6.5 Data catalogs

Choosing an approach to structure the inventory of your data assets

Why should I do this?

To provide a central accessible repository of metadata and information about the various data assets, thus making it easy for users to discover, understand and effectively use the available data resources.

How do data catalogs relate to FAIR?

Findability: Publishing a granular metadata schema corresponding to your dataset can enhance its findability on dataset search engines.

Interoperability: Cataloging can increase semantic interoperability by building a coherent administrative structure around your data that can work alongside the CDM for easier collaboration between team members and synergy with other projects.

Reusability: By enhancing findability, cataloging methods innately also enhance reusability, by making it easy for others to use your dataset in their work.

Download this data catalogs factsheet for more insights.

What is a data catalog?

A data catalog is a structured inventory of data assets within an organization or for a particular project.

1) If you are a Program Officer (PO), you may want to share this page directly with your grantee, so they can act on it.

2) If you are a grantee, ensure you have technical team members involved in this process. While the content is accessible to both technical and non-technical members, technical expertise will be required to make decisions for the investment in this step.

3) If you have not already downloaded 'Project SIS' or 'Waterways', the illustrative scenarios provide examples on how each theme is navigated. These scenarios are frequently referred to across the content in Step 6 to help you understand how different aspects within a theme are applied.

Things to consider for your investment:

Determine the scope

At this stage, the lead grantee needs to be able to identify where the data is coming from, the format it is in, and what it consists of. In addition, the lead grantee needs to have a good idea of the likely journey that the data will go through, asking questions such as:

Does all the data originate with this project?

Is some data being collected from other projects?

Is some data being purchased or sourced from elsewhere?

This will allow the lead grantee to plan how the data and descriptive metadata will be captured.

Decide on a schema for descriptive metadata

The schema is a set of metadata properties that will be used primarily to ensure that the dataset is findable by those that need to access it. This is also called descriptive metadata, and if the lead grantee is not quickly able to put a schema together, the best alternative is to borrow a schema from elsewhere.

Building your own schema means identifying the key metadata properties that will be used during the course of the project. This can be an iterative process and will probably involve most of the project team, particularly the project manager, the project officer, and the research scientists.

DCAT is a vocabulary designed to facilitate finding metadata. It is a set of the words that are likely to be used to describe the data held inside a data catalog (i.e., metadata registry and repository). This is a good place to start in selecting particular properties that will be used to describe the dataset. A lead grantee could simply use DCAT, or utilize it as a starting point for building a schema.

Some example properties that maybe attached to each dataset are:

Title

Description

Access rights

Contact point

Date of first release

Date of latest release

All of these can be found in DCAT. One may want to include properties from other vocabularies or ontologies, such as 'was attributed to', which is defined in PROV, or simply 'publisher', which is defined in Dublin Core. By deriving and referencing metadata terms and definitions from existing schemas, vocabularies and ontologies one is making the metadata and data more interoperable and reusable.

In some cases, one may want to add properties that are not found in these common reference sources, or one may wish to change aspects of their definition. One would expect to arrive at a schema of 20-50 descriptive elements, divided into sections, with the schema defined in a formal standard format, such as the JSON schema shown below:

```
{
  "$schema": "http://json-schema.org/draft-07/schema",
  "$id": "https://raw.githubusercontent.com/USER/schemata/schema/dataset/dataset.schema
  "type": "object",
  "title": "HDR UK Dataset",
  "description": "SAMPLE Dataset Schema",
  "version": "2.0.1",
  "required": [
    "identifier",
    "version",
    "revisions",
    "issued",
    "modified",
    "summary",
    "accessibility"
  ],
  "additionalProperties": false,
  "properties": {
    "identifier": {
      "$id": "#/properties/identifier",
      "title": "Dataset identifier",
      "$comment": "http://purl.org/dc/terms/identifier ",
      "examples": [
        [
          "226fb3f1-4471-400a-8c39-2b66d46a39b6"
        ]
      ],
      "description": "System dataset identifier",
      "anyOf": [
        {
          "$ref": "#/definitions/uuidv4"
        },
        {
```

```
        {
          "type": "string",
          "format": "uri"
        }
      ]
    },
    "version": {
      "$id": "#/properties/version",
      "title": "Dataset Version",
      "description": "Dataset metadata version",
      "$ref": "#/definitions/semver",
      "examples": [
        "1.1.0"
      ]
    },
    "revisions": {
      "$id": "#/properties/revisions",
      "title": "Dataset Revisions",
      "description": "Revisions of Dataset metadata",
      "type": "array",
      "items": [
        {
          "$ref": "#/definitions/revision"
        }
      ]
    },
    "issued": {
      "$id": "#/properties/issued",
      "title": "Creation Date",
      "$comment": "dcat:issued",
      "description": "Dataset Metadata Creation Date",
      "type": "string",
      "format": "date-time"
```

```
    },
    "modified": {
      "$id": "#/properties/modified",
      "title": "Modification Date",
      "$comment": "dcat:modified",
      "description": "Dataset Metadata Creation Date",
      "type": "string",
      "format": "date-time"
    },
    "summary": {
      "$id": "#/properties/summary",
      "title": "Summary",
      "description": "Summary metadata must be completed by Data Custodians onboarding r
      "$ref": "#/definitions/summary"
    },
    "documentation": {
      "$id": "#/properties/documentation",
      "title": "Documentation",
      "description": "Documentation can include a rich text description of the dataset (
      "$ref": "#/definitions/documentation"
    },
    "coverage": {
      "$id": "#/properties/coverage",
      "title": "Coverage",
      "description": "This information includes attributes for geographical and tempora
      "$ref": "#/definitions/coverage"
    },
    "provenance": {
      "$id": "#/properties/provenance",
      "title": "Provenance",
      "description": "Provenance information allows researchers to understand data with
      "$ref": "#/definitions/provenance"
    },
```

