6.7 Data privacy and security

Choosing an approach to structure the inventory of your data assets

Why should I do this?

To protect personal information and to prevent unauthorized access or misuse.

How does data privacy and security relate to FAIR?

Accessibility: Login systems work to build accessibility, where authorized people can securely work while data privacy constraints are well-respected, therefore being inclusive of stakeholders' privacy requirements. It is important to recognize that accessibility does not mean 'opening datasets to all', but rather 'opening datasets to the right people'.

Reusability: If your dataset has been cleared of all sensitive data, therefore adhering to privacy and security requirements, only then can it be released as part of a product for reuse. If privacy and security is not considered when releasing your product, you are personally liable for any resulting damage.

Download this data privacy and security factsheet for more insights.

What is data privacy and security?

Data privacy and security refers to the protection of personal information, ensuring that sensitive data is handled, stored and processed securely to prevent unauthorized access or misuse. It encompasses a range of principles and practices that safeguard individuals' privacy in the digital age.

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 of access

The lead grantee needs to be able to identify the necessary access levels of individual users to perform their project roles, and to provide users with appropriate access to delimit their particular responsibility regarding the data.

What the policymaker, scientist or farmer needs to see of agriculture data is determined by their relationship to the actual 'on-the-ground' work. Login and security is more than deciding who can access the data and who cannot. It also allows each user to see the data, analysis and results through their particular window of involvement.

The lead grantee must be able to answer the following overall questions:

- Who has responsibility for onboarding the data, providing particular datasets (without needing access to other datasets)?
- Who has responsibility for accessing all the data in order to process, enrich and analyze it?
- Who are the data consumers, and what is their requirement in accessing the data, to immediately understand the implications of the research for their own field of expertise?

This process will allow the lead grantee to plan how the login and security will be set up.

Establish a simple login system

To avoid requiring multiple logins and security steps across different systems holding different parts of the data, it is vital to hold the data centrally, so that users navigate a single login portal giving them rights to edit or access the data, according to their roles. There is always a balance to be struck between requiring the user to log in as a time-consuming process of remembering passwords, and protecting data. Using ESRI as a GIS interface in Waterways allows users to be given access to edit particular maps or datasets, while the output can be publicly visible or restricted to invited users.

A good system will draw a line between the datasets that need protection to be accessed and edited, and the output that is freely available for farmers and other data consumers to access.

Assess the users and the data

This can be done in a variety of ways. However, a simple spreadsheet along the lines illustrated here may help:

Users	Datasets	Content	Ownership	Licenses

Create a table of users and data access permissions

By making a simple table of roles through the lifecycle of the data from onboarding to analysis to use, the lead grantee is able to better oversee how the data lives in different stages by allowing different modes of access.

Users	Stage	Access
Data providers	Onboarding data	Entering data into the system
Data analysers	Processing, enrichment and analysis of data	Comprehensive permissions to edit and create overlays of all entered data
Data consumers	Data product	Specific access to output that is of direct concern to their expertise and activity.

Sharing ownership of research across different users

In summary, the lead grantee sets in place the permissions of which user can see and edit what data. They must:

- Decide who will have access to the data.
- Set a system of limited access according to the user level of permissions.
- Keep a public interface open to access the general message of the data product realized through the value chain of processing, enrichment and

analysis.

Such a system ensures a collective ownership of the research by different user groups, who are able to interact with different aspects of it.

Identify which users have which roles

Again, a spreadsheet can be used for this:

Users	Dataset	Role
John	soil-testing-0	curator
George	soil-testing-1	viewer
Paul	gis-dataset0	curator
Sally	gis-dataset1	curator
John	soil-intermediate	curator
George	gis-intermediate	viewer
Mia	soil-intermediate	curator
Mia	gis-intermediate	curator
Mia	gis-soil-analysis-0	curator
Ringo	gis-soil-analysis-1	curator

We are assigning a role of 'curator', which means that the user has permissions to read the data and to write (i.e., change the content), or we are assigning the role of 'viewer', which means that the user has permission only to read the data.

Define groups for each of the datasets, and which dataset belongs to which group

This can be defined using the previous spreadsheet, depending on the content. For instance, if there are only the first three datasets, one would not need to carry out this stage. However, if we consider the whole spreadsheet, then we may see that it is easier to define groups and add permissions to the groups rather than the individual user.

