Assessing fertilizer effectiveness using FAIR soil testing data (Project SIS)

An Introduction to Project SIS 2

Planning Project SIS 3

- Context: Data actors 3
- Context: Data inputs 4

1. Data onboarding 5
2. Data processing 11
3. Data enrichment 13
4. Data analysis 15
5. Data products 16

Reflection 18

References 19

An introduction to Project SIS

Project SIS is fictional but based on existing projects around the world and can be used to understand the processes and planning required for the **collection and management of data**. It is based in the imaginary country 'Dataland' and administered by the imaginary NGO 'SoilScience'.

Dataland is a small country with a landscape dominated by mountains and highlands. In recent years, crop yields have been hampered by nutrient losses from soil erosion as a result of farming on steep slopes. These losses could be mitigated by providing farmers with crop- and soil-specific fertilizers that are well adapted to the agricultural climate in Dataland.

The goal of this project is to create a centralized Soil Information System (SIS), a data repository containing pertinent soil and agronomy data to enable the future discovery of the optimal crop- and soil-specific fertilizer regimes for farmers in Dataland's highlands. Such a repository will be able to inform other agricultural decisions that farmers will have to make in the future.

The project will be carried out by researchers at SoilScience, managed by a grantee, and supported by Dataland's Ministry of Agriculture with some local farmers. The outcome will be a bank of data that can be used for further research.

The following chapters are written from the perspective of a grantee as they plan the FAIR technical implementation of Project SIS. They begin by providing some context on the actors in the ecosystem and data that will be used, and then follow the DVC as a framework to structure the plan for the project while consulting Step 6 resources.

Planning Project SIS

Context: Data actors

In Step 2.1 Identify personas and their value exchanges, we identified the key actors that will each have roles in Project SIS regarding data. They are summarized here to provide necessary context for the technical plan.

Data providers

- **Third-party publishers (TPPs)**: A number of companies and government-funded organizations provide data that will be used in the SIS.
- **Government or Survey Staff (GSS)** in the Ministry of Agriculture **(MOA)**: Other data will be collected by staff at the MOA, who will survey the agricultural landscape of Dataland, interview farmers, and test soil with handheld devices that can be processed in government laboratories.
- **Local farmers (LFs)**: The people growing the crops will authorize the GSS to survey their land and provide information in interviews.
- **Project partners (PPs)** employed by the MOA: Partners in Dataland will organize the surveys and overall data collection, coordinating with SoilScience for the timely delivery of data to the SIS.

Data processors

- **Soil researchers at MOA laboratories in Dataland**: Soil samples are tested on handheld devices, but the test results must be interpreted, validated, and analyzed before being sent to SoilScience for integration into the SIS.
- **Researchers at SoilScience**: SoilScience researchers will compile the data from Dataland with data from TPPs to build the SIS.
- **Project Officer at SoilScience**: The grantee who manages all aspects of Project SIS, but will also be involved in data processing and analysis.

Data consumers

- **Project partner (PPs)**: Partners in Dataland with technical skills and an understanding of soil science will be able to use data collected and stored in the SIS to recommend policy changes to the MOA.
- **Local farmers (LFs)**: Local farmers will hopefully benefit from the data in the SIS, if not directly then by initiatives run by PPs and the wider MOA.
- **Researchers in agricultural science**: The SIS will be an invaluable resource for those researching agricultural science, and although focused on Dataland, the data will be applicable in numerous contexts.

Context: Data inputs

We identified at a high level what kind of data will be collected, and how it will be used, stored, and shared, in Step 3.0 Identifying data assets. More technical information is summarized here.

The primary source of data is the ground survey conducted by GSS. At systematically chosen locations, surveyors will take soil samples and test them with handheld devices [1] for nutrients, entering the results into online forms that upload the data to a central location for analysis. Some of these soil samples will be given IDs and stored to provide physical evidence for the data collected, as well as a historic record of soil composition at the time of sampling.

Other datasets will be bought or obtained free of charge from TPPs for use in the SIS. These will be meteorological data (from Visual Crossing [2]) and commercial satellite data, likely from APIs or downloaded as CSVs from websites.

GSS will also collect qualitative data from local farmers via interviews about agricultural success over the years.

1. Data onboarding

For the initial onboarding of data into the project from the above data providers, we have identified the following themes as relevant:

Compile the aspirational insights in the table below to create feasible FAIR aligning principles:

	APIs	CDMs	Data catalogs	Data licenses	Data standards	Privacy and security
1. Data onboarding	✓	✓	✓	✓	✓	✓

Common Data Models (CDMs)

Given the aim of this project is to combine data from TPPs and GSS, we will design a CDM to blueprint 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.
The overall structure, planned using 6.3 Common Data Models (CDMs), is shown in figure 1.

