6.3 Common Data Models (CDMs)

Establishing a standardized, consistent and shared data structure designed to organize and describe data in a way that facilitates interoperability and data integration across various applications and domains

Why should I do this?

To establish a common understanding and representation of data, reducing the complexity and effort required for data sharing and integration.

How do CDMs relate to FAIR?

> **Interoperability:** Building a CDM to plot the infrastructure of your project ensures everyone in the team knows where data is coming from, and at what point in the project datasets and systems are linked together.
>
> **Accessibility:** CDMs provide clear architectures for projects that can be used so that all researchers within the project team are on the same page, which can mitigate the effects of problems such as language barriers.

Download this CDM factsheet for more insights.

What is a CDM?

A CDM is a standardized, consistent and shared data structure designed to organize and describe data in a way that facilitates interoperability and data integration across various applications and domains. The goal of a CDM is to establish a common understanding and representation of data, reducing the complexity and effort required for data sharing and integration.

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 stakeholders

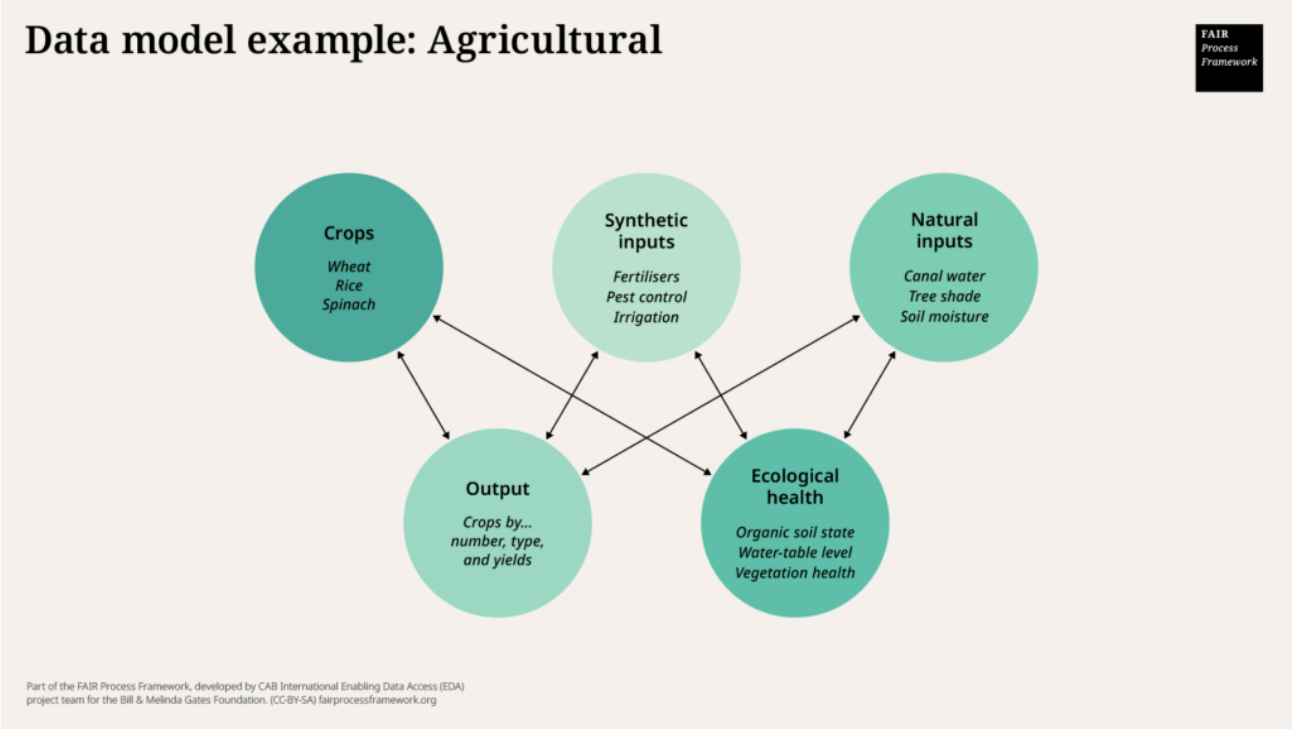
This step applies to a large project with a variety of different stakeholders. If there are, for instance, several organizations taking part in the study, you will need to list them and agree exactly what each will contribute.

