6.4 Data standards

Establishing conventions or rules that specify how data is structured, represented and exchanged within a particular context or domain

Why should I do this?

To ensure consistency and interoperability, enabling different systems and applications to communicate and share data effectively. Before being able to comply with data standards, there is a need for your investment to identify what is relevant for your project.

How do data standards relate to FAIR?

Interoperability: With your dataset adhering to a certain ontology, linkage with other datasets that follow the same ontology is very easy.

Reusability: Likewise, if your dataset adheres to an ontology, it is easy to integrate into other projects that might be using the same ontology, or something similar.

Download this data standards factsheet for more insights.

What are data standards?

Data standards are established conventions or rules that specify how data is structured, represented and exchanged within a particular context or domain. Data standards can cover various aspects of data, including its format, syntax, semantics and transmission protocols.

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:

Identify key stakeholders

The lead grantee must decide how data standards are relevant to this project. This will depend on the type of project.

Is the project building on previous research, and is it wanting to add an additional perspective of knowledge into an existing body of work? For instance, say that a crop is to be grown and tested under a new set of conditions to complement previous research. In this case, the Programme Manager (PM) might want the data standards to be rigorously defined and structured. A good tool in this case is the Agronomy Ontology (AgrO), where the logic and data elements are defined, and into which new research can find place in relation to past findings. In this case, it is good to start from a predetermined ontology, so that the output of the project adheres to clear data standards that allow for integration of results.

Is the project taking a more creative step open to new understandings coming from the research? Is the project seeking to examine a new question that needs fluidity to come up with answers? An example of this is Waterways, where the complexity of inputs into farmer crop yields needs a blank canvas to be investigated. This requires an openness not to fit the data and logic into an existing structure but to let the data identify the pivotal concepts in the problem. In this case, it is important that the researcher chooses a more open set of data standards, such as arranging data on a topological map, in order to draw conclusions. The data standards are still important in making sure the input from the datasets are consistent, but should not predefine the scope of the research.

In each of these cases, data standards play a critical role, requiring different handling to best fit the analysis into accessible and shareable outcomes.

Apply data standards throughout the project

Data standards help form the course of research and set the foundation on which the processing of data can be securely managed.

In the case of a very familiar context for the research, data can be fitted into an existing logic and model that immediately joins the processing and the results into a larger body of knowledge. A good tool in such cases is Global Agricultural Research Data Innovation Acceleration Network (GARDIAN), developed by CGIAR, which will guide all the steps of data manipulation in the vocabulary of AgrO as foundation. At each stage of the project, the tools, data and processing steps can be checked to make sure the experiment can interchange its results with peers. In this way, the research is at once relevant to a wider debate, and can build up knowledge across a wide range of research projects.

In the case of research seeking to make totally new understandings about a region or inquiry, the data standards need to be much looser. One does not know in

advance the logical structure into which the data will fit. The data standards are now more focused on maintaining integrity across the different data providers contributing to the project, to realize an internal consistency of data use. This is the case in Waterways, where it is only at the end of the research that the pivotal relation of water availability to crop health and ground temperature makes itself known. The data foundation is in this case the map of GIS, upon which the data comes together into a whole picture distinctive to the area.

Comply with data standards

Data standards are vital in integrating different users' ways of storing information. A well structured data standard will also be extendable if new fields have to be added, or if compared to outside datasets. However the data situation might change in the future, the data standard gives a reference that can hold the whole picture of what is happening. This creates a foundational structure into which data can be embedded, and presumes a logical form for analysis.

AgrO takes much of the headache of handling data out of the hands of the researcher. Once you make the mapping from your current research to the existing template of definitions, the associated GARDIAN toolkit can navigate the data value chain reliably and routinely. Much of the difficulty of engaging with data is incorporated into standard tools that are ready to be applied at each stage of the process. The data adheres as far as possible to standards from other work, so that the results are at once applicable and shareable with related conclusions in similar trials. The ontology naturally steers the research to comply with the data standards chosen.

If researchers want a more flexible approach to how the data interprets the situation, they need to pay more attention to complying with data standards, as mistakes can easily happen. For instance, in Waterways, the village names are listed in advance to avoid ambiguity when asking the farmer where they are from. KML (Keyhole Markup Language) is chosen as a standard format for all satellite and field data as this format is widely recognized and can be easily imported into geographical map programs. The geographical data model used has its own pre-set data standards that the project must honor.