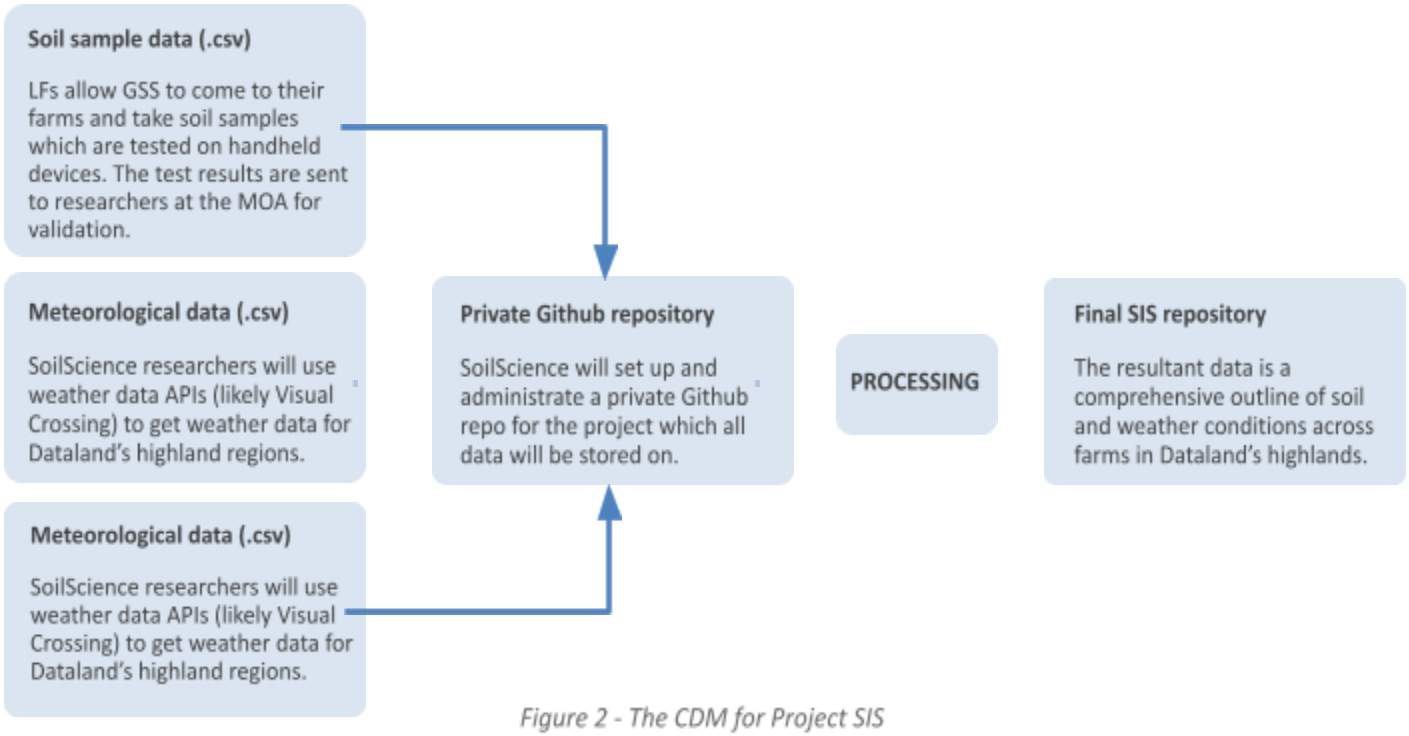

Figure 2 - The CDM for Project SIS

Data catalogs

As per the CDM, the input datasets will be .csv files. To ensure efficient findability and accessibility, each will be stored in a folder with metadata in a JSON. We have designed a descriptive metadata schema (based on the suggested approach provided in 6.5 Data catalogs) 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 last modified.
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 the 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

Data licenses

TPP data like that received from Visual Crossing must be used in a legal way as dictated by the dataset's licenses (6.6 Data licenses).

We have examined the license for use of Visual Crossing data (in its terms of use [3]) to ensure the data can be used for our purposes. In doing so, we identified the type of membership we require, and the cost this would incur.

We are in the process of exploring whether information on the Normalised Difference Vegetation Index (NDVI) [4] of regions in Dataland's highlands would be valuable to include in the SIS. If so, this data would be bought as an asset from a commercial satellite source, in which case data licensing problems may arise: sharing proprietary data of satellite companies with the final users of SIS may require a specific legal contract between SoilScience and said companies. If that is not possible, NDVI data will not be onboarded to the project at all.

Similarly, consent is required from LFs to collect data from their land (with regard to soil samples) and their experiences (with regard to interviews). We will draft a data sharing agreement for LFs to sign in order to provide consent, with its terms published on the project website as part of the license/terms of use for SIS.

