



Development and Governance of FAIR Thresholds for a Data Federation

PRACTICE PAPER

MEGAN WONG

KERRY LEVETT

ASHLIN LEE

PAUL BOX

BRUCE SIMONS

RAKESH DAVID

ANDREW MACLEOD

NICOLAS TAYLOR

DEREK SCHNEIDER

HELEN THOMPSON

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

The FAIR (findable, accessible, interoperable, and re-usable) principles and practice recommendations provide high level guidance and recommendations that are not research-domain specific in nature. There remains a gap in practice at the data provider and domain scientist level demonstrating how the FAIR principles can be applied beyond a set of generalist guidelines to meet the needs of a specific domain community.

We present our insights developing FAIR thresholds in a domain specific context for self-governance by a community (agricultural research). 'Minimum thresholds' for FAIR data are required to align expectations for data delivered from providers' distributed data stores through a community-governed federation (the Agricultural Research Federation, AgReFed).

Data providers were supported to make data holdings more FAIR. There was a range of different FAIR starting points, organisational goals, and end user needs, solutions, and capabilities. This informed the distilling of a set of FAIR criteria ranging from 'Minimum thresholds' to 'Stretch targets'. These were operationalised through consensus into a framework for governance and implementation by the agricultural research domain community.

Improving the FAIR maturity of data took resourcing and incentive to do so, highlighting the challenge for data federations to generate value whilst reducing costs of participation. Our experience showed a role for supporting collective advocacy, relationship brokering, tailored support, and low-bar tooling access particularly across the areas of data structure, access and semantics that were challenging to domain researchers. Active democratic participation supported by a governance framework like AgReFed's will ensure participants have a say in how federations can deliver individual and collective benefits for members.

CORRESPONDING AUTHOR:

Megan Wong

Federation University, AU
mr.wong@federation.edu.au

KEYWORDS:

agriculture; AgReFed; FAIR data; community; governance; RM-ODP

TO CITE THIS ARTICLE:

Wong, M, Levett, K, Lee, A, Box, P, Simons, B, David, R, MacLeod, A, Taylor, N, Schneider, D and Thompson, H. 2022. Development and Governance of FAIR Thresholds for a Data Federation. *Data Science Journal*, 21: 13, pp. 1–12. DOI: <https://doi.org/10.5334/dsj-2022-013>

1. CONTEXT AND CONTRIBUTION

The agriculture data landscape is complex comprising of a range of data types, standards, repositories, stakeholder needs, and commercial interests, creating data silos and potential 'lock-ins' for consumers (Kenney, Serhan & Trystram 2020; Ingram et al. 2022). There is an urgent need to work toward clear, ethical, efficient agricultural data sharing practices (Jakku et al. 2019; Wiseman & Sanderson 2018) with improvements to discoverability, accessibility, interoperability, and quality of data across the value chain (Barry et al. 2017; Perrett et al. 2017; Sanderson, Reeson & Box 2017). A priority stakeholder question across the agri-tech sector is 'how do we create systems whereby people feel confident in entering and sharing data and in turn how to create systems to govern data for the benefit of all?' (Ingram et al. 2022: 6).

Agricultural data stakeholders span the public and private sector including farmers, traders, researchers, universities, consultants, and consumers. Their varied needs around data type, trustworthiness, timeliness, availability, and accuracy shape the many data capture, storage, delivery, and value-add products emerging across the public and private sector (Allemang & Bobbin 2016; Kenney, Serhan & Trystram 2020). Data providers require confidence in data infrastructure governance before they share their data including ethics of ownership, access, and control. Strong value propositions are also key. This helps grow participants via a 'network effect' increasing infrastructure value further (Chiles et al. 2021; Ingram et al. 2022; Sanderson, Reeson & Box 2017).

Offerings of the many data infrastructures vary and include data deposition for persistence, citation, publisher and funding requirements (see Datacite, 2022); increasing collaborative opportunities; regulatory compliance; on-farm operations; leveraging standardisation, quality assurance and quality control pipelines and specialist analysis capacity (e.g., Harper et al. 2018; Wicquart et al. 2022); running simulations through Virtual Research Environments (Knapen et al. 2020); cross domain data integration (Kruseman et al. 2020) and linking data and models to knowledge products and decision support tooling (Antle et al. 2017).

If the goal is to make data trusted, discoverable, and re-usable across the sector (Pearson et al. 2021; Ernst & Young 2019) then a single platform is unlikely to meet all (public, private, commercial) needs (Pearson et al. 2021; Ingram et al. 2022). Sector concerns include lock-ins and stifling innovation (Ingram et al. 2022). So, a grand challenge is how can data be discovered and interoperate between so many different databases and infrastructures? One solution is a decentralised federated approach where there is no single master data repository or registry (Harper et al. 2018). Instead, a network of independent databases and infrastructures can deliver data through a shared platform using standard transfer protocols (via Application Programming Interface, API). The data remains with providers, as can the access controls.

Data federation is not novel, and many of the FAIR principles (Wilkinson et al. 2016) underpin data federations' functions. Some examples include Earth System Grid Federation (Petrie et al. 2021), materials science data discovery (Plante et al. 2021), and OneGeology (One Geology 2020). In agriculture, there is AgDataCommons (USDA 2021), proposed UK Food 'Data Trust' (Pearson et al. 2021), AgINFRA (Drakos et al. 2015), and CGIAR Platform for Big Data In Agriculture (CGIAR 2021). Many of these data federation initiatives specify standards for description and exchange of data, focus on a particular data type of provider and/or provide a central intermediate space to standardise data. We believe agriculture required a different approach given the diversity of data stores including ways data is structured, described, and delivered; differences in organisational and research requirements and norms; and economic, trust and Intellectual Property concerns connected to agricultural data in general.