Once you know this, you can discuss to determine if all parties already use a CDM or ontology, or whether there is one they can all adopt. If so, then the next step will be much simpler.

Identify the right data model

A data model is a behind-the-scenes organizational template of how the data entering the project is to be managed. The terms 'common data model', 'ontology' and 'terminology' are often used interchangeably during the early stages of project planning. We are talking here about a CDM that can be based on an ontology or terminology, and may indeed be a subset of an ontology. Though it is not the same thing, it can be thought of as a template for all the data elements that are going to be used in the project.

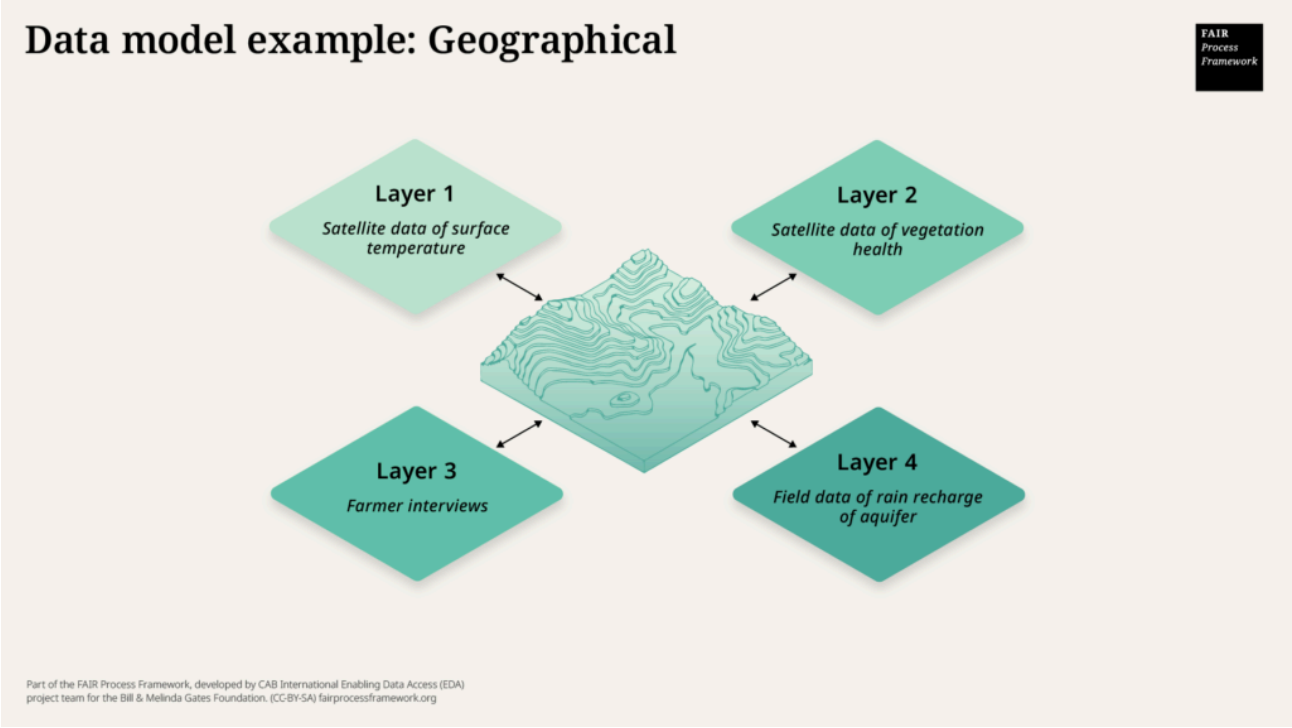
Without a CDM, different datasets will sit in different folders on the computer or the storage cloud, with no way of comparing or analyzing them. The project manager and team will decide in advance the unique ordering of the data to allow for an effective analysis.