In any research project, work with data standards must be attentive to ensure all data applies to the standards chosen. A data standard ensures that data from the producers is consistent in allowing the research to bring novel discoveries to the data product consumers.

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 processing

SoilScience research has previously used the AgrO data standard, so it is likely that we will use it again for this project. The Plant Experimental Conditional Ontology, a sub-component of AgrO, will be particularly useful for standardizing the results of tests on soil samples. We are in the process of checking if it has a suitable ontology for meteorological data, but this should not be an issue for this project, as meteorological data is likely to come from one source only (Visual Crossing).

Further processing will be done to ensure numerical data is formatted to three decimal places, incomplete entries are cleaned out of the dataset, and anomalies are identified and double-checked. The standardization processes will mostly be done with an IPython Notebook. The aggregation and averaging of meteorological data will be completed to a uniform standard in an IPython notebook, resulting in measures like 'aggregate rainfall', 'average sun per day' and 'average wind speed'—all useful measures to understand soil erosion in Dataland's highlands.

2. Data products

Within the metadata schema, the data standards will need to be made clear for users (in the 'ontology' field).

Waterways

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

1. Data processing

Satellite data can come as KMLs, shapefiles, or ordinary CSVs, whereas ground field data from WRO will likely come as a CSV. Data processing at this stage of the project will involve standardizing all files so that they can be easily integrated into a topographical map via Geographic Information System (GIS) software.

2. Data enrichment

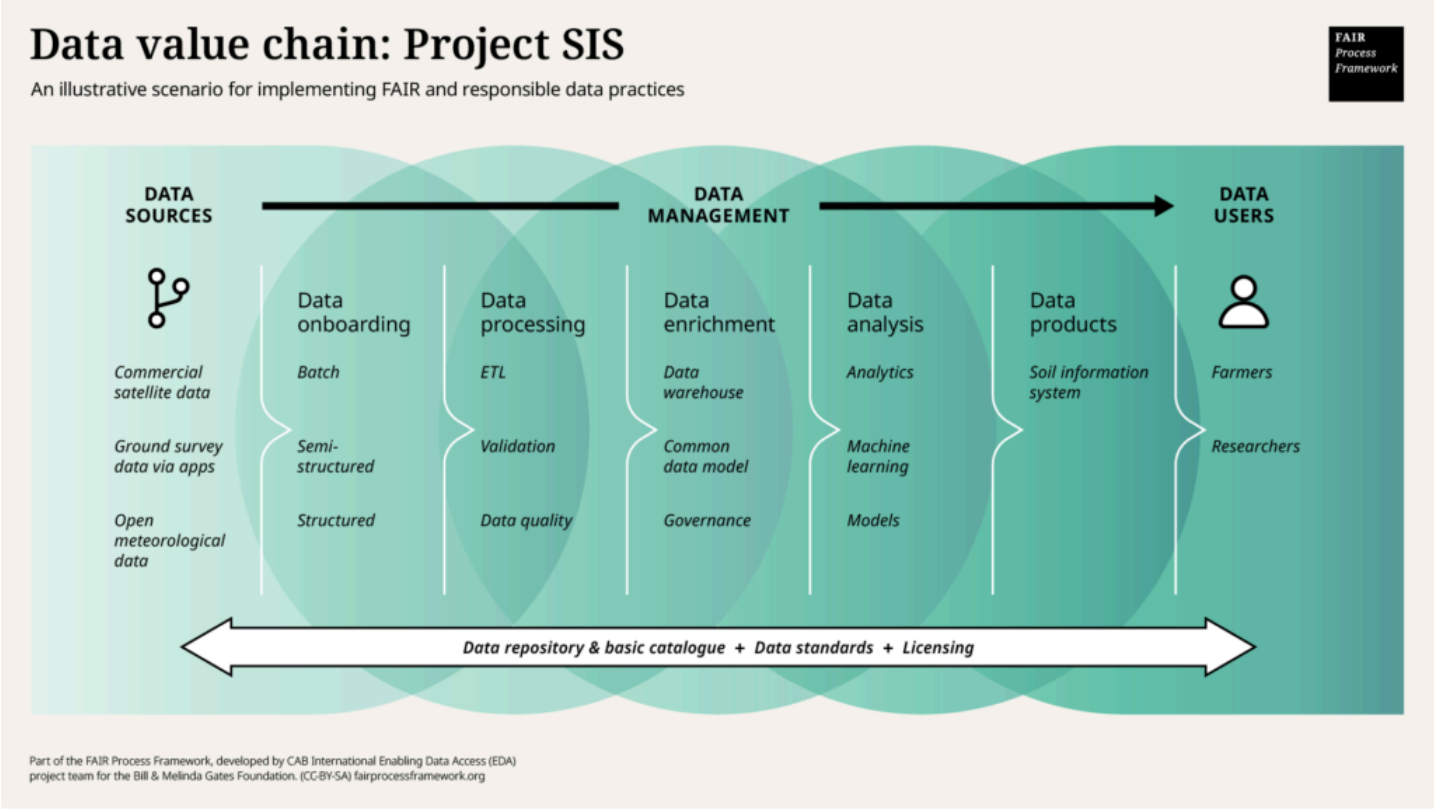
The interview transcripts will be stored in the SoilScience repository in a folder that is open to the public. Links to individual transcripts will be created and attached to the topographical map.

