Agricultural Research Federation (AgReFed) Data Policy and Process Guidelines

Paul Box[1], Kerry Levett[2], Bruce Simons[3], Megan Wong[3]

[1] CSIRO
[2] Australian Research Data Commons (ARDC)
[3] Federation University Australia

Version 1.2 2019-05-30

AgReFed
Agricultural Research Federation

AgReFed data Processes and Policicy Guidelines: From ‘my data’ to Our FAIR data

- These pages are exerts from the AgReFed FAIR and Trusted Policies which have been tested and developed in consultation with founding AgReFed partners*
- It serves as quick guide for those interested in participating in AgReFed to understand the principles guiding AgReFed FAIR and Trusted data for Agricultural Research
- For review by the AgReFed Technical Committee and AgReFed Council through the enactment phase (July 2019 –Oct 2019)

* Box, Paul; Levett, Kerry; Simons, Bruce; Wong, Megan. Guidelines for the development of a Data Stewardship and Governance Framework for the Agricultural Research Federation (AgReFed). Sydney: CSIRO; 2019.https://doi.org/10.25919/5cf179ba35db9

Citation

Box, Paul; Levett, Kerry; Simons, Bruce; Wong, Megan (2019) Guidelines for the development of a Data Stewardship and Governance Framework for the Agricultural Research Federation (AgReFed) Version 1.1. CSIRO.

Acknowledgements

This research was supported by the Australian Research Data Commons (ARDC). ARDC is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program (NCRIS).

Partners in the project were the University of Adelaide, the University of Western Australia, the Western Australian Department of Primary Industries and Regional Development (DPIRD), the University of New England, Federation University Australia and the Commonwealth Scientific and Industrial Research Organisation. The authors would like to acknowledge all individuals from these partner organisation that provided valuable input into this document, including technical input from Andrew MacLeod, significant feedback on drafts by Helen Thompson and Peter Wilson and early version input by Nigel Prior and Kay Steel.

Alignment processes and policies

From ‘my data’ to Our FAIR data

- Individual providers’ heterogeneous data and provisioning arrangements can be brought into **alignment** with agreed AgReFed levels of FAIRness and repository trustedness
- **AgReFed Trusted Repository Policy** and **AgReFed FAIR Data Policy** are used to define the qualifying levels required for data to be provided as AgReFed Data

AgReFed FAIR Data policy (p5-6) is based on FAIR principles.

The AgReFed FAIR Data Self-assessment is based on the ARDC FAIR Data Self-assessment Tool but incorporates maturity models to measure and track increases in interoperability delivered through services.

The development of the Assessment is explained further in the Appendix, Box *et al.* p38.

AgReFed Trusted Repository Policy (p7) is based on CoreTrustSeal Data Repositories Requirements.

The development of the Assessment is explained further in the Appendix, Box *et al.* p39.

Rowing: AgReFed FAIR Data Policy

The AgReFed FAIR Data Policy

1. AgReFed FAIR Data Self-assessment (p6) will be used as the data alignment process for AgReFed
2. Current thresholds required for qualification as AgReFed FAIR data are shown in the table on page 6
3. Standards for data provision including vocabulary standards will be defined as part of the AgReFed FAIR Data Policy settings
4. **Additional data structure semantics and syntax standards may be specified as part of the AgReFed FAIR Data Policy**
5. The Federation Data Steward will review and approve (or reject) data providers' AgReFed FAIR Data Self-assessments
6. Any disputes in relation to the validation of assessment will be escalated to the Federation Technical Committee for review and decision
7. The FAIR assessment process and settings (including qualifying threshold levels) may be modified by the Federation Technical Committee