From 2018 we piloted a community-governed federation approach (the Agricultural Research Federation, AgReFed) (Box et al. 2019a). Participants provisioned their data holdings from their own choice of data repository aligned to their organisation's capabilities and requirements of their research field. Concurrently, they aligned with collective expectations for FAIR data. This required developing acceptable levels of FAIR data to be implemented and governed by AgReFed participants. Current practices adopt FAIR as high-level guiding principles (Wilkinson et al. 2016) or generalist practice recommendations (Bahim et al. 2020). This case study addressed this gap in an agricultural-specific implementation of FAIR in practice. We:

- I. co-developed FAIR threshold criteria for participants to deliver data through a federation
- II. though a consensus process integrated these FAIR thresholds into a framework for ongoing governance by a research domain community, for generating individual and collective benefit and growth of a data federation

2. THE USE CASES

The datasets of the pilot included point observations, spatial, temporal, on-ground, sensor, and remote sensed data. The data described plants (yield, crop rotation, metabolomic, proteomic, hyperspectral), soil and climatic variables from across Australia (Table 1).

IN-TEXT ABBREVIATION	DATA PRODUCTS' NAME	DATA PRODUCT TYPE	DATA PROVIDER TO AGREED (* INDICATES BOTH DATA PROVIDER AND USER OF THE DATASET OR COLLECTION)
SH (Soil Health)	Corangamite Soil Health Monitoring Program Data https://doi.org/10.25955/5c1c6b8f4d8d2 (Corangamite Catchment Management Authority, 2019)	Dataset and service	Federation University, Centre for eResearch and Innovation (CeRDI).
SMN-1 (Soil Moisture Network 1)	Soil moisture probe network, SFS (Southern Farming Systems https://doi.org/10.25955/5cdcff6168a76 (Southern Farming Systems, 2011)	Dataset	Federation University, CeRDI.
WT (Wheat Trials)	Waite Permanent Rotation Trial https://doi.org/10.4225/08/55E5165ECOD29 (Sanderman et al. 2015)	Dataset	*University of Adelaide, School of Agriculture, Food and Wine
NS (NatSoils)	Soil SITES database (NatSoil) https://doi.org/10.25919/5c36d77a6299c (CSIRO 2013)	Dataset and service	Commonwealth Scientific and Industrial Research Organisation (CSIRO)
SLG (Soil and Landscape Grid)	Soil and Landscape Grid National Soil Attribute Maps (3" resolution), Release 1 Collection. Sample see Rossel et al. (2014) https://doi.org/10.4225/08/546ED604ADD8A	Data product (maps), collection and service	CSIRO
FT (Frost Trials)	Crop Variety Frost Trial data collections https://doi.org/10.26182/5cedf001186f3 (Taylor et al. 2019)	Dataset collection	*University of Western Australia (UWA) and Department of Primary Industries and Regional Development (DPIRD)
SMN-2 (Soil Moisture Network 2)	SensorNets – SMART Farms Soil Moisture Network https://doi.org/10.4226/95/5b10d5ca18aef (Schneider et al. 2018)	Dataset	*University of New England (UNE)

Table 1 The data providers and their data products.

The data providers defined a set of research use cases for the data in Table 1 (see MacLeod et al. 2020: 29–31), identifying the current and anticipated data users and their ideal user experience. We then identified the requirements of the AgReFed platform and the (meta)data needed to deliver the use cases, and the FAIR principles that supported these requirements as follows:

- Allow the datasets and the services delivering the data to be discovered through metadata. Ideally the ability to discover should be persistent and through multiple avenues (Findable Q1, Q2, Q3 and Accessible Q4 and Q7, Table 2).
- Support appropriate data reuse and access controlled from the providers' infrastructure through licencing, data access controls and attribution (Accessible Q5 and Q6 and Reusable Q12 and Q14, Table 2).
- Allow the data to be queried on user-defined parameters including temporal and spatial properties, what is being measured (e.g., 'wheat', 'water'), the observed property being measured, the result, the procedure used to obtain the result, and the units of measurement (Interoperable Q9 and Q10 and Accessible Q6, Table 2).

- Allow a subset of the data to be visualised through the platform and downloaded in a useable format (e.g., CSV). This requires a web service interface (Accessible Q6 and Interoperable Q8 and Q9, [Table 2](#)).
- Allow the combining of data from different datasets (Interoperable Q8 and Q9). This requires the ability to map terms in the data to external vocabularies and semantics (e.g., replacing local descriptive terms with published controlled vocabulary concepts, such as ‘m’ or ‘metre’ with ‘<http://qudt.org/vocab/unit/M>’) (Interoperable Q10, [Table 2](#)).
- Allow locality to be interoperable between datasets (for example latitude and longitude with coordinate reference system) (Interoperable Q9 and Q10, [Table 2](#)).

Data collection and service records need to be discoverable through the federation’s platform ([AgReFed, 2021](#)). AgReFed currently harvests from Research Data Australia ([MacLeod et. al. 2020](#)). Therefore, it is an additional requirement that minimum metadata is entered into or harvestable by Research Data Australia ([Box et al. 2019a: 36–37](#)).

Table 2 AgReFed Version 1 FAIR thresholds for participation ([Box et al. 2019a: 22](#)).

Light grey indicates the AgReFed minimum acceptable requirements (‘Minimum thresholds’) and dark grey the ideal (‘Stretch targets’). The start-status and end-status indicate the progression of FAIR maturity. Data products are **SH** Soil Health; **SMN-1** Soil Moisture Network 1; **WT** Wheat Trials; **NS** NatSoils; **SLG** Soil Landscape Grid; **FT** Frost Trials; **SMN-2** Soil Moisture Network 2.¹ The minimum metadata requirement for data collections and services ([Box et al. 2019a: 36–37](#)).² ‘Machine-readable’ defined in terms of both syntax and structure, that is, as the representation of data products in a standard computer language that is structured in a way that is interpretable by machines.