A schema could be defined in another suitable language, such as XML Schema or OWL, or it could simply be written in a word document, although this latter option is not practical for machine reading. Using a specific language such as JSON Schema allows the ambiguity present in natural language to be eliminated. The key metadata is then available in a machine readable format, which can be read into the data catalog, or at least translated in an unambiguous fashion.

It is expected that some of the descriptive metadata will be common to many contexts, while some may be context-sensitive, and defined with a subject area in mind. Most contexts already have sample metadata schemas mapped out that can be used as a basis for the descriptive metadata schema.

Onboarding metadata is discussed in Project SIS and in Waterways.

In Project SIS, we have a basic metadata schema that in some data catalogs, such as MDW Exchange (see Figure 3.1), will be displayed as a dashboard, while in others, such as the OSF repository, it will be displayed in different areas (see Figure 3.2).

First see the descriptive metadata, assembled in an excel spreadsheet:

ReleaseDate	08/02/2024
EffectiveFrom	15/02/2024
AppliesTo	CHN
ConformanceDate	29/02/2024
Publisher Name	CABI
Description	https://www.cabi.org
Stage	alpha
DataStatus	Secured
Title	CABI-Handheld
Abstract	This provides one core input data sample for the Scenario A Sample Project.
Contact Point	projectManager@cabi.org
Keywords	["Soil Lat Lon ECa"]
License	https://creativecommons.org/licenses/by-nc/4.0/