See page 6 for detail

Principle (for AgReFed)	Increasingly FAIR ->					
FINDABLE						
Q1 The data product has been assigned (an) identifier(s)	No identifier	Local identifier	Web address (URL)	Globally unique, citable and persistent identifier (e.g. DOI, PURL, or Handle)		
Q2 The data product identifier is included in all metadata records/files describing the data	No	Yes				
Q3 The data product is described by a metadata record that facilitates discovery, access and reuse of the data.	The data is not described	Brief title and description	Brief title and description, and multiple other fields filled out, albeit briefly.	Comprehensively (including all AgReFed required fields*) using a formal machine-readable metadata schema.		
Q4 The data product is described by a metadata record that is indexed in a searchable registry or repository...	The data is not described in any registry or repository	Local institutional repository	Domain-specific repository	Generalist public repository	Data is in one place but discoverable through several places (i.e. other registries, RDA, Google Data Search)	
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.	No metadata record	Access to metadata only	Unspecified access conditions e.g. "contact the data custodian to discuss access"	Embargoed access after a specified date	A deidentified 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
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	By individual arrangement	File download from online location	Non-standard web service (e.g. OpenAPI/Swagger/Informal API)	Standard web service API (e.g. OGC)	
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 exists)	Unsure	Yes			
INTEROPERABLE						
Q8 The data products are available in (an) open (file) format(s)	Data are mostly available only in a proprietary format	Data are available in an open format	Data are available in an open, documented, widely-used standard format (i.e. NetCDF, CSV, JSON, XML, etc)			
Q9 The data is machine readable (see Glossary for definition)	The data are unstructured	The data are structured and machine-readable (i.e. csv, JSON, XML, RDF, database files, etc)				
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)	Data elements are described (so that a human user can correctly interpret the data), but no standards have been used in the description	Published vocabularies / ontologies / schema (without global identifiers) are used	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.		
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	The metadata record includes URI links to related metadata, data and definitions	Qualified links to other resources are recorded in a machine readable format, e.g. a linked data format such as RDF			
REUSABLE						
Q12 Machine-readable data licenses are assigned to each data product, and are stated in the metadata record.	No license is applied	Non-standard license applied, without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Non-standard license applied, WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	
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.	No provenance information is recorded	Partially recorded	Comprehensively recorded in a text format (i.e. TXT or PDF)	Comprehensively recorded in a machine readable format (i.e. in metadata record's schema or PROV, or in RDF, JSON, NetCDF, XML, etc)		
Q14 The preferred citation for the data product is provided in metadata record	No	Citation does not include identifier	Citation includes identifier			

AgReFed FAIR Data Policy qualifying thresholds:

The green cells indicate the proposed minimum acceptable level that data must comply with before it can be 'published' as AgReFed Data:

- Where different shades of green are shown, the lightest green indicates minimum acceptable level, and the darkest green indicates stretch goal

* Question 3 specifies minimum metadata requirements for collections and services (see appendix – pages 7-8)

AgReFed FAIR Data assessment – Initial settings

Principle (for AgReFed)	Increasingly FAIR -->				
FINDABLE					
Q1 The data product has been assigned (an) identifier(s)	No identifier	Local identifier	Web address (URL)	Globally unique, citable and persistent identifier (e.g. DOI, PURL, or Handle)	
Q2 The data product identifier is included in all metadata records/files describing the data	No	Yes			
Q3 The data product is described by a metadata record that facilitates discovery, access and reuse of the data.	The data is not described	Brief title and description	Brief title and description, and multiple other fields filled out, albeit briefly.	Comprehensively (including all AgReFed required fields*) using a formal machine-readable metadata schema.	
Q4 The data product is described by a metadata record that is indexed in a searchable registry or repository…	The data is not described in any registry or repository	Local institutional repository	Domain-specific repository	Generalist public repository	Data is in one place but discoverable through several places (i.e. other registries, RDA, Google Data Search)
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.	No metadata record	Access to metadata only	Unspecified access conditions e.g. "contact the data custodian to discuss access"	Embargoed access after a specified date; or A deidentified 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
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	By individual arrangement	File download from online location	Non-standard web service (e.g. OpenAPI/Swagger/informal API)	Standard web service API (e.g. OGC)
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 exists)	Unsure	Yes		
INTEROPERABLE					
Q8 The data products are available in (an) open (file) format(s)	Data are mostly available only in a proprietary format	Data are available in an open format	Data are available in an open, documented, widely-used standard format (i.e. NetCDF, CSV, JSON, XML, etc)		
Q9 The data is machine readable (see Glossary for definition)	The data are unstructured	The data are structured and machine-readable (i.e. csv, JSON, XML, RDF, database files, etc)			
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)	Data elements are described (so that a human user can correctly interpret the data), but no standards have been used in the description	Published vocabularies / ontologies / schema (without global identifiers) are used	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.	
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	The metadata record includes URI links to related metadata, data and definitions	Qualified links to other resources are recorded in a machine readable format, e.g. a linked data format such as RDF		
REUSABLE					
Q12 Machine-readable data licenses are assigned to each data product, and are stated in the metadata record.	No license is applied	Non-standard license applied, without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Non-standard license applied, WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record
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.	No provenance information is recorded	Partially recorded	Comprehensively recorded in a text format (i.e. TXT or PDF)	Comprehensively recorded in a machine readable format (i.e. in metadata record's schema or PROV, or in RDF, JSON, NetCDF, XML, etc)	
Q14 The preferred citation for the data product is provided in metadata record	No	Citation does not include identifier	Citation includes identifier		

- Where different shades of green are shown, the lightest green indicates minimum acceptable level, and the darkest green indicates stretch goal

* Question 3 specifies minimum metadata requirements for collections and services (see appendix – pages 36-37)

Rowing: AgReFed Trusted Data Policy and initial settings

AgReFed Trusted Repository Policy

1. The AgReFed Trusted Repository Self-assessment process will be used as the AgReFed alignment process
2. Assessment (scope and requirements) may be modified by the AgReFed TC
3. Current thresholds for qualification are proposed as shown in the table on this page. If R2, R3, R4, R11 and R13 to R16 requirements are met, the repository qualifies as an AgReFed Trusted Repository
4. An additional requirement that repository metadata must be harvestable by Research Data Australia (RDA)
5. The Federation Data Steward will review and qualify (or reject) Data Provider Communities' AgReFed Trusted Repository Self-assessment.
6. Any disputes in relation to the validation of assessment will be escalated to the AgReFed TC for review and decision
7. Qualifying threshold levels can be reset by AgReFed TC