	START-STATUS	END-STATUS
FINDABLE		
Q1. The data product has been assigned (an) identifier(s)		
No identifier	FT	
Local identifier		
Web address (URL)	SH, SMN-1	
Globally unique, citable, and persistent identifier (e.g., DOI, PURL, or Handle)	WT, NS, SLG, SMN-2	SH, SMN-1, WT, NS, SLG, FT, SMN-2
Q2. The data product identifier is included in all metadata records/files describing the data		
No	SH, SMN-1, FT, SMN-2	
Yes	WT, NS, SLG	SH, SMN-1, WT, NS, SLG, FT, SMN-2
Q3. The data product is described by a metadata record		
Not described	SH, SMN-1, FT	
Brief title and description	SMN-2	
Brief title, description, and other fields	WT, NS	
Comprehensively ¹ in a formal metadata schema	SLG	SH, SMN-1, WT, NS, SLG, FT, SMN-2
Q4. The data product is described by a metadata record that is indexed in a searchable registry or repository.		
Not indexed	SH, SMN-1, FT	
Local institutional repository		
Domain specific repository		
Generalist public repository		
Discoverable through several places (i.e., other registries, Research Data Australia, Google Data Search)	WT, NS, SLG, SMN-2	SH, SMN-1, FT, WT, NS, SLG, SMN-2
ACCESSIBLE		
Q5. How accessible is the data? The access method(s) must be explicitly stated in the metadata record e.g., if any authentication is needed, or there are any restrictions to access.		
Not accessible	SH, SMN-1	
Access to metadata only		
Through unspecified access conditions e.g., ‘contact the data custodian to discuss access’	NS, FT, SMN-2	SMN-2
Embargoed access after a specified date; or a de identified version of the data is publicly accessible		
Fully accessible public, or to persons who meet and follow explicitly stated conditions and processes, e.g., ethics approval for sensitive data	WT, SLG	SH, SMN-1, NS, FT, WT, SLG,

(contd.)

	START-STATUS	END-STATUS
Q6. Data are available for reuse via a standardised communication protocol, such as file download over https, or a web service		
No access to data	SH, SMN-1, FT	
By individual arrangement	SMN-2	SMN-2
File download online	WT, SLG (partial)	
Non-standard web service (e.g., OpenAPI/Swagger/informal API)		WT, FT
Standard web service API (e.g., OGC)	NS, SLG (partial)	SH, SMN-1, NS, SLG (full)
Q7. The repository/registry agrees to maintain the persistence of the metadata record, even if the data product is no longer available		
No, or not applicable if no metadata record	SH, SMN-1, FT	
Unsure	WT	
Yes	NS, SLG, SMN-2	SH, SMN-1, NS, SLG, FT, SMN-2, WT
INTEROPERABLE		
Q8. The data products are available in (an) open (file) format(s)		
Data are mostly available only in a proprietary format	WT, FT	
Data are available in an open format	SH, SMN-1	
Data are available in an open, documented, widely used standard format (e.g., NetCDF, CSV, JSON, XML)	NS, SLG, SMN-2	SH, SMN-1, WT, NS, SLG, FT, SMN-2
Q9. The data is machine-readable?		
The data are unstructured	SMN-1, WT, FT	
The data are structured and machine-readable (e.g., csv, JSON, XML, RDF, database files)	SH, NS, SLG, SMN-2	SH, SMN-1, WT, NS, SLG, FT, SMN-2
Q10. The data are semantically interoperable, because they use standard, accessible ontologies and/or vocabularies to describe the data elements/variables.		
Data elements are not described (i.e., fields or objects are labelled with codes or not at all)	SMN-2	
Data elements are described (so that a human user can correctly interpret the data), but no standards have been used in the description	SH, SMN-1, WT, FT	SMN-2
Recognised standards have been used in the description of data elements, but no published vocabularies with resolvable URIs	NS, SLG	SLG, FT
Published vocabularies using resolvable global identifiers linking to explanations are used, so that the data can be read and understood by machines as well as humans.		SH, SMN-1, NS, WT
Q11. The relationships to other data and resources (e.g., related datasets, services, publications, grants, etc) are described in the metadata or data, to provide context around the data		
There are no links to other metadata or data	SH, SMN-1, FT, SMN-2	SMN-2
The metadata record includes URI links to related metadata, data, and definitions	WT, NS	NS
Qualified links to other resources are recorded in a machine-readable format, e.g., a linked data format such as RDF	SLG	SH, SMN-1, WT, SLG, FT
REUSABLE		
Q12. Machine-readable data licenses are assigned to each data product, and are stated in the metadata record		
No licence applied	SH, SMN-1, FT, SMN-2	FT (standard licence but not in metadata record)
Non-standard license applied, with a machine-readable license/license deed URL	WT	
Standard license applied, without a machine-readable license deed URL		
Standard license applied, with a machine-readable license/license deed URL	NS, SLG	SH, SMN-1, WT, NS, SLG, SMN-2
Q13. The provenance of the data product is described in the metadata i.e., project objectives, data generation/collection (including from external sources) and processing workflows.		
None recorded	SH, FT, SMN-2	FT
Partially recorded	SMN-1, WT	SMN-2
Comprehensively recorded in a text format (e.g., TXT or PDF)	NS, SLG	WT, NS, SLG
Comprehensively recorded in a machine-readable format (e.g., in metadata record's schema or PROV, or in RDF, JSON, NetCDF, or XML)		SH, SMN-1
Q14. The preferred citation for the data product is provided in metadata record		
No	SH, FT, SMN-1	
Citation but with no persistent identifiers		
Citation with persistent identifiers	WT, NS, SLG, SMN-2	SH, SMN-1. WT, NS, SLG, FT, SMN-2

3. DEVELOPING AND TESTING THE FAIR THRESHOLDS

The development of the FAIR Thresholds for AgReFed participation were co-developed by the participating research data experts and data providers. Baseline assessments of the 'FAIRness' of providers' data ('start-status' in [Table 2](#)) were made using the Australian Research Data Commons (ARDC) FAIR data self-assessment tool ([Schweitzer et al. 2021](#)). The manual ARDC self-assessment tool articulates various levels of FAIR maturity (or 'FAIRness') of (meta)data from not at all discoverable or machine understandable, through to fully understandable by both humans and machines. It also serves as an education resource for providers working to improve the FAIR maturity of their holdings.