Figure 4.1

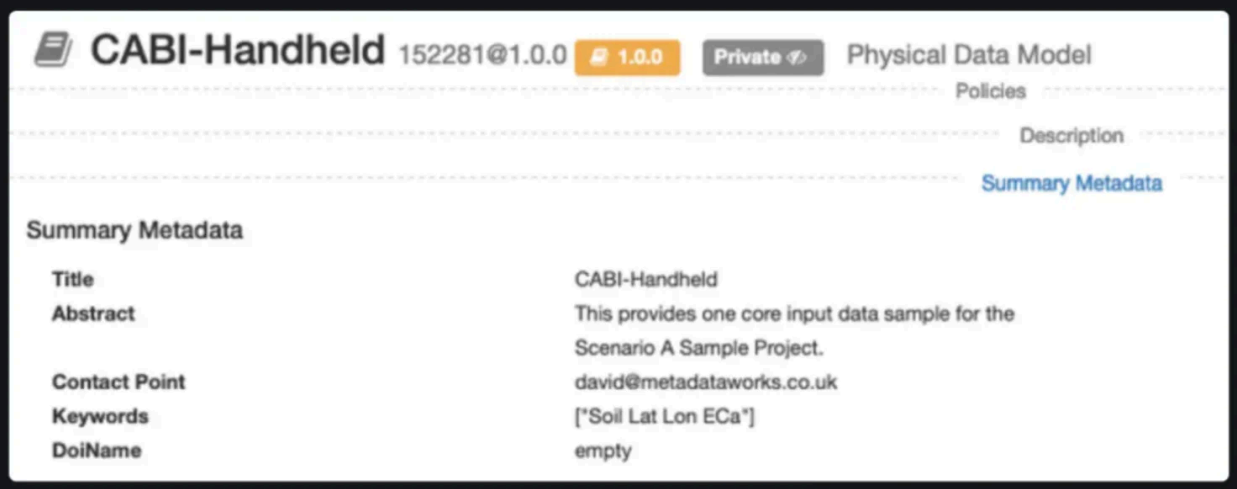


Figure 4.2

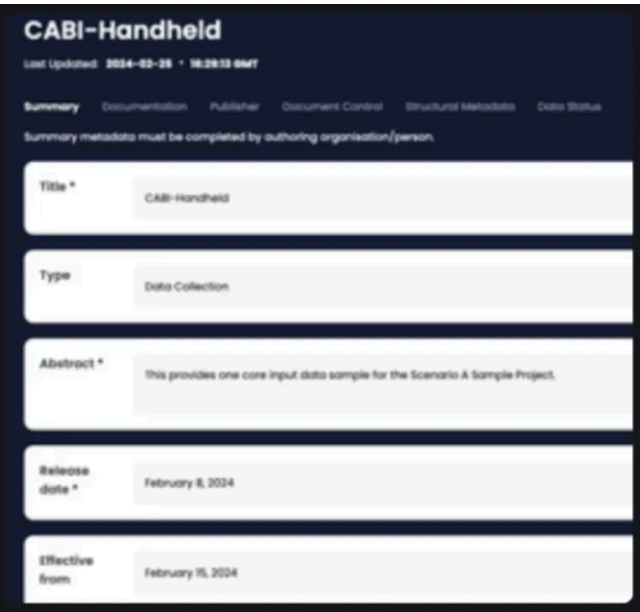


Figure 4.3

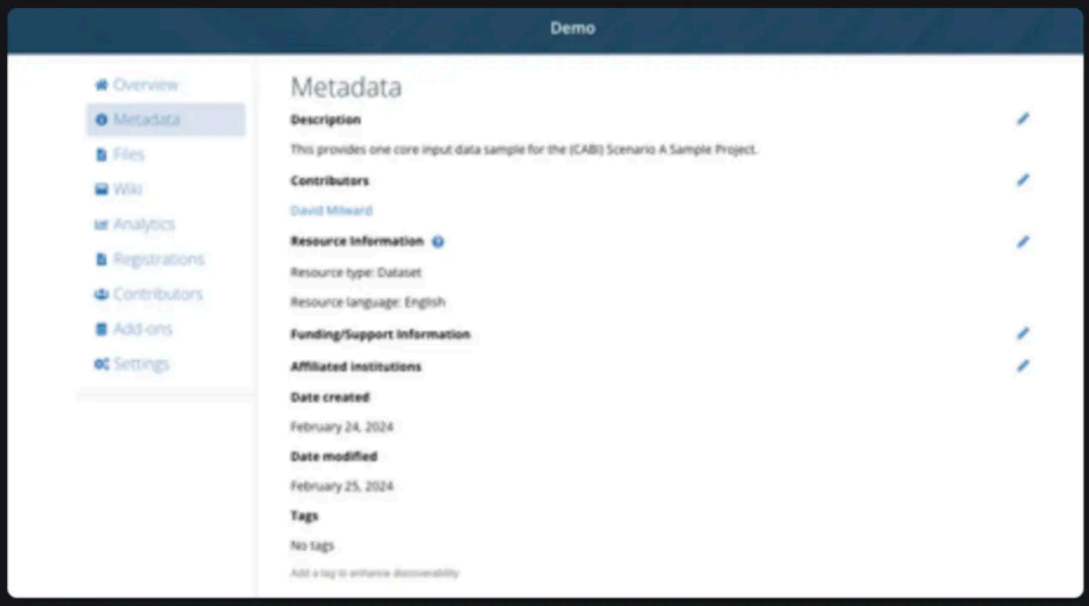


Figure 4.4

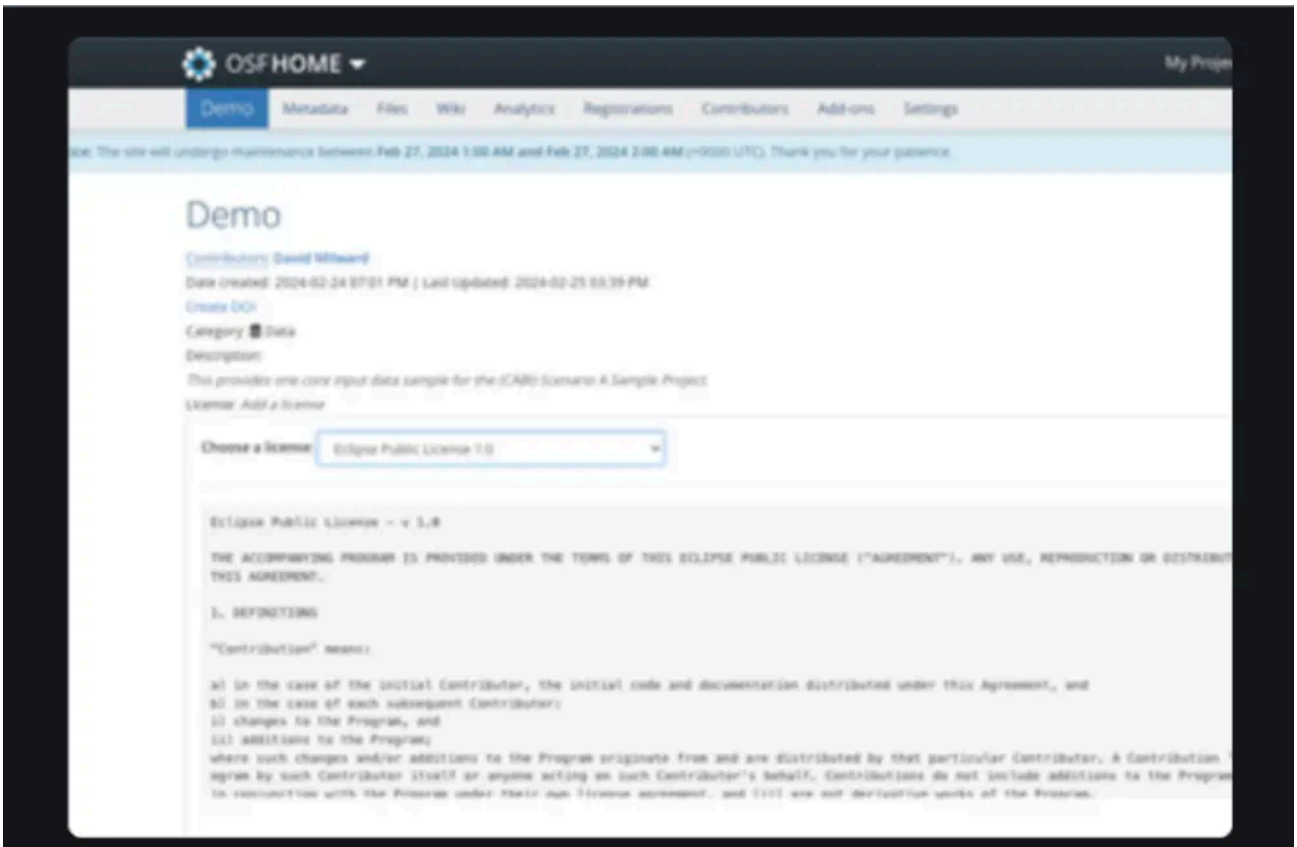


Figure 4.5

Decide on a vocabulary

The goal of this step is to ensure that each data element is named in an unambiguous way, and is ideally linked to common open vocabulary. This may not be possible with data obtained from third-party sources, but it is possible for any data generated in the project. It allows the process of collecting and then cataloging data to be clearer and better defined, which in turn improves interoperability, communication and data mapping.

As a first example, we can see that DCAT is a vocabulary for describing data catalogs. It is a set of words that are used in this context. In the same way, the context of any project using agricultural data can be best described using a vocabulary of terms relating to the context. Each dataset should only use terms that come from the controlled vocabulary. That way, data being collected must be described in a particular way.

For instance, if one is collecting geographical data, then it may be sensible to restrict users to using terms found in the Getty Thesaurus of Geographic Names (TGN), which removes any ambiguity in describing a particular region. Alternatively, one can build up a vocabulary for use with the project, using definitions taken from online terminologies such as these. The easiest way to

start would be to select an available online vocabulary, and then add as a base conceptual model to the catalog, so that when data elements are needed, they can be built using this vocabulary. A selection of open vocabularies is given below:

1. Agronomy Ontology (AgrO)
AgrO describes the agronomic practices, techniques and variables used in agronomic experiments.

2. AgroPortal
AgroPortal is a repository of agricultural ontologies and controlled vocabularies. It provides a platform for accessing, annotating and aligning ontologies related to agronomy and agriculture.