See below:

Users	Dataset	Role	Group	Permission
John	soil-testing-0	curator	soil-in	read, read-write
George	soil-testing-1	viewer	soil-in	read
Paul	soil-testing-1	curator	soil-in	read
Paul	gis-dataset0	curator	gis-in	read, read-write
Janis	gis-dataset1	curator	gis-in	read, read-write
John	soil-intermediate	curator	analysis	read, read-write
Paul	gis-intermediate	curator	analysis	read, read-write
George	gis-intermediate	viewer	analysis	read
Tina	soil-intermediate	curator	analysis	read, read-write
John	gis-intermediate	curator	analysis	read, read-write
Tina	gis-soil-analysis-0	curator	analysis	read, read-write
Ringo	gis-soil-analysis-1	curator	analysis	read, read-write

For projects with only three or four datasets and two or three users, we can use the roles systems, assigning roles to users for each dataset. However when we start getting to five to 10 datasets and more than three users, it is generally easier to assign both users and datasets to groups, and manage the privacy concerns using groups. A user can then be limited to the 'analysis' group and then they will only have access to datasets for analysis and will not be able to even see the input datasets. For example, in the table above, Tina may be set up so that they can only see or work on one type of data, such as GIS data, as is the case with Janis.

Identify any personal identifiable information

For most of the satellite and field data used, Personal Identifiable Information (PII) will not apply. However, a farmer interview in Waterways is a good example where the farmer should not be identifiable. First, this PII vulnerability should be listed in the metadata. Then the interviewer needs to anonymize the identifying farmers' names so they are not recognizable.

Identify any restrictions which need to be added to any of the datasets

There will be restrictions on some of the data used. For instance, in Waterways, if one is using ESRI GIS, then anyone who wants to edit the data will need to be an authorized user under the ESRI subscription of the organization. This authorization of users for certain datasets requires that permissions be obtained to meet the restrictions.

If using Google Maps or another free GIS system, there would be no restrictions on inviting users to edit the 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 a get feel for the scenario, read here.

1. Data onboarding

The number of stakeholders in project SIS will require a keen eye over privacy and security of data. For onboarding, SoilScience researchers will have the responsibility of bringing in TPP data, but on-the-ground data collection is spread between the GSS teams and the project partners. Security of data was a concern raised during the initial proposal of the project, given the vulnerable status of farmers in Dataland. Malicious actors that came across their interviews or soil sample test results could potentially take advantage of the farmers with fraudulent schemes or predatory competition. To avoid this problem, only PPs and authorized staff at the MOA will be able to onboard (raw) data collected to the project's centralized repository, with further responsibility for ensuring deletion of any data stored in unsecured, local devices (like those used to record interviews) and keeping oversight on GSS and farmers.
This system will be governed by a simple login system. As per the CDM, the centralized repository data will be onboarded to is a private Github repository, with access only provided to SoilScience, PPs, and authorized staff at the MOA. SoilScience will audit the work of all parties in data collection to ensure privacy and security.

2. Data processing

All PII will be removed from the data received from Dataland, as recommended in the data privacy and security workbook. This includes deletion of names, contact details, and specific location

(although latitude and longitude will be reserved for dataset linkage in the next stage before deletion). For the sake of version control, this processing is not done by overwriting the raw data sources, but by making a 'clean' copy of each one. However, after the processing is complete, the raw data sources will be deleted from the repository.

3. Data enrichment

After the datasets are linked, the latitude and longitude columns will have served their purpose. Although it would be nice to keep their measurements for each soil sample and its corresponding meteorological data points for reproducibility, releasing latitude and longitude in the final SIS dataset would put farmers at risk. However, to retain a level of geographic granularity, latitude and longitude will be used to categorize entries in the dataset by sub-regions in the Dataland Highlands. We will likely use sub-regions based on the data points' relation to landmark geographic features—for example, some data will be categorized as 'West side of Mt. Data'. Once these categorizations are made, latitude, longitude, and linking fields can be removed from the dataset. It is likely that we will delete these closer to time of product release so that we can validate our work and link further datasets if required.

4. Data analysis

More high-level analysis will be run on the data, as well as the entire repository, to ensure it is clean and, importantly, that farmers' details have been deleted properly. Only authorized researchers at SoilScience will be allowed to do so, although Dataland's MOA will be consulted on findings afterward.