Following this baseline assessment, providers determined where improvements could be made to move their data products along the FAIR continuum. Solutions were identified that met their own organisation's goals, capabilities, and end users' needs. These were combined with requirements in the Use Cases to identify 'Minimum thresholds' of data maturity required to support key platform functionality for (meta)data discovery, access and reuse through AgReFed. 'Stretch targets' were also defined to communicate to the agricultural research community the level of data maturity that enables maximal data integration and (re)use (see the shading in [Table 2](#)).

As well as the addition of 'Minimum thresholds' and 'Stretch Targets', the content of the ARDC FAIR tool was modified somewhat to assist with ease of interpretation ([Table 2](#)). Changes made in response to user feedback included:

- Examples of some possible information and technology solutions were worked into the questions and answers.
- The concept of 'comprehensive' metadata was clearly specified for both data collection and service records (see [Box et al. 2019a: 36–37](#)).
- Preferred citation in the metadata was added as an AgReFed requirement (Q14)
- The openness of the file format was separated from the machine readability of the data (Q8). The term 'Machine-readable' was defined in terms of both syntax and structure, that is, as the representation of data products in a standard computer language structured in a way that is interpretable by machines (Q9).
- A challenge for data providers was that their (meta)data were not only individual datasets contained in a single file but multiple collections, derivations (e.g., maps) and data service endpoints. Hence the ARDC FAIR assessment was refocused from the word 'data' to 'data product', being the data collection or product that is provided to users, along with any associated metadata or services required for its delivery. Here for simplicity of a manual assessment Q1 – 4, 7, 10, 12 and 13 are focused on assessment of the metadata and 5, 6, 8, 9 and 10 on the data. It is acknowledged that data and metadata can be assessed for FAIRness independently (see [Bahim et al. 2020](#)) and the feasibility of assessing this way for AgReFed's purposes should be evaluated in the future.

The ARDC and the Centre for eResearch and Digital Innovation (CeRDI) supported the data providers to improve the level of FAIR maturity of their data across a twelve-month period (2018–2019). The baseline assessments, progress to the final states and the information and technology solutions used at those states are available as supplementary data ([Levett et al. 2022](#)). Some notable experiences informed the AgReFed FAIR 'Minimum thresholds' to 'Stretch targets' ([Table 2](#)). These included:

- Providers' exemplar data products each had different FAIR starting points (See start-status, [Table 2](#)).
- Improvements to metadata records to meet AgReFed 'Minimum threshold' requirements were possible with organisational library and ARDC support (Findability Q1 to Q4) so 'Minimum thresholds' were set high for Q1 to Q3.
- Access requirements and licencing varied. These were accommodated across the thresholds of Q5 and Q12.

- Data format and structure (Interoperability Q8 and Q9) and data access method (Accessibility Q6) varied between providers, as did FAIR solutions. The solutions varied depending on the data types and the organisational/research group aspirations, skills, and IT support available. So, examples of acceptable solutions were given for Interoperability Q8 and Q9, and ‘Minimum thresholds’ to ‘Stretch targets’ highlighted for Accessibility Q6. Provider examples included a data service provider converting sensor data from web viewable-only-HTML to O&M structured data in machine-readable format (JSON), delivered by Sensor Things API via Frost-server. Agronomic researchers converted data in Microsoft Excel tables to PostGreSQL and MySQL databases with partial O&M design patterns. These delivered JSON and CSV by Swagger PostgREST API and OpenAPI.
- The semantic interoperability (Q10) of the data products was initially highly variable. However, no data providers utilised vocabularies that were FAIR (Cox et al. 2021) or near to FAIR. This was a ‘Stretch target’ for providers, reflected in the sliding scale from the ‘Minimum threshold’ (Q10). Providers described data with the URIs of external machine-readable vocabularies from within their database headers or lookup tables. These were expressed through the API endpoints. Challenges included finding and selecting vocabularies including evaluating authority and persistence; and the need to create (e.g., Cox & Gregory, 2020) and therefore upskill.
- Provenance was recorded in different formats, reflected in ‘Minimum threshold’ to ‘Stretch target’ (Reusable Q13). Improvements were inconsistent and further work is needed defining ‘comprehensive’ content.

4. FAIR THRESHOLD GOVERNANCE AND IMPLEMENTATION

The FAIR thresholds were presented to the AgReFed Council and approved through consensus (see [AgReFed Council Terms of Reference](#), Wong et al. 2021) for integration into AgReFed’s Membership and Technical Policy (Wong et al. 2021; MacLeod et al. 2020). An in-depth discussion of AgReFed’s architecture is not the focus of this practice paper and is reported elsewhere (Box et al. 2019a). However, we provide an overview in the context of how founding members implemented the governance around the FAIR thresholds.

AgReFed’s operation and design is a federated architecture (Figure 1) (Box et al. 2019a). It draws on a Service Orientated Architecture Reference Model design for Open Distributed Processing (RM-ODP) (ISO/IEC 1996), with the addition of a unique ‘Social Architecture’ viewpoint to structure social aspects of the system – such as governance. AgReFed’s Social Architecture adopts a democratic cooperative governance approach. It is led by its members to meet shared goals of self-governance, trust through active participation, and self-determination (Buchanan 1965; Pentland & Hardjono 2020). Governance, Roles and Responsibilities are defined in the Social (Membership, Financial and Strategic) and Technical Policies (Wong et al. 2021, MacLeod et al. 2020) that determine operation of AgReFed including the implementation and governance of FAIR thresholds.