3. Agricultural Thesaurus and Glossary (AGROVOC)
AGROVOC is a multilingual thesaurus covering agriculture, forestry, fisheries and related domains. It is developed and maintained by the Food and Agriculture Organization (FAO) of the United Nations.

4. Plant Ontology (Planteome)
Planteome is a controlled vocabulary and ontology that provides standardized terms for describing plant structures, growth stages, and developmental processes.

5. Environment and Sustainability Ontology (EnvO)
EnvO is an ontology that covers environmental and habitat-related concepts. It is relevant to agriculture for describing the environmental context of farming practices.

6. Crop Ontology
Crop Ontology is designed to standardize and integrate crop-related information, including traits, germplasm, and experimental conditions.

7. Semantic Sensor Network Ontology (SSN)
SSN is an ontology for describing sensors, observations, and measurement processes. It is applicable in agriculture for modeling sensor data related to environmental monitoring.

The choice of vocabulary may limit and restrict the choice of data catalog. Ideally, freedom to choose or import the ontology of choice will make the job of

defining the data heading (i.e., the data elements) much easier, and will enable automatic interoperability.

Some data catalogs will already have ontologies built in, such as the agronomy ontology, either in a fixed form or as a base so that other definitions can be added. Built-in vocabularies and flexible vocabularies make the process of building a vocabulary much easier.

Identify classifications

This step involves considering which datasets and data elements should be available to whom. Are there data elements that need to be private and not public? Should some people in your team only have access to some of the data? Start by making a list of different levels of access, and then assign people or roles to each of the different levels. Also, different datasets can be added to different groups, and you may want to assign different groups to different types of dataset; for example, perhaps only some members of the team will handle the third-party data.

Typically, one would use a sensitivity-based scheme for the metadata in the data catalog such as the one shown below:

> Public: Freely accessible by anyone. Descriptive metadata (as defined in the schema) will be available, whereas other aspects, including the raw data, will probably require access control.
>
> Internal: Accessible only within an organization or specific group. Metadata might disclose internal processes or personnel information.
>
> Confidential: Requires authorized access, due to its sensitive nature. Metadata could reveal proprietary information or trade secrets.
>
> Restricted: Highly limited access due to criticality. Metadata could expose security or personal data.

In many cases, only two or three of these levels will be needed, although in certain circumstances, one may want to add regulatory aspects, especially if dealing with data falling under a specific framework, such as GDPR or HIPAA. It is also key to consider data lineage, especially when handling data from third parties that may have additional licensing imposed on any merged datasets.

If the data is available under a public, share-all type license, one may make it available in an open repository. If not, one will need to add in a role-based access security model.

Choose a data catalog

Key decisions are fairly complex and are based around the project itself.

As a guide, consider the issues below:
1. Does the data catalog handle all the descriptive metadata in my schema?
2. Can the data catalog relate different data elements (fields) to a common data model?
3. Can the data catalog relate different data elements to a common vocabulary or ontology?
4. Can the data catalog record both structural and descriptive metadata?
5. Can the data catalog provide, or link to, a DOI registry?
6. Does the data catalog have a good mechanism for tagging data elements and searching for them?
7. Does the data catalog have a good mechanism for accessing the data and metadata using an API?
8. Does the data catalog have a role-based security model?

We briefly examine the OSF data repository and the Global Agricultural Research Data Innovation Acceleration Network (GARDIAN) in Step 6.4, since they are widely used in the sector, with some references to the Metadata Exchange (MDW), a catalog used in the UK medical and government sector.

The following comprises a quick guide to how these different aspects are handled in each of the three data catalogs.

Issue	GARDIAN	OSF	MDW
Can the DC handle any user-defined descriptive schema?	No, a standard metadata schema is provided, based on Agronomy ontology	No, a minimal schema is provided although further items can be added in WIKI	Yes, client-based schema or standards based schema can be used
Can fields (data elements) be related to a standard?	Yes, but only the Agronomy standard	Not as far as we are aware	Yes, fields can be related to any standard or data model
Can fields (data elements) be related to a common vocabulary or ontology?	Yes, but only the Agronomy ontology	Not as far as we are aware	Yes, fields can be related to any ontology or vocabulary
Can the data catalog record both structural and descriptive metadata?	Not fully	No	Yes
Can the data catalog provide or link to a DOI registry?	Yes	Yes	Yes if configured
Does the data catalog have a good mechanism for tagging data elements and searching for them?	Not yet examined	Some tagging available	Very comprehensive search, filtering and tagging
Does the data catalog have a good mechanism for accessing the data and metadata using an API?	Not yet examined	Yes, out of the box	Yes, but needs extra configuration
Does the data catalog have a role-based security model?	???	Security available but not RBA	Full Role based access available

Set up a data catalog or data repository

Data Catalogs differ enormously in their features, technology set, and the way they are set up, as the previous section shows. However, there are some key points to bear in mind when setting up a data catalog and/or data repository. Most have been covered in the previous sections, but are repeated below.