In some cases, pre-existing agronomy data models can be used (see factsheet for a list). The data model should naturally suggest itself according to the type of analysis required:

> A quantitative relation—e.g., the effect of fertilizer on output.
>
> A temporal basis of tracking trends—e.g., the climate variation over the long term.
>
> A geographical basis of a topological relation—e.g., the managing of crops in a complex environment.
>
> A semantic basis of identifying patterns—e.g., finding new patterns of crop-climate dynamics.

The data model is a crucial understanding before embarking on research, and must be suitable for the research to be undertaken. Here is an example of a geographical data model that is typical of Geographic Information System (GIS) map approaches:



To realize a good data model, one must focus on the question that one is bringing to the data.

> Is it simply a question of quantifying a relationship as in Figure 2.1 of inputs, crop yield and ecological health as is relevant to Project SIS? In this case, the data model can be precise, listing the variables one wants to measure and the relationships one wants to test. The research is working towards a numerical analysis of measured quantities and outputs.
>
> Or is it a question of place: how does the location and the climate influence the output of crops? Or a question of finding patterns in bringing together different datasets—as happens more in Waterways? In this case, the data model is related to the geography, and the data is overlaid on a topological map, to discover the form of its pattern.
>
> The data model can be quantitative, as in Figure 2.1, where the elements of research are already decided. Or more open, happening through the geographical map, in Figure 2.2.

Identify entities, data elements and relationships

1. Entities

What they are: Real-world objects or concepts that you need to store data about, represented as tables in a database. They usually correspond to nouns.

How to identify:

Review business processes, user requirements, and existing data sources. Look for nouns that represent things important to the system.

Consider both physical objects (customers, products, locations) and abstract concepts (orders, appointments, events).

Examples: Customer, Order, Product, Employee, Department.

2. Data elements (Attributes)

What they are: The specific pieces of information you want to store about each entity, represented as columns in a database table.

How to identify:

Analyze the characteristics, properties or descriptors of each entity. Think about what attributes you would use to differentiate one instance of an entity from another.

Review your business requirements to ensure you are capturing all essential data elements.

Examples:

Customer: Customer_ID, Name, Address, Phone_Number, Email.

Order: Order_ID, Order_Date, Customer_ID, Product_ID, Quantity, Total_Price.

3. Relationships

What they are: Links or connections between entities that describe how they interact, represented as lines or connections between tables in a database. Relationships are often expressed with verbs.

How to identify:

Analyze how entities relate to each other. Determine how instances of one entity are associated with instances of another.

Consider relationships in terms of:

Cardinality (one-to-one, one-to-many, many-to-many): The number of instances of one entity that can be associated with the other.

Optionality (mandatory vs. optional): Whether the relationship is optional or required for entities to exist.

Examples:

A company has one CEO (one-to-one)

A company has many products (one-to-many)

Many companies can have the same set of many products (many-to-many)

Overall process

1. **Gather information:** Understand business goals, processes, data sources, and user requirements.
2. **Identify key entities:** List the core concepts relevant to the problem domain.
3. **Define attributes:** List relevant data elements associated with each entity.
4. **Establish relationships:** Determine the connections between entities, and specify the cardinality and optionality of each relationship.
5. **Review and refine:** Discuss the initial draft with stakeholders and make adjustments based on feedback.

It would be a good idea to establish a good naming convention for all data elements, entities and relationships.

Identify links to existing data standards

Once the CDM is outlined, the next step is to see if data elements can be linked to relevant data standards. In this case, one would look at:

OGC standards for the GIS data: Ensure consistent and interoperable spatial data representation.