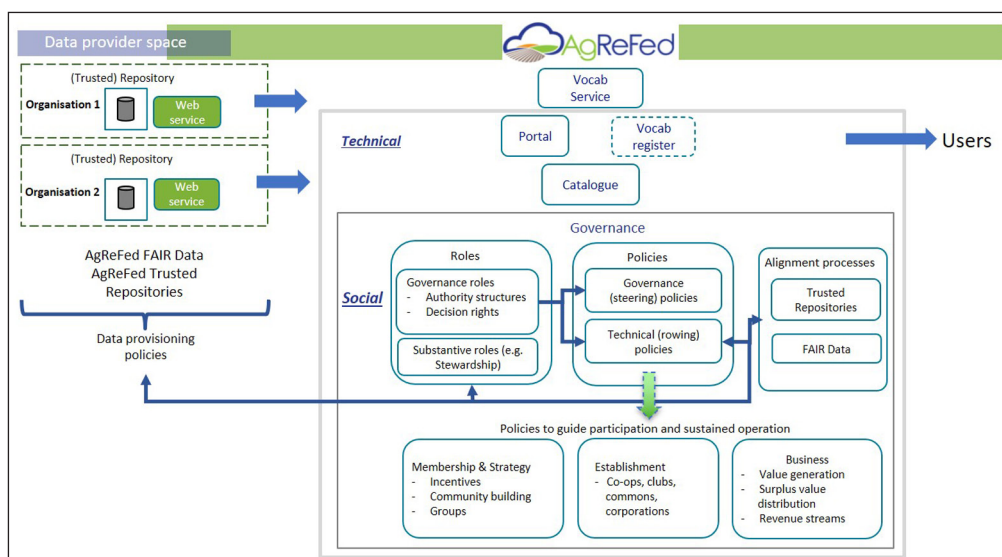


Figure 1 The FAIR alignment process within the AgReFed architecture. FAIR thresholds are part of the alignment process for organisations to participate in the federation. They are integrated into Governance and Technical Policies, and Roles and Responsibilities (Box et al. 2019a: 7).

The recommended process is that applications for membership to AgReFed are assessed by a Federation Data Steward (Box et al. 2019a: 11, Wong et al. 2021: 22). They assess if the provider/provider community meets the Membership Policy including whether the thresholds are met. The Technical Committee work with the Federation Data Steward and potentially Data Standards and Vocabularies Steward or delegated expert advisors/groups to ensure the partners' solutions align with and are integrated into Technical Policy.

The provider has now demonstrated alignment with collective expectations for FAIR data (Box et al. 2019a). They nominate a Data Provider Collection Custodian and member to Council and (optionally) Technical Committee. Their (meta)data is made discoverable/harvestable to AgReFed and they are now AgReFed members. All members have equal participation and decision rights. In this way, the community participates in the governance of the FAIR threshold settings including how they are maintained and implemented.

5. REFLECTIONS AND NEXT STEPS

The manual AgReFed FAIR thresholds assessment, with practical examples and definitions, was useful for helping data providers conduct meaningful assessments of their data across the full continuum of data maturity. It was also useful for developing and implementing works plans. However, to enable transparency, repeatability, and scalability of assessments across the agricultural domain some improvements could be made. Where (meta)data and services are machine actionable, automated assessment could be used to support scalability and repeatability (see Devarju et al. 2021). In contrast, manual assessment will still be required for less mature meta(data) or where a more nuanced interpretation is required for example of content to enhance re-usability. In the current platform phase (ARDC, 2020) we plan to improve repeatability of the FAIR threshold assessment by integrating a hybrid (semi-automated) approach (Peters-von Gehlen et al. 2022). To improve transparency and repeatability the evidence required for both manual and machine assessment will need to be specified and may include, as examples, screen shots and automated assessment outputs.

Our experience highlighted the expertise of a Federation Data Steward will be essential for assisting partners with FAIR threshold assessments. As the federation grows the assessments will encompass more standards and technology solutions used by different communities (for example see FAIRsharing, Sansone et al. 2019). If various solutions align with or should be integrated into AgReFed Technical Policy this will need to be evaluated by the steward in consultation with the Technical Committee. Keeping up to date with current developments such as the FAIR Data Maturity Model (Bahim et al. 2020) will ensure the relevance and currency of the policy and thresholds. A dedicated Federation Standards and Vocabulary Steward (Box et al. 2019a: 18) would be valuable here for brokering conversations with expert domain communities and working groups that can advise or make delegated decisions.

Here, we focused on defining FAIR data thresholds. However, we recognise that repositories that the data is served from should be 'FAIR data enabling' as a critical component of the broader 'FAIR data ecosystem' (Collins et al. 2018; Devaraju et al. 2021). There are various ways of assessing or accrediting repositories relating to areas of security risk management, organisational and physical infrastructure, and digital object management (Lin et al. 2020). As a preliminary trial, we included assessment of a 'pass' or 'fail' of several CoreTrustSeal requirements deemed necessary for persistent delivery of trusted of agricultural data whilst not being onerous and disincentivising participation (Box et al. 2019a: 23). Our early experience showed that research scientists and even data managers had difficulty knowing if their repositories complied. Furthermore, there were challenges knowing what to assess if the data products were served from multiple repositories. AgReFed could play a role helping providers assess and choose repositories that meet community expectations. We look forward to learning about the solutions of other domains here.

Our experience was that the starting point FAIRness of pilot participants' data varied as did their priorities, capacities, and solutions for improving. To ensure these viewpoints were encapsulated, setting the FAIR thresholds and their governance and implementation was done through consensus with providers. It is envisaged that this active participation will help ensure the settings are realistic and promote trust and self-determination giving providers incentive

to participate. The thresholds aimed to strike a balance between the realities and priorities of providers so as not to disincentivise participation whilst also aiming to inspire, support and educate for fully FAIR data and meet end-users needs.