Data content

Develop a common data model (CDM): Define the structure and meaning of your data using a well-documented CDM. This ensures consistency, facilitates data sharing, and enables easier analysis across different experiments.

Adhere to existing data standards: Consider incorporating established data standards relevant to agriculture, such as the Agronomy Ontology or SensorML. This enhances interoperability and exchange with other research projects.

Metadata management

Capture comprehensive metadata: Describe your data comprehensively using relevant metadata standards like the Dublin Core or the Agricultural Data Commons (ADC) metadata schema. This makes your data discoverable and interpretable by others.

Standardize metadata: Define clear guidelines and templates, such as a separate metadata schema for capturing metadata to ensure consistency and completeness across datasets.

Data organization and access

Logical organization: Organize data logically, considering factors like experiment type, crop variety, location and date. This facilitates easier browsing and retrieval.

Version control: Implement version control to track changes made to data and ensure reproducibility of research findings.

Access control: Define appropriate access levels to ensure data security and privacy while enabling controlled sharing with collaborators or the broader research community.

Documentation and user support

Clear documentation: Provide clear documentation explaining the data catalog/repository, including:

The purpose and scope of the data collection.

Description of the data model and any applied standards.

Instructions on accessing and using the data.

Sustainability and interoperability

Long-term sustainability: Plan for the long-term sustainability of your data catalog/repository, including considerations for hardware, software, and personnel resources to maintain and update it.

Interoperability: Ensure your data catalog/repository is interoperable with other agricultural research platforms or data portals. This enables broader data sharing and collaboration within the research community.

Additional considerations

Security: Implement robust security measures to protect sensitive data, such as user authentication, authorization and encryption.

Compliance with regulations: Ensure compliance with any relevant regulations regarding data privacy and biosecurity, especially when dealing with sensitive research data.

Illustrative scenarios

Overview

©Gates Archive/Mansi Midha

Refer to the illustrative scenario that you have downloaded to see how this has been considered.

Ensure any work notes or decisions taken are being documented as this would be useful to refer to at later stages or for someone new joining the team.

Project SIS

Only the specific theme related content has been highlighted here. To get a feel for the scenario, read here.

1. Data onboarding

As per the CDM, the input datasets will be .csv files. To ensure efficient findability and accessibility, they will each be stored in a folder with metadata in a JSON. We have designed a descriptive metadata schema for this purpose, making all further steps of the project easier by providing easy inventory and discovery of datasets within the centralized repository. This schema will be carried across the CDM for all datasets.

Metadata Keyword	Description
CreationDate	The time the dataset was first created
EffectiveFrom	The time of the earliest data point in the dataset
LastModified	The time the dataset was modified last
AppliesTo	The project the dataset applies to (Project SIS)
Source	Who the data is coming from
Coverage	The geographical region(s) or subregion(s) that the dataset covers
Type	The type of data (soil, meteorological, interview)
State	The state of the dataset, i.e. whether it is raw, clean, processed/linked, inspected, or final
Listed	Whether the dataset is listed publicly (after the SIS is released as a product)
Title	The name of the dataset
Abstract	A one-sentence description of the dataset
Contact	The email address of someone responsible for this dataset
Columns	A list of the column names of the dataset
License	The license the dataset has for use
ProjectURL	The URL of the project, for a user to find more information about data collection strategy

Table 1 - The Descriptive Metadata Schema

2. Data processing

Once the datasets are updated, their metadata will require updating too. The 'state' will be 'clean' or 'processed', and the column names will be different, as well as the modification date.

3. Data enrichment

The metadata for the final dataset will need to be updated.

4. Data products

The metadata will be audited to ensure accuracy and useability for potential users of SIS. This also involves checking that metadata is easy to discover and read on the platform that SIS is hosted on. Prior to release, we might be able to bring together a focus group of researchers to understand if the final metadata schema needs any modification.

Waterways

Only the specific theme related content has been highlighted here. To get a feel for the scenario, read here.