	Requirement	AgReFed policy
R1	**Mission/Scope**: The repository has an explicit mission to provide access to and preserve data in its domain	
R2	**Licenses**: The repository maintains all applicable licenses covering data access and use and monitors compliance	Yes
R3	**Continuity of Access**: The repository has a continuity plan to ensure ongoing access to and preservation of its holdings	Yes
R4	**Confidentiality/Ethics**: The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms	Yes
R5	**Organisational Infrastructure**: The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission	
R6	**Expert Guidance**: The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant)	
R7	**Digital Object Management**: The repository guarantees the integrity and authenticity of the data	
R8	**Appraisal**: The repository accepts data and metadata based on defined criteria to ensure relevance and understand ability for data users	
R9	**Documented Storage Procedures**: The repository applies documented processes and procedures in managing archival storage of the data	
R10	**Preservation Plan**: The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way	
R11	**Data Quality**: The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations	Yes
R12	**Workflows**: Archiving takes place according to defined workflows from ingest to dissemination	
R13	**Data Discovery and Identification**: The repository enables users to discover the data and refer to them in a persistent way through proper citation	Yes
R14	**Data Reuse**: The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data	Yes
R15	**Technical infrastructure**: The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its users.	Yes
R16	**Security**: The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users	Yes

Minimum metadata elements for collection records

Minimum metadata requirements for AgReFed Collection Descriptions in Research Data Australia, to meet AgReFed FAIR Data Policy (Q3).
Fields based on RIF-CS schema (as used by Research Data Australia).
See https://documentation.ands.org.au/display/DOC/Collection for more information on each field.

Information type (field)	Meaning
Metadata publisher	The organisation that is contributing the metadata record
Identifier	A unique identifier for the resource, i.e. DOI
Metadata source	The primary/authoritative source of truth for the metadata record, as represented by a URI.
Collection Type	The type of collection being described, i.e. collection, dataset, software, etc
Title	The name or title of the collection, should be descriptive and unique, avoid acronyms.
Parties	A related person or organisation linked to the collection (include ORCID if possible) e.g. creator, owner, manager.
Location	Online location (DOI, Handle or URL) of the **metadata record** OR to **download** the resource
Related Service	Include a link to the AgReFed portal RDA record (workflow TBA); or to other Services.
Citation	The preferred form for citing a collection to enable data to be referenced.
Access Rights	Collection access conditions. Specify one of: **open**, **conditional** or **restricted**.
Licence	License conditions associated with the collection; a standard licence, e.g. creative commons is preferred.
Description	A summary description of the collection. Provide sufficient information to enable a user to assess suitability of the data for reuse for their purpose.
Subject	Keywords or terms to describe the topic of the resource. Include at least one ANZSRC-FOR code. Additionally, AGRIVOC terms should be used.
Spatial coverage (if relevant)	The geometry for the location the resource relates to.
Temporal coverage (if relevant)	The time period the resource relates to, in W3C Date/Time Format.

Minimum metadata elements for service records

Minimum metadata requirements for AgReFed Collection Descriptions in Research Data Australia, to meet AgReFed FAIR Data Policy (Q3). Fields based on RIF-CS schema (as used by Research Data Australia).

See https://documentation.ands.org.au/display/DOC/Collection for more information on each field.

Information type (field)	Meaning
Metadata publisher	The organisation that is contributing the metadata record
Identifier	A unique identifier for the resource, i.e. DOI
Metadata source	The primary/authoritative source of truth for the metadata record, as represented by a URI.
Service Type	The type of service being described, from this list.
Service name	The name or title of the service, should be descriptive and unique, avoid acronyms.
Parties related to this service	A related person or organisation linked to the service (include ORCID if possible) e.g. owner, manager.
Service location	An electronic address (e.g. access URL) where the service may be accessed.
Related Collections	All collections that are related to, or may be accessed by, the AgReFed portal.
Access Rights	Service access conditions. Specify one of: **open**, **conditional** or **restricted**.
Description	A summary description of the collection. Provide sufficient information to enable a user to assess suitability of the data for reuse for their purpose.
Subject	Keywords or terms to describe the research focus of the service. Include at least one ANZSRC-FOR code. Additionally, AGRIVOC terms should be used.
Spatial coverage*	The geometry for the location the resource relates to (a point or a polygon).
Temporal coverage*	The time period the resource relates to, in W3C Date/Time Format.
Related information*	Related resources such as publications (via DOIs), websites (via URLs), funding info, etc


AgReFed
Agricultural Research
Federation