Waterways

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

1. Data onboarding

Our main concern with regard to privacy and security is ensuring that the data is only accessible to core people in the project. Although the data is not necessarily sensitive, we would like to keep things organized and safe, especially if required by the satellite data licenses.

To do so, we are designing a login system with four levels, as suggested in the instructions.

Uploading: WRO researchers can upload their collected ground field data, and edit it if required. Similarly, Project Partners (PPs) can upload transcripts from the interviews with farmers.

Uploading and analyzing: SoilScience researchers can upload satellite data and edit it. Meanwhile, they can view and edit other data in the repository and then use all of the datasets to build the topographical map and empirical analysis.

Content management: The PO at SoilScience will be the only person with authorization to delete files in the repo and change the overall folder structure.

Viewing: Consumers of the final outputs (i.e., users of the topographical map) will only have permission to view the map.

2. Data enrichment

Building the points from latitude and longitude could potentially create a privacy risk. Although interview data will likely not be sensitive in nature, it might still put farmers in vulnerable positions. Further PII, like the farmers' names could have similar effects, and must therefore be identified. Via a contract, we will have consent from interviewed farmers to publish this information, but we believe education is far more important: farmers should know what their interviews are contributing to and, moreover, how they might be at risk if their data is online. Our project partners in Waterways are experienced in providing this sort of education.