1. Data onboarding

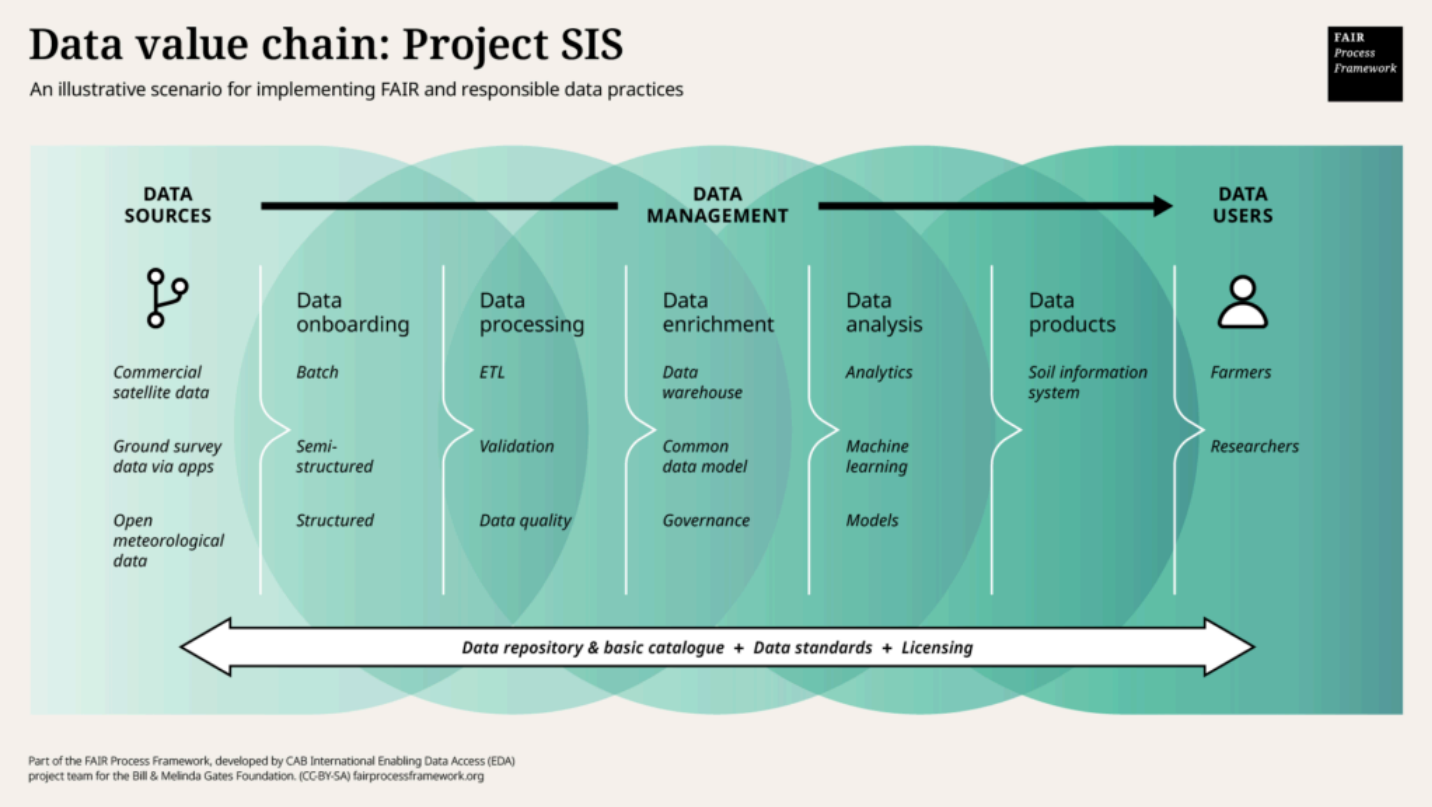
We will not build a specific metadata schema for this project as there are only four data sources, none of which will be published as part of the final outputs. Nonetheless, in our repository we will keep a diary and readme for ease of use and potential onboarding of new researchers.

