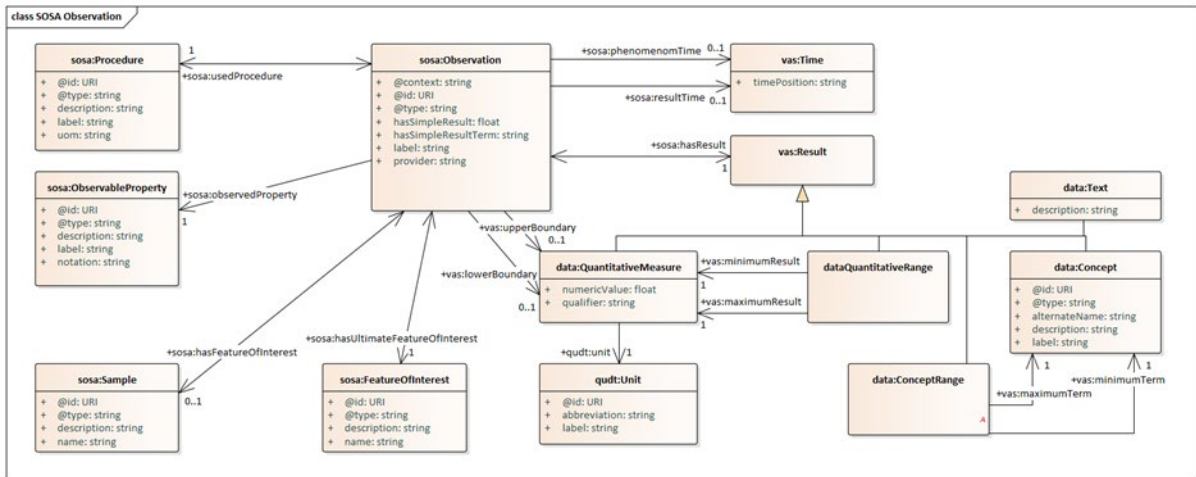
### Example request URL

<https://id.cerdi.edu.au/ccma/data/sosa/observation/>

### Query String (Filter in ODATA syntax)

This query returns all observations for Total Organic Carbon using the Dumas high-temperature combustion procedure where the result in tonnes/hectare is greater than 30 and was observed after 1 January 2015.

[https://id.cerdi.edu.au/ccma/data/sosa/observation?usedProcedure->>"@id"=eq.http://www.anzsoil.org/def/au/scma/6B2a&observedProperty->>"@id"=eq.http://environment.data.gov.au/def/property/carbon\\_organic\\_concentration&hasSimpleResult=gt.30&resultTime->>timePosition=gte.2015-01-01&limit=20](https://id.cerdi.edu.au/ccma/data/sosa/observation?usedProcedure->>)



## END POINT – SAMPLING

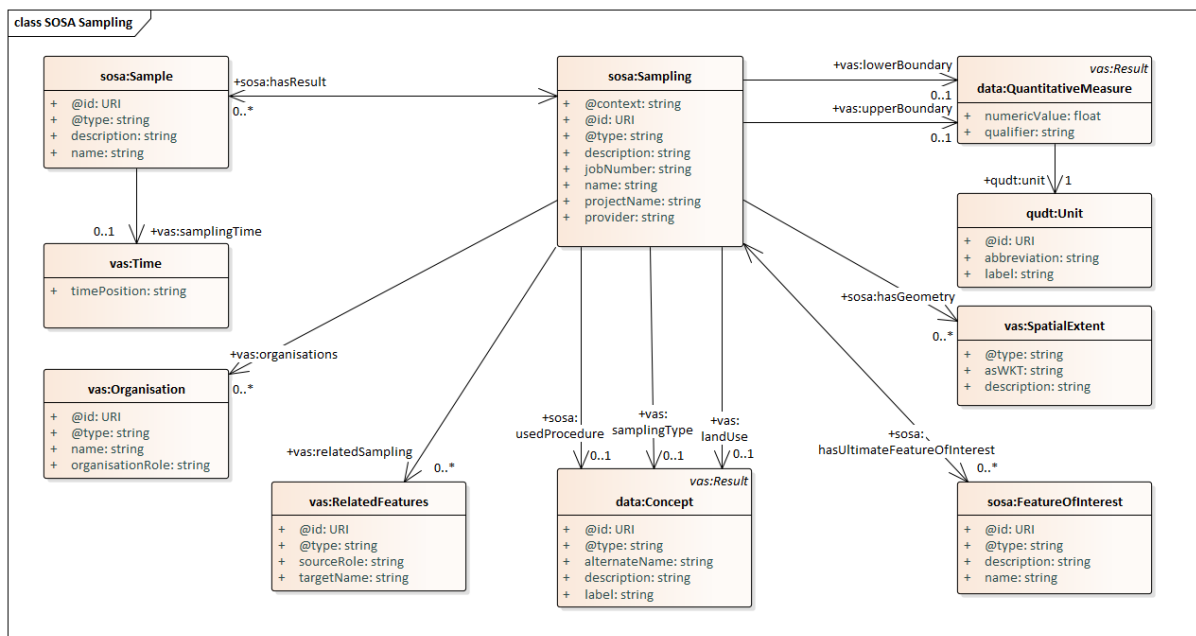
Provides a listing of Sampling features, (details of the of how and where the real world feature was sampled).

**Path:**

/[provider]/data/sosa/sampling

**Example request URL**

<https://id.cerdi.edu.au/ccma/data/sosa/sampling/>





## END POINT – SAMPLE (NOT IMPLEMENTED)