FAO vocabulary for any crop details: Standardize crop types for easier analysis and comparison across projects.

SensorML for any sensor data being used: Describe sensor properties for better data interpretation.

AgSW or Agronomy Ontology for any other data elements: Enable integration with other agricultural data platforms.
Once we have links, we can then ensure any rules for the value domain can be enforced.

Identify any layering of datasets and commonalities

Datasets can be layered based on different levels of abstraction and detail. Here is an example:

Level 1: Raw sensor data
This layer consists of the unprocessed data points directly captured by the sensors, including timestamps, sensor IDs, and raw sensor values (e.g., temperature readings in volts).

Level 2: Processed sensor data
This layer involves processing the raw data from level 1 to convert it into meaningful units (e.g., temperature in degrees Celsius) and potentially applying calibration or filtering techniques.

Level 3: Derived features
This layer utilizes the processed data from level 2 to calculate additional features or metrics relevant to agriculture. Examples include:

Vegetation indices: Derived from combinations of different sensor readings to assess plant health and growth.

Soil moisture index: Calculated using soil moisture sensor data to indicate irrigation needs.

Level 4: Crop and field management data
This layer encompasses data related to crop management practices, including planting dates, fertilizer application records, pest control activities, and yield data.

Commonalities
Identifying commonalities across different datasets can be crucial for integrating data and creating insights. Here are some examples:

Spatial data: Sensor data and field information both have a spatial component, often represented by geospatial coordinates. This allows for spatial analysis and visualization of data across different layers—for example, correlating sensor readings with specific locations within a field.

Time-based data: All data points will likely have timestamps, allowing for temporal analysis and comparisons, such as tracking changes in sensor readings over time or observing yield trends across different planting seasons.

Unique identifiers: Assigning unique identifiers to entities like sensors, fields and crops enables consistent referencing and data integration across different datasets.

Benefits of identifying layering and commonalities

Understanding these aspects helps you to:

Organize data effectively: Categorizing data into layers based on abstraction helps with data management, and facilitates analysis at the appropriate level of detail.

Facilitate data integration: Identifying commonalities like spatial and temporal references allows you to combine data from different sources for comprehensive analysis.

Extract valuable insights: By analyzing different layers and combining data points based on commonalities, you can gain deeper understanding of crop health, soil conditions, and factors influencing yield.

Organize research around the data model

The data model is used to ensure that the research is organized with respect to the key questions that need to be answered.

In some research, you already know in advance the key variables you want to measure and the relationships you want to test. In this case, you will already be able to set out a data model with the different elements—say, mineral inputs, soil moisture and crop yields—and the question of the relationship of dependence between them. The data model is self-evident, as the research is to test the effect of minerals and moisture on yields.

In other situations, the research is seeking to find patterns in time. The data model of climate research, for instance, must accurately show trends in weather patterns and how this affects farming practices. The data should be organized to evidence long-term patterns of change. The data model must be expanded to include the relation to time, as a central element in the research.

The data model can also be based on the geographical context. This is the case in Waterways, where the different datasets are layered over the topological map. The layering includes:

The geographical spread of the satellite and field measurements.

The topographical features, such as the canal.

The farmer interviews, which are represented by virtual pins, to locate the information in context.

The data model can be set up to investigate and analyze semantic relations.

The data model is both the organizational template in which the data is gathered, and the frame through which the data product is communicated. The data model is a critical theme in the forming of an analysis.

In Waterways, the data analysis relates the semantic interpretation of the interviews, the satellite data and the topographical features as telling a story together. It is important that one does not interpret this semantic web beforehand, but allows the research to discover the connections.

In the case of Waterways, the surprising relation of the canal water to the surface temperature and crop health was the semantic key to the data model puzzle. The data analysis is only able to find the relation between these elements in the process of questioning the semantic relation between the different layers of 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

Given the aim of this project is to combine data from TPPs and GSS, we will design a CDM to create a blueprint for our data management. Although our data will be geographic in nature, Project SIS will not be producing a topographic map. Instead, the eventual output will be a dataset.