Improving the FAIRness of data took resourcing, so value propositions are required for providers to have confidence in participation. Benefits to founding partners included being an exemplar of FAIR best practice at the institutional level, making access and re-use easier for end-users, and being able to combine data types for research insights (see [Use Case stories, AgReFed 2021](#)). Providers benefited from metadata guidance through education resources, library, and licencing support. Expert assistance including from providers' organisational IT was required for data structuring, access through APIs, and finding, selecting, creating, and applying vocabularies. In one case (SMN-2) institutional IT resourcing for data service work was a challenge but raised the prioritisation of upgrades now being worked on. Data federations can support collective advocacy, relationship brokering, and tailored support across these areas.

The provision, assembling and demonstration of tooling resources for data providers' various needs, priorities and capabilities would also lower the cost of delivering FAIR data, thereby incentivising federation participation. Examples across the data management cycle include data management plans, data collection tools (e.g., [Devare et al. 2021](#)), data deposition tools (e.g., [Shaw et al. 2020](#)) and example protocol/reference implementations (for example [FAIR Data Points 2022](#)). This is a focus of AgReFed's next phase. Virtual research environments with example workflows are also being integrated. Furthermore, the federation can continue to align/encourage membership with intermediates or broker platforms that offer value in specific fields of research including in data standardisation.

The current phase has focused on research institutes. Expanding participation to co-operatives, Research Development Corporations, industry, and farmers as envisaged by members ([Box et al. 2019b](#)) will require incentivisation. The governance structure of AgReFed enables the community to make policy adjustments to support this. For example, alternative funding models may be leveraged, such as user-pays for certain services and data in the competitive space. Stakeholders can bring assets aside from data to the table to help meet the varied needs of participants. Recognising this, membership was recently expanded to providers of tooling, infrastructure, and other resources. Active participation through the federation will help ensure individual and collective benefits are delivered across the agricultural research sector, including through FAIR and trusted data.

DATA ACCESSIBILITY STATEMENT

Datasets informing this research are available by open licence permitting unrestricted access (<https://doi.org/10.25955/5c1c6b8f4d8d2>, <https://doi.org/10.4225/08/55E5165ECOD29>, <https://doi.org/10.25919/5c36d77a6299c>, <https://doi.org/10.4225/08/546ED604ADD8A>, <https://doi.org/10.26182/5cedf001186f3>) unless otherwise stated (<https://doi.org/10.4226/95/5b10d5ca18aef>, <https://doi.org/10.25955/5cdcff6168a76>). The FAIR assessments are available at <https://doi.org/10.5281/zenodo.6541413>.

ACKNOWLEDGEMENTS

We gratefully acknowledge the assistance from Catherine Brady and Melanie Barlow (ARDC) for metadata and services support; the data contributions of Dr Ben Biddulph (Department of Primary Industries and Regional Development, WA), Southern Farming Systems and Corangamite Catchment Management Authority; technical development by Andrew MacLeod, Scott Limmer and Heath Gillett (Federation University), Linda Gregory (CSIRO, National Soil Data and Information), Daniel Watkins (University of New England); vocabulary support by Simon Cox (CSIRO, Environmental Informatics) and policy work and manuscript feedback from Dr Joel Epstein. Thank you to all those providing review of AgReFed documents cited herein including Dr Andrew Treloar (ARDC), Prof. Harvey Millar (The University of Western Australia), Peter Wilson (CSIRO, National Soil Data, and Information), Assoc. Prof. Peter Dahlhaus and Jude Channon (Federation University), Prof. Matthew Gilliam (University of Adelaide), Dr Bettina Berger (University of Adelaide), Dr Kay Steel (Federation University) and Dr Rachelle Hergenhan (University of New England).

FUNDING INFORMATION

This research was supported by the Australian Research Data Commons (ARDC) Agriculture Research Data Cloud project (DC063) and ARDC Discovery Activities (TD018). The ARDC is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Methodology and framework was developed by P. Box derived from previous work and experiences, with contributions to development and implementation of the framework by B. Simons, K. Levett, A. MacLeod and M. Wong and testing and feedback by R. David, D. Schneider and N. Taylor. A. Lee, P. Box, H. Thompson, B. Simons, A. MacLeod and M. Wong made contributions to writing social and or technical policies. M. Wong led article writing, with significant writing contributions from K. Levett and A. Lee. All authors contributed to writing including revising critically for intellectual input. P. Box refined methodological scope and significance in early drafts.

AUTHOR AFFILIATIONS


Megan Wong  orcid.org/0000-0002-2991-2308
Federation University, AU

Kerry Levett  orcid.org/0000-0001-5963-0195
Australian Research Data Commons, AU


Ashlin Lee  orcid.org/0000-0003-0524-121X
Commonwealth Scientific and Industrial Research Organisation, AU

Paul Box  orcid.org/0000-0002-7073-8858
Commonwealth Scientific and Industrial Research Organisation, AU

Bruce Simons  orcid.org/0000-0001-7519-5875
Federation University, AU

Rakesh David  orcid.org/0000-0002-3306-7581
University of Adelaide, AU

Andrew MacLeod  orcid.org/0000-0003-4834-5481
Federation University, AU

Nicolas Taylor  orcid.org/0000-0003-2004-5217
The University of Western Australia, AU