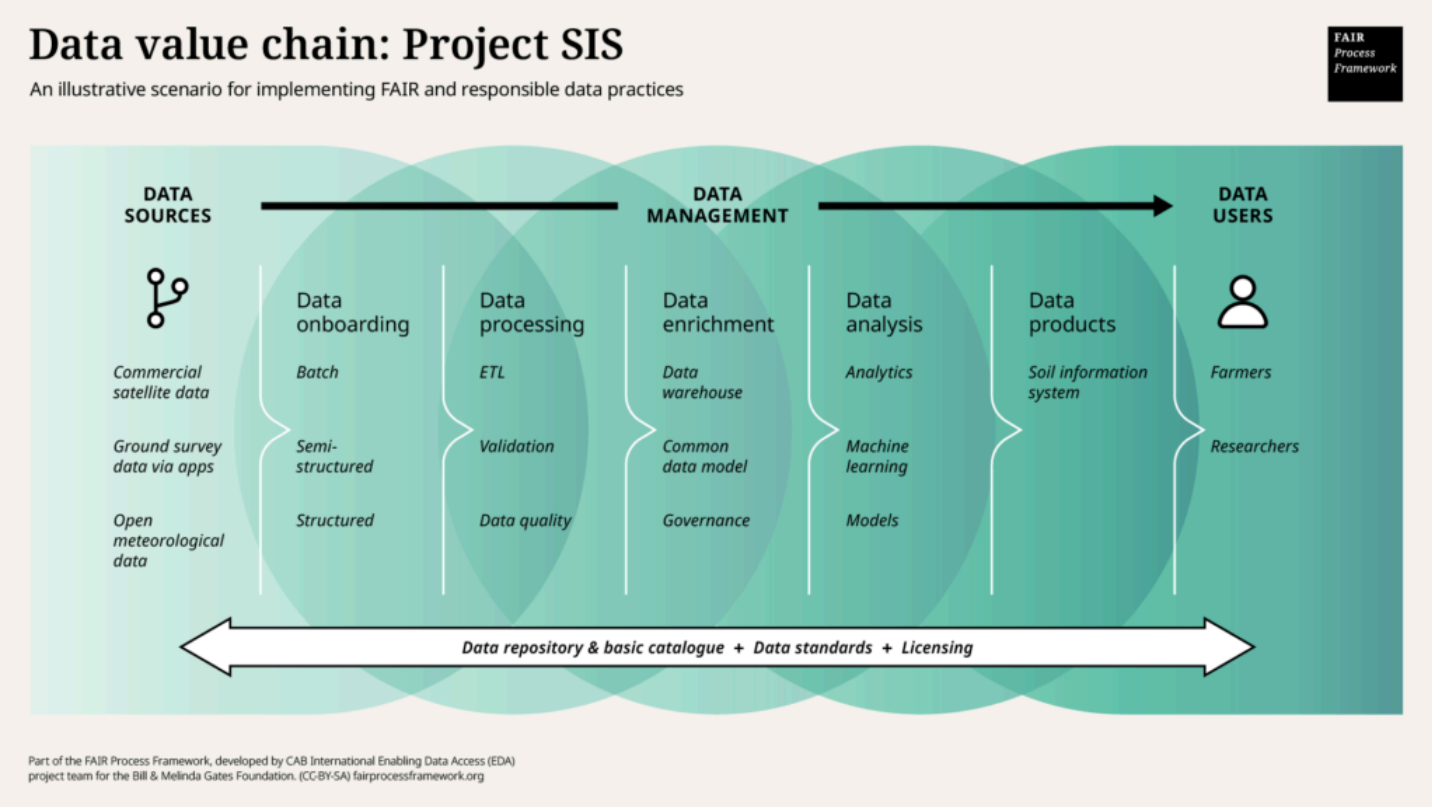
3. Data analysis

The only people authorized to perform the data analysis will be researchers at SoilScience. This will once again be governed by our login system.

The theme of data privacy and security 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).

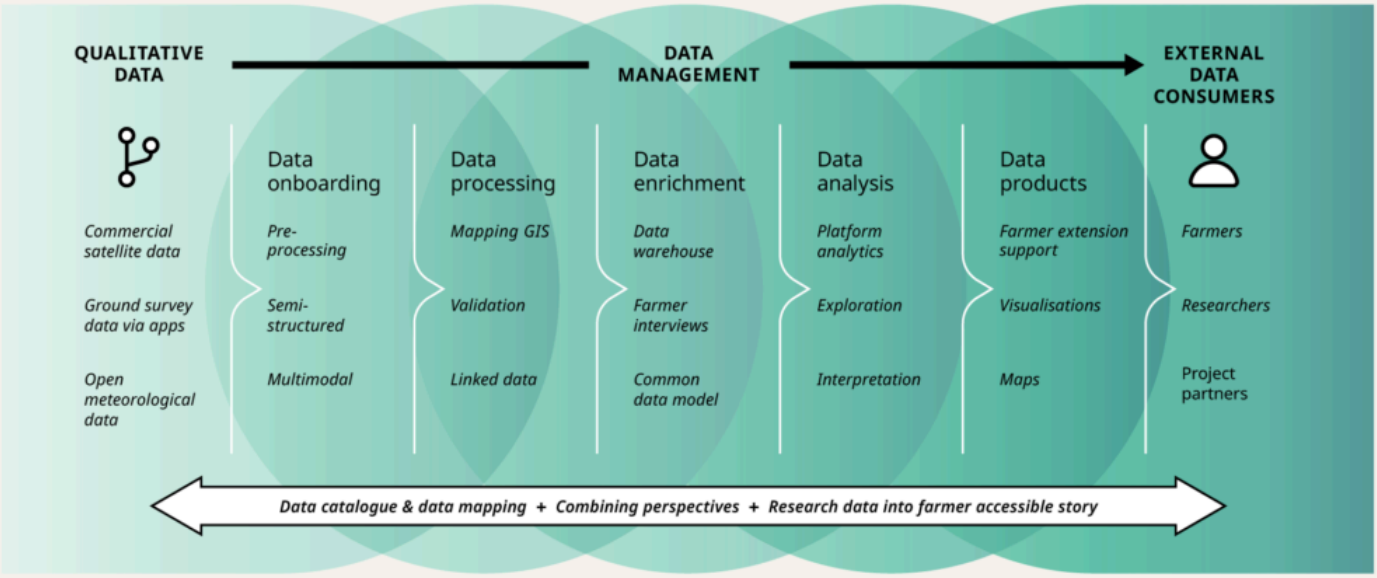
The DVC is a way of viewing the process of running a project from the point of view of the data, thereby identifying how it is onboarded, processed, enriched, analyzed and released in a product. In doing so, the DVC shows the moving parts in project implementations, making it a useful framework regarding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



Data value chain: Waterways

An illustrative scenario for implementing FAIR and responsible data practices

FAIR Process Framework



The moment you visit a farmer in her field and explain why you're here and why you would like, for example, to take a soil sample to analyze it, she has to give consent to it. If you miss out on that, you know, you might not be able to put the data out in the public.

Christian Witt, Program Officer, Bill & Melinda Gates Foundation

Learn more

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.