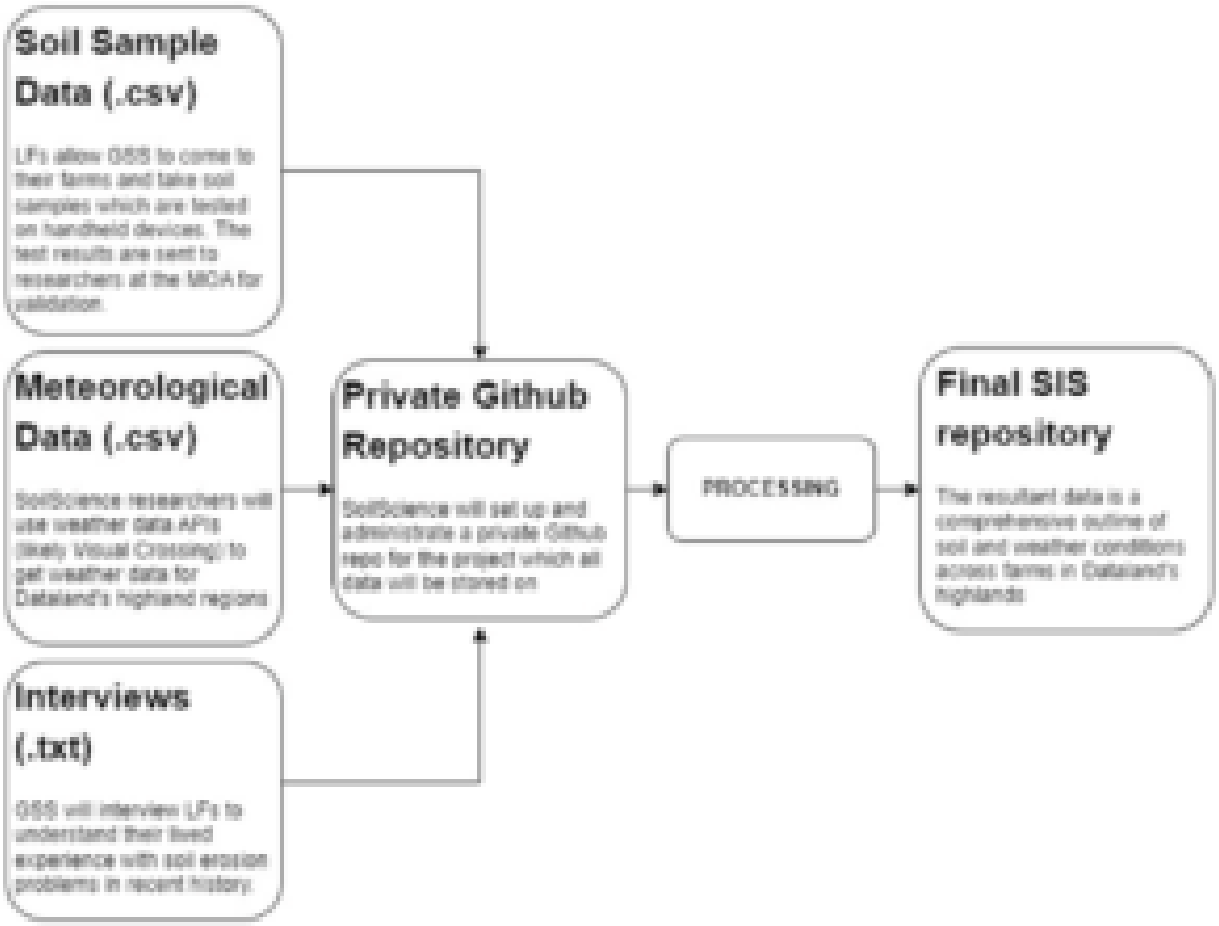


Figure 2 - The CDM for Project SIS

2. Data enrichment

The CDM is consulted to ensure that data is being stored in the right folders within the repository before and after enrichment.