The means of this attachment is likely a shapefile, not a KML, though we will explore our options once we have the interviews. The interviews will not be time-differentiated, so the main strength of KMLs is not necessary, while they will be displayed as 'points' in planar geometry with which shapefiles work well. The points are simply the latitude and longitude of the center of each interviewed farmer's farm.

The theme of data standards 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).

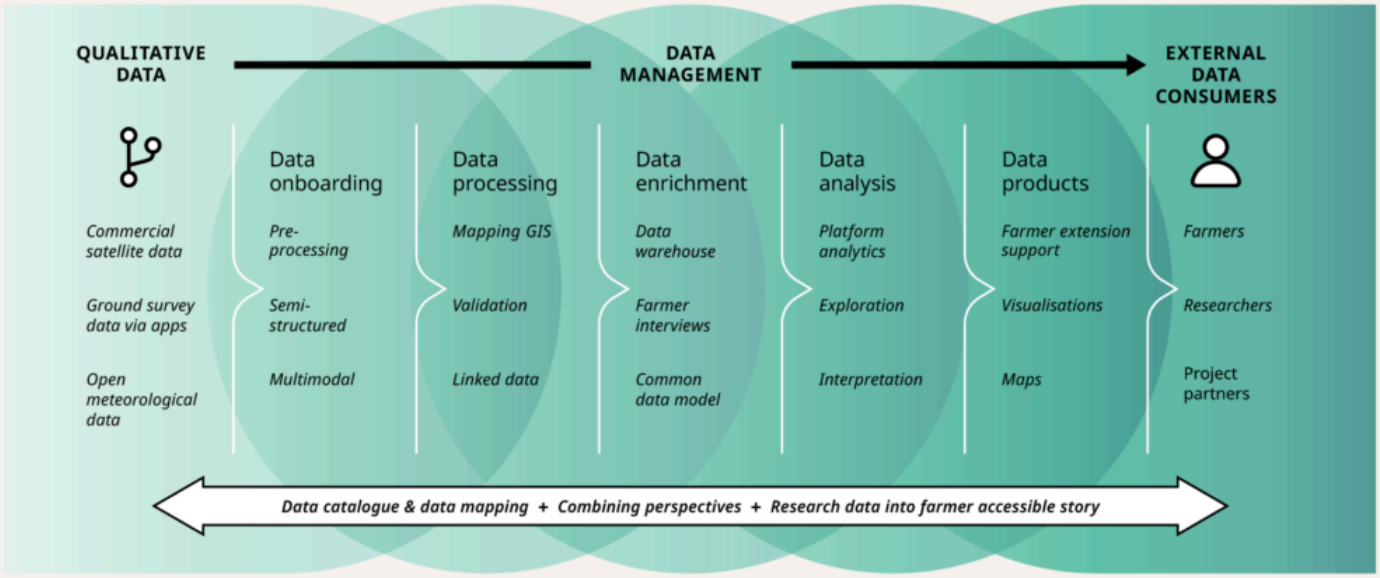
The DVC is a way of viewing the process of running a project from the point of view of the data, thereby identifying how it is onboarded, processed, enriched, analyzed and released in a product. In doing so, the DVC shows the moving parts in project implementations, making it a useful framework regarding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



FAIR data fundamentally aims to improve data access, increasing the data available for AI to learn from. This is critical, as more data-intensive AI innovation (such as generative AI—for example, ChatGPT—or more generally language-learning

models or LLMs) becomes commonplace in Agricultural Development.

Ameen Jauhar, Data Governance Lead, CABI

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.