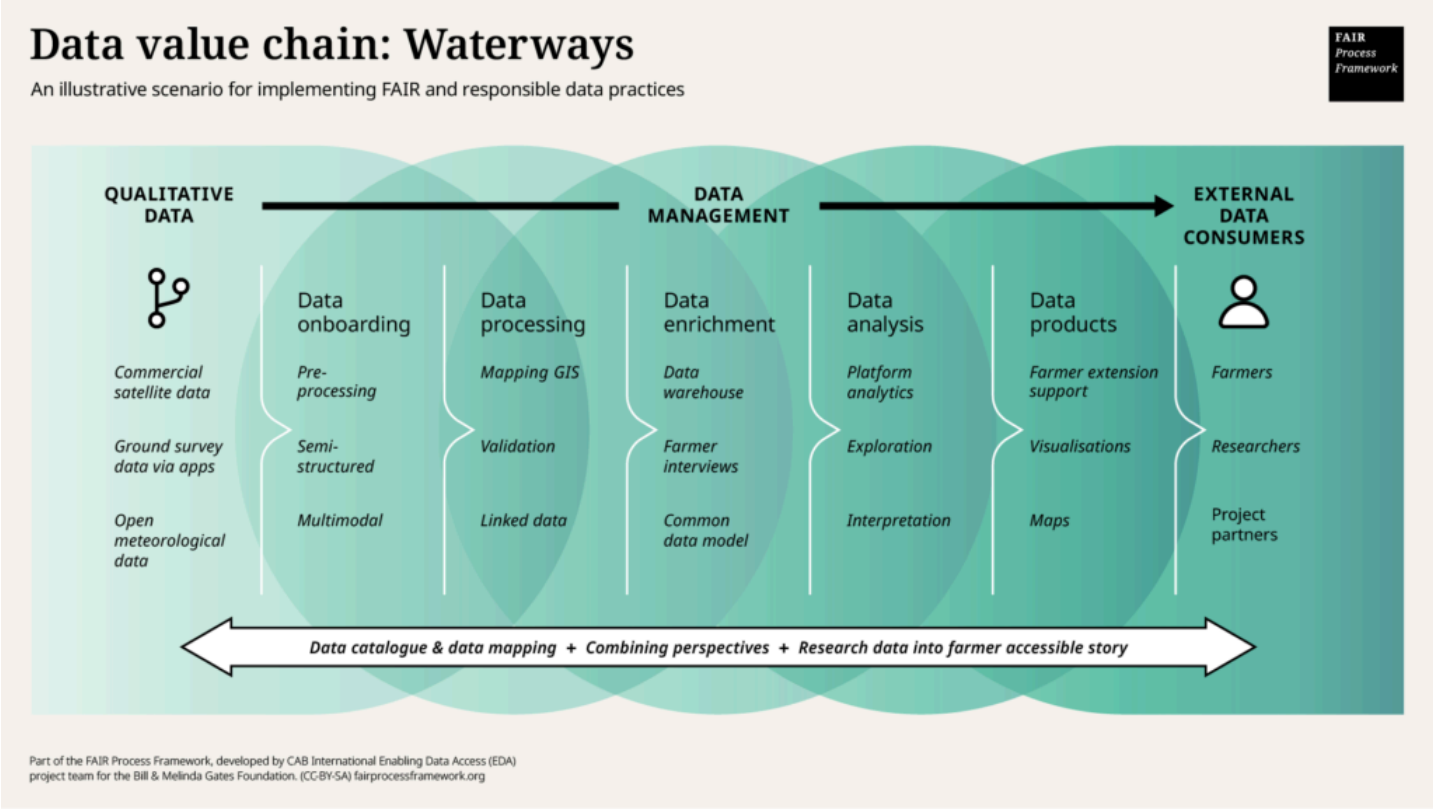
2. Data enrichment

Cataloging interviews and attaching data to them is extremely important, to provide context. The simple recording of details like the date of the interview, the location of the farm, and how long the farmer has been working in Waterways, builds a bigger picture for the interview content to sit in. This therefore enriches the interview data and makes it easier to use in the topographical map and as a starting point for empirical analysis. Cataloging will also enable further analysis of the interviews with perhaps textual analysis techniques.

The theme of data catalogs can be important at different stages of your project, whether or not you expect that to be the case. To help you incorporate them into your project planning, this section provides suggestions about where you should think about the theme, structured using the stages from the Data Value Chain (DVC).

The DVC is a way of viewing the process of running a project from the point of view of the data, thereby identifying how it is onboarded, processed, enriched, analyzed and released in a product. In doing so, the DVC shows the moving parts in project implementations, making it a useful framework regarding the general steps of any project working with data.





FAIR data principles ensure that data is shared in a manner that enhances reuse by both humans and machines, fostering further research and innovation.

Learn more

Acknowledgements

FAQs

FAIR Process Framework has been developed by the Enabling Data Access (EDA) project team at CABI and is funded by the Bill & Melinda Gates Foundation to support the foundation's Open Access Policy. The FAIR Process Framework is a tool to assist partners in developing data access and management plans (DMAPs) that incorporate FAIR and responsible data practices. Except where otherwise noted, the content on this website is licensed under a Creative Commons Attribution 4.0 International License.