Waterways

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

1. Data onboarding

Before any data is actually onboarded, we will design a CDM that will govern the flows of data within our project. We can split the project into two sub-sections: the building of a comprehensive topographic map (to present data we collect) and the publishing of an academic paper with analysis of the data, which should provide recommendations for crop selection in Waterways.

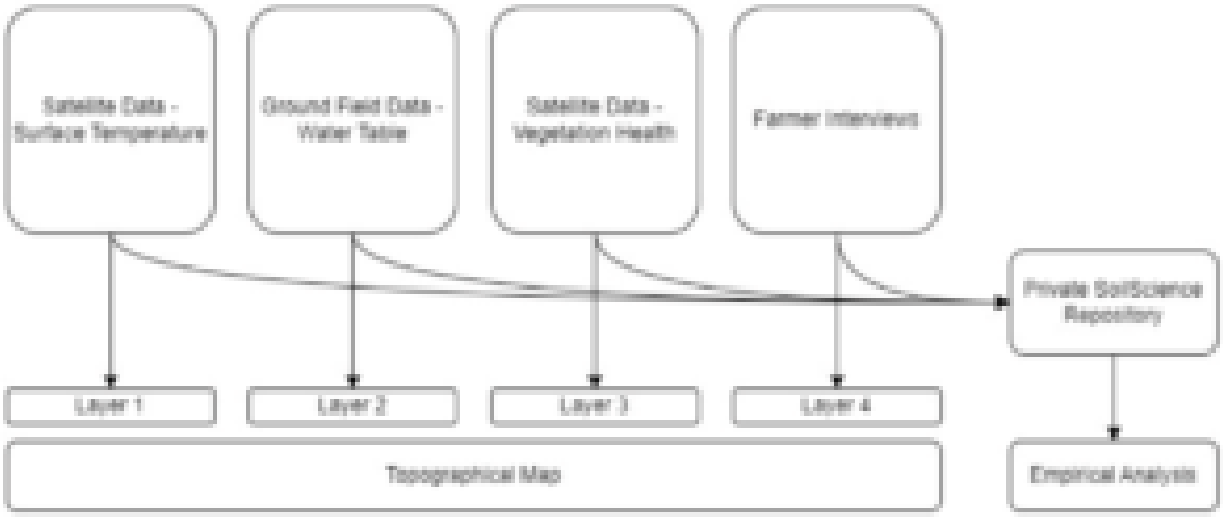


Figure 2 - The CDM for Waterways World

A CDM can be very simple, listing the elements of the research and the relationships to be tested in a quantitative research experiment. As in Scenario A, this might be the predominant question that the scientists are asking.