Derek Schneider  orcid.org/0000-0002-1897-4175
University of New England, AU

Helen Thompson  orcid.org/0000-0001-7698-450X
Federation University, AU

REFERENCES

- AgReFed.** 2021. *AgReFed – Agricultural Research Federation*. Agricultural Research Federation. Available at: <https://www.agrefed.org.au> [Last accessed 11 November 2021].
- Allemang, D** and **Bobbin, T.** 2016. *A Global Data Ecosystem for Agriculture and Food*. Oxfordshire: GODAN, CABI. Available at: https://www.godan.info/sites/default/files/documents/Godan_Global_Data_Ecosystem_Publication_lowres.pdf [Last accessed 17 December 2021].
- Antle, JM,** et al. 2017. 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>
- ARDC.** 2020. *AgReFed: A platform for the transformation of agricultural research*. Australian Research Data Commons. DOI: <https://doi.org/10.47486/PL005>
- Bahim, C,** et al. 2020. The FAIR data maturity model: An approach to harmonise FAIR assessments. *Data Science Journal*, 19(1): 1–7. DOI: <https://doi.org/10.5334/dsj-2020-041>

- Barry, S**, et al. 2017. *Precision to Decision – Current and Future State of Agricultural Data for Digital Agriculture in Australia*. Available at: <https://www.crdc.com.au/precision-to-decision> [Last accessed 11 November 2021].
- Box, P**, et al. 2019a. *Guidelines for the development of a Data Stewardship and Governance Framework for the Agricultural Research Federation (AgReFed)*. Sydney: Commonwealth Scientific Industrial Research Organisation. DOI: <https://doi.org/10.25919/5cf179ba35db9>
- Box, P**, et al. 2019b. *White Paper for the enactment phase of the Agricultural Research Federation (AgReFed)*. Sydney: Commonwealth Scientific Industrial Research Organisation. DOI: <https://doi.org/10.5281/ZENODO.3706374>
- Buchanan, JM**. 1965. An Economic Theory of Clubs. *Economica*, 32: 1–14. DOI: <https://doi.org/10.2307/2552442>
- CGIAR**. 2021. CGIAR Platform for Big Data in Agriculture <https://bigdata.cgiar.org/> [Last accessed 15 November 2021].
- Chiles, RM**, et al. 2021. Democratizing ownership and participation in the 4th Industrial Revolution: challenges and opportunities in cellular agriculture. *Agriculture and Human Values*, 38: 943–961. DOI: <https://doi.org/10.1007/s10460-021-10237-7>
- Collins, S**, et al. 2018. *Turning FAIR into reality*. Brussels: European Commission. DOI: <https://doi.org/10.2777/1524>
- Cox, S**, and **Gregory, L**. 2020. *RDF representation of ASLS soil profile classification*. v1. Australia: CSIRO. DOI: <https://doi.org/10.25919/5f42f324b2ef8>
- Cox, SJD**, et al. 2021. Ten simple rules for making a vocabulary FAIR. *PLOS Computational Biology*, 17(6): 1009041. DOI: <https://doi.org/10.1371/journal.pcbi.1009041>
- Corangamite Catchment Management Authority**. 2019. *Corangamite Soil Health Monitoring Program Data. Version 1.0*. Mt Helen, Australia: Federation University. DOI: <https://doi.org/10.25955/5c1c6b8f4d8d2>
- CSIRO**. 2013. *CSIRO National Soil Site Database. Version 1*. Australia: CSIRO. DOI: <https://doi.org/10.25919/5c36d77a6299c>
- Datacite**. 2022. *Repository Finder*. The Enabling FAIR Data Project and FAIRsFAIR Project. Accessible at: <https://repositoryfinder.datacite.org/> [Last accessed at 02 March 2022].
- Devaraju, A**, et al. 2021. From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. *Data Science Journal*, 20(4): 1–14. DOI: <https://doi.org/10.5334/dsj-2021-004>
- Devare, M**, et al. 2021. AgroFIMS: A Tool to Enable Digital Collection of Standards-Compliant FAIR Data. *Frontiers of Sustainable Food Systems*, 5: 1–12. DOI: <https://doi.org/10.3389/fsufs.2021.726646>
- Drakos, A, Protonotarios, V and Manouselis, N**. 2015. agINFRA: A research data hub for agriculture, food and the environment. *F1000Research*, 4. DOI: <https://doi.org/10.12688/f1000research.6349.2>
- Ernst and Young**. 2019. *Agricultural Innovation – A National Approach to Grow Australia’s Future*. Accessible at: <https://www.awe.gov.au/sites/default/files/sitecollectiondocuments/agriculture-food/innovation/summary-report-agricultural-innovation.PDF> [Last accessed 24 March 2022].
- FAIR Data Points**. 2022. Fair Data Points. Available at <https://www.fairdatapoint.org/> [Last accessed 24 March 2022].
- Harper, L**, et al. 2018. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database*. DOI: <https://doi.org/10.1093/database/bay088>
- Ingram, J**, et al. 2022. What are the priority research questions for digital agriculture? *Land Use Policy*, 114. DOI: <https://doi.org/10.1016/j.landusepol.2021.105962>
- ISO/IEC**. 1996. *Information technology -- Open Distributed Processing – Reference Model: Architecture*. ISO/IEC 10746-3: 1996. Geneva, Switzerland: ISO/IEC.
- Jakku, E**, et al. 2019. “If they don’t tell us what they do with it, why would we trust them?” Trust, transparency and benefit-sharing in Smart Farming. *NJAS – Wageningen Journal of Life Sciences*, 100285: 90–91. DOI: <https://doi.org/10.1016/j.njas.2018.11.002>
- Kenney, M, Serhan, H and Trystram, G**. 2020. Digitization and Platforms in Agriculture: Organizations, Power 2020 Asymmetry, and Collective Action Solutions. *SSRN*, 1–50. DOI: <https://doi.org/10.2139/ssrn.3638547>
- Knapen, MJR**, et al. 2020. AGINFRA PLUS: Running Crop Simulations on the D4Science Distributed e-Infrastructure. In: *Environmental Software Systems. Data Science in Action. ISESS 2020. IFIP Advances in Information and Communication Technology*, vol 554. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-39815-6_8
- Kruseman, G**, et al. 2020. CGIAR modelling approaches for resource- constrained scenarios: II. Models for analyzing socioeconomic factors to improve policy recommendations. *Crop Science*, 60(2): 568–581. DOI: <https://doi.org/10.1002/csc2.20114>
- Lin, D**, et al. 2020. The TRUST Principles for digital repositories. *Scientific Data*, 7(1): 1–5. DOI: <https://doi.org/10.1038/s41597-020-0486-7>
- Levett, K, Wong, M and MacLeod, A**. 2022. Testing of AgReFed FAIR data Minimum Thresholds and Stretch Targets (Version 1). *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.6541413>