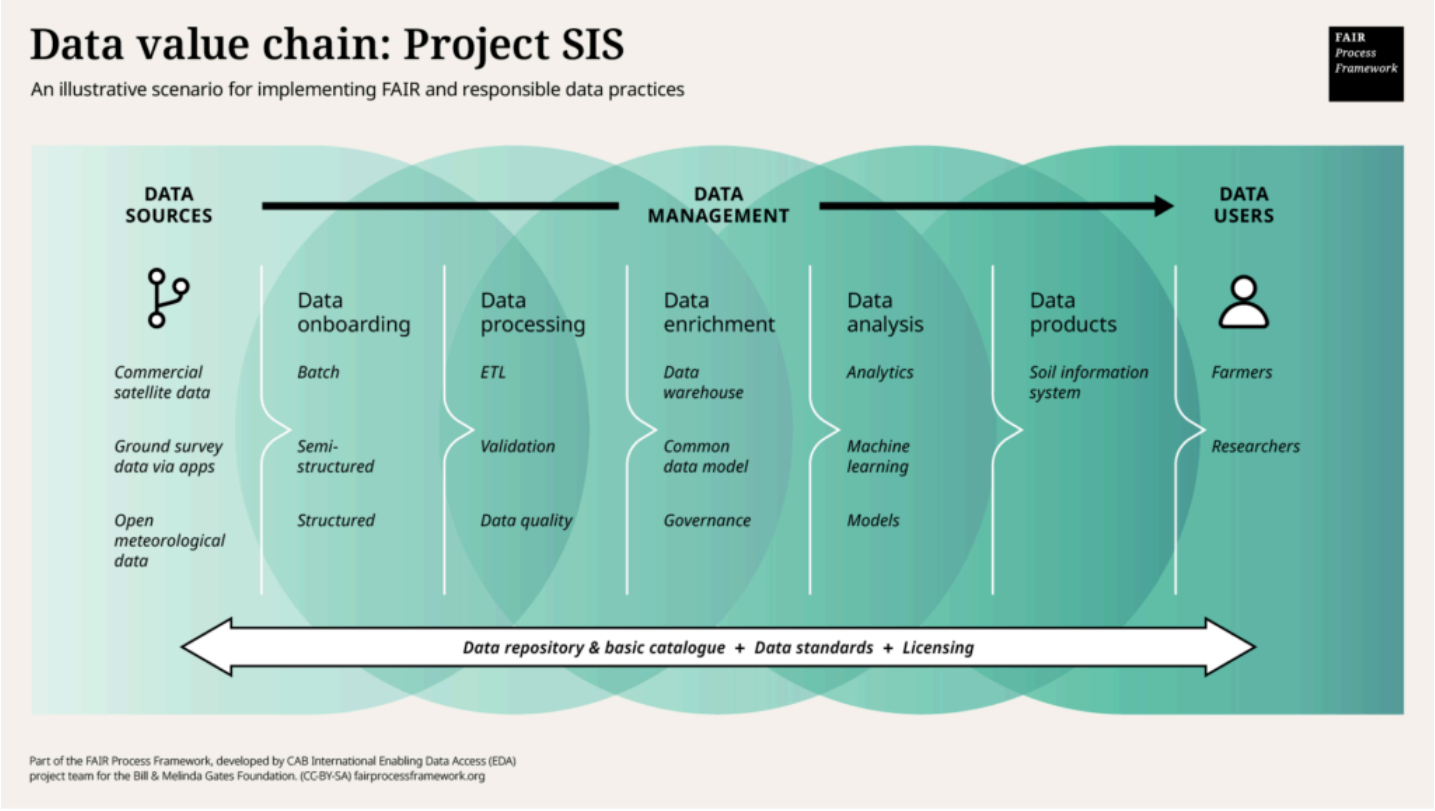
Where there is a lot of complexity in the situation, as is increasingly the case where the climate has become very unpredictable, the data model needs to be more open. In Scenario B, a geographical data model was layered on the ground map of the area, the satellite and field data, the canal, and interviews with the farmers, to arrive at the real picture of the contrasting yields of crops across villages. In either case, the data model is a foundation to the research that can be expanded to deal with related situations, or new elements to be researched. The basic frame of a data model does not change if one repeats the experiment in a different location or with slightly changed parameters. If a new data source is found, and more satellite data is available, the data model allows this to be seamlessly included in the research.

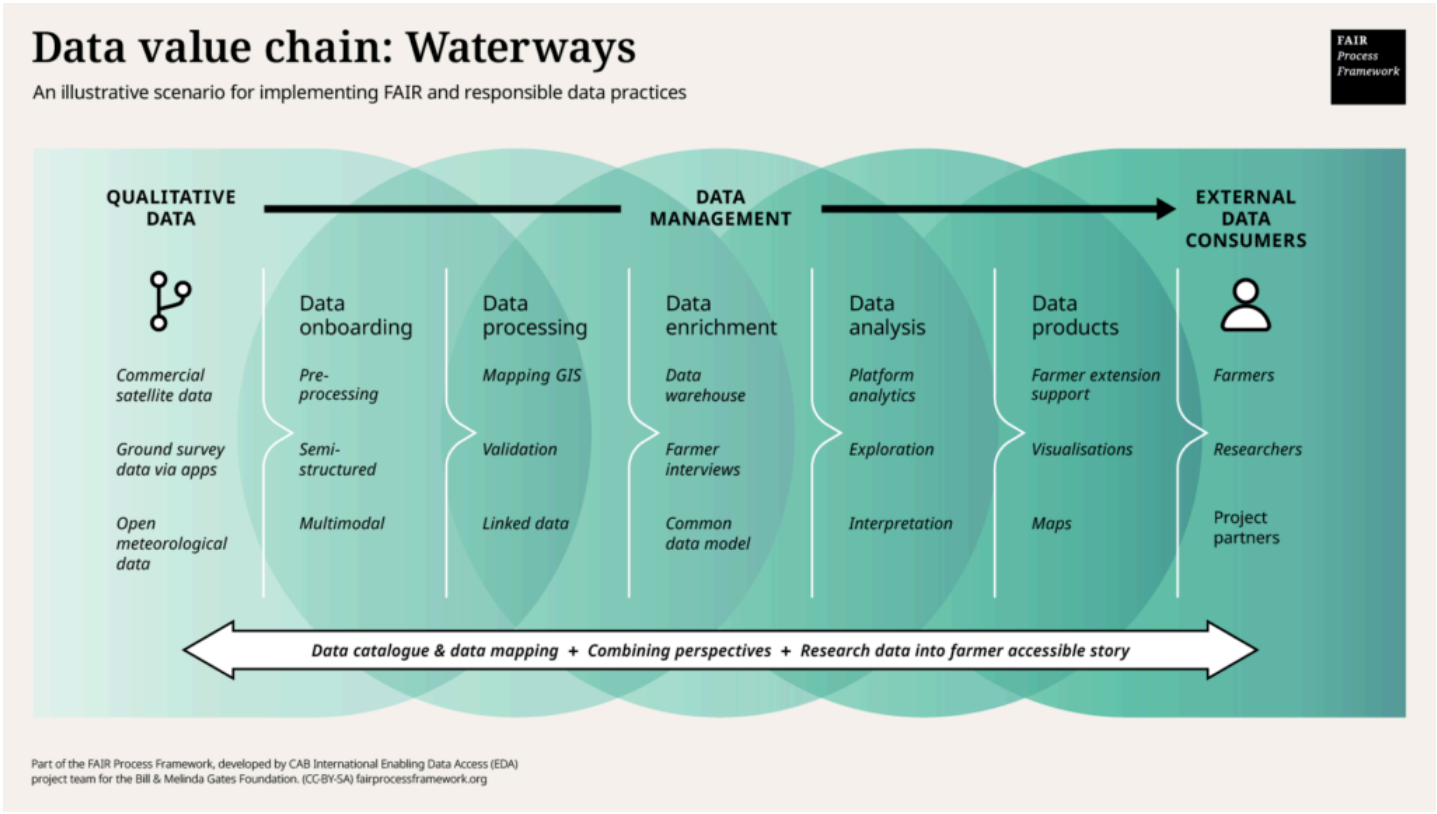
The data model, when properly chosen, is at the heart of a FAIR approach. The data is findable, accessible, interoperable, and reusable, to a large extent based on the CDM that allows one experiment to communicate and relate its results, both to other researchers and to the array of stakeholders.

The theme of CDMs can be important at different stages of your project, whether or not you expect that to be the case. To help you incorporate it into your project planning, this section

provides suggestions about where you should think about the theme, structured using the stages of 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.





It’s important we don’t lose any more data.

Martin Parr, Director, Data Policy & Practice, CABI

Learn more

Acknowledgements

FAQs

Glossary

Accessibility

Privacy & cookies

T&Cs

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.