- MacLeod, A**, et al. 2020. *The Agricultural Research Federation (AgReFed) Technical and Information Policy Suite*. The Agricultural Research Federation. DOI: <https://doi.org/10.5281/ZENODO.3993784>
- One Geology**. 2020. One Geology. Available at <https://www.onegeology.org/> [Last accessed 24 March 2022].
- Pearson, S**, et al. 2021. *Food Data Trust: A framework for information sharing*. United Kingdom: Food Standards Agency and University of Lincoln. FSA, project ref FS301083. DOI: <https://doi.org/10.5281/zenodo.4575565>
- Pentland, A** and **Hardjono, T**. 2020. Data Cooperatives. In *Building the New Economy*. DOI: <https://doi.org/10.21428/ba67f642.0499afe0>
- Perrett, E**, et al. 2017. *Accelerating Precision Agriculture to Decision Agriculture – Analysis of the Economic Benefit and Strategies for Delivery of Digital Agriculture in Australia*. Sydney: Australian Farm Institute. Available at: <https://www.crdc.com.au/precision-to-decision>.
- Peters-von Gehlen, K**, et al. 2022. Recommendations for Discipline-Specific FAIRness Evaluation Derived from Applying an Ensemble of Evaluation Tools. *Data Science Journal*, 21: 1–21. DOI: <https://doi.org/10.5334/dsj-2022-007>
- Petrie, R**, et al. 2021. Coordinating an operational data distribution network for CMIP6 data. *Geoscientific Model Development*, 14: 629–644. DOI: <https://doi.org/10.5194/gmd-14-629-2021>
- Plante, RL**, et al. 2021. Implementing a registry federation for materials science data discovery. *Data Science Journal*, 20: 1–9. DOI: <https://doi.org/10.5334/dsj-2021-015>
- Rossel, V**, et al. 2014. Soil and Landscape Grid National Soil Attribute Maps – Available Water Capacity (3” resolution) – Release 1. Version 4. Australia: CSIRO. DOI: <https://doi.org/10.4225/08/546ED604ADD8A>
- Sanderman, J**, et al. 2015. *Waite Permanent Rotation Trial. Version 4*. Australia: CSIRO. DOI: <https://doi.org/10.4225/08/55E5165ECOD29>
- Sanderson, T, Reeson, A, and Box, P**. 2017. *Cultivating Trust: Towards an Australian Agricultural Data Market*. Commonwealth Scientific Industrial Research Organisation. DOI: <https://doi.org/10.21820/23987073.2017.10.62>
- Sansone, S**, et al. 2019. FAIRsharing, a cohesive community approach to the growth in standards, repositories and policies. *Nature Biotechnology*, 37: 358–367. DOI: <https://doi.org/10.1038/s41587-019-0080-8>
- Schneider, D**, et al. 2018. *SensorNets – SMART Farms Soil Moisture Network*. Australia: University of New England. DOI: <https://doi.org/10.4226/95/5b10d5ca18aef>
- Schweitzer, M**, et al. 2021. *au-research/FAIR-Data-Assessment-Tool: Release v1.0*. Australian Research Data Commons. DOI: <https://doi.org/10.5281/zenodo.4971127>
- Shaw, F**, et al. 2020. COPO: a metadata platform for brokering FAIR data in the life sciences [version 1; peer review: 1 approved]. *F1000Research*, 9(495). DOI: <https://doi.org/10.12688/f1000research.23889.1>
- Southern Farming Systems (SFS)**. 2011. *Southern Farming Systems Moisture Probe Network Data*. Australia: Federation University DOI: <https://doi.org/10.25955/5cdcff6168a76>
- Taylor, N**, et al. 2019. *UWA/DPIRD Frost Nursery Trial 2018*. Australia: The University of Western Australia. DOI: <https://doi.org/10.26182/5cedf001186f3>
- USDA**. 2021. Ag Data Commons. U.S. Department of Agriculture <https://data.nal.usda.gov/> [Last accessed 24 March 2022].
- Wicquart, J**, et al. 2022. A workflow to integrate ecological monitoring data from different sources. *Ecological Informatics*, 68. DOI: <https://doi.org/10.1016/j.ecoinf.2021.101543>
- Wilkinson, MD**, et al. 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(160018). DOI: <https://doi.org/10.1038/sdata.2016.18>
- Wiseman, L** and **Sanderson, J**. 2018. Legal and trust issues in Australian agriculture. In: *40th Annual Conference Australian Society of Sugar Cane Technologists*. ASSCT. Handle: 10072/379876.
- Wong, M**, et al. 2021. *Agricultural Research Federation (AgReFed) Steering Policies, Roles and Responsibilities (Version 1.1)*. Agricultural Research Federation. DOI: <https://doi.org/10.5281/zenodo.5205273>

TO CITE THIS ARTICLE:

Wong, M, Levett, K, Lee, A, Box, P, Simons, B, David, R, MacLeod, A, Taylor, N, Schneider, D and Thompson, H. 2022. Development and Governance of FAIR Thresholds for a Data Federation. *Data Science Journal*, 21: 13, pp. 1–12. DOI: <https://doi.org/10.5334/dsj-2022-013>

Submitted: 21 December 2021

Accepted: 13 May 2022

Published: 09 June 2022

COPYRIGHT:

© 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.