APIs

Visual Crossing has an API with information available on its website [5], much like other sources of meteorological data used in agricultural science (6.2 Application Programming Interfaces (APIs). The API will provide meteorological data for areas we are interested in as a .csv file. As per the terms of use mentioned above, our use of the API will have certain constraints, although the project is in an advantageous position in this regard, as we will not be requesting data frequently.

Privacy and security

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, due to 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 MOA staff will be able to onboard (raw) data collected to the project's centralized repository. They will have further responsibility for ensuring deletion of any data stored in unsecure, local devices (like those used to record interviews) and keeping oversight of GSS and LFs. This system will be governed by a simple login system (as provided in the suggested approach in 6.7 Data privacy and security).

As per the CDM, the centralized repository data will be onboarded to a private Github repository, with access limited 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.

How do these themes bring a project closer to FAIR?

Three FAIR principles are being prioritized here. Applying a data cataloging method makes data **findable**, ensuring that it can be easily discovered through unique identifiers and metadata, reducing duplication of efforts and saving time. Cataloging also increases semantic **interoperability** by building a coherent administrative structure that works alongside the CDM for easier collaboration between team members and synergy with other projects. Finally, the login system works to build **accessibility**, where authorized people can securely work while data privacy constraints are well-respected and, being inclusive of stakeholders' privacy requirements.

Appendix: What the data will look like

Meteorological data returned from our Visual Crossing API queries will take the form of a CSV with the following columns:

- datetime
- tempmax
- tempmin
- temp
- feelslikemax
- feelslikemin
- feelslike
- def
- humidity
- precip
- precipprob
- precipcover
- preciptype
- snow
- snowdepth
- windgust
- windspeed
- winddir
- sealevelpressure
- cloudcover
- visibility
- solarradiation
- solarenergy
- uvindex
- severerisk
- sunrise
- sunset
- moonphase
- conditions

The results of soil sample testing will come as a CSV with the following columns:

- id
- collector_id
- latitude
- longitude
- datetime
- texture
- moisturecontent
- salinity
- ph
- iron
- nitrogen
- som
- potassium
- phosphorous

2. Data processing

Data processing in this project will involve the standardization, cleaning, de-identification and validation of data onboarded. For the TPP meteorological datasets, historical weather data will be aggregated or averaged over the 24 months before soil samples were collected. These processes will mostly be completed by SoilScience researchers. The following themes were found to be relevant:

	APIs	CDMs	Data catalogs	Data licenses	Data standards	Privacy and security
2. Data processing			✓		✓	✓

Data standards

SoilScience research has previously used the Agronomy Ontology [6] data standard, so it is likely that we will use it again for this project, applying it as presented in 6.4 Data standards. The Plant Experimental Conditional Ontology, a sub-component of the Agronomy Ontology, 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 for this project it should not be an issue, as meteorological data is likely to come from one source only (Visual Crossing). Further processing will 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.

Privacy and security

All personally identifiable information (PII) will be removed from the data received from Dataland, as recommended in 6.7 Data privacy and security. 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 does not overwrite the raw data sources, but instead makes a 'clean' copy of each one. However, after the processing is complete, the raw data sources will be deleted from the repository.

Data catalogs

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

How do these themes bring a project closer to FAIR?

The data standards introduced in this step build **interoperability** once again by molding input datasets into vocabularies and ontologies that have been built and agreed over the course of recent agricultural research. In doing so, we facilitate smooth integration of data into our analysis methods, even in other projects where this data can be reused, thereby meeting the FAIR principle of **reusability**. Removing the PII for privacy and security reminds us that **accessibility,** in the FAIR sense, does not mean that all of the data must be freely accessible to all. It is our responsibility to ensure that accessibility aligns with stakeholders' privacy. Updating metadata via data catalogs after processing (and the subsequent steps of this work) enhances the **findability** of our data, both internally for our researchers to easily be able to find what they need, but also externally when SIS is released.

3. Data enrichment

Data sources will be linked to one another as part of SIS, which means that third-party meteorological data will be linked to the data collected on the ground, culminating in a granular dataset on agricultural conditions across Dataland's highlands (as per the CDM). The method will be as follows:

- Each incoming dataset contains one latitude and one longitude column. We will first combine these columns into one 'linking field' where, for each row, the entry in its linking field is a tuple of its latitude and longitude entries.
- With this singular linking field, the datasets can be linked to one another using a merge function. Care will be taken to preserve the order of columns.
- Latitude, longitude and the linking field will be deleted from the data to preserve privacy of local farmers.

	APIs	CDMs	Data catalogs	Data licenses	Data standards	Privacy and security
3. Data enrichment		✓	✓			✓

CDMs

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

Data catalogs

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

Privacy and security

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 LFs 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

SoilScience will perform preliminary analysis on the final dataset. The following themes will be considered.

	APIs	CDMs	Data catalogs	Data licenses	Data standards	Privacy and security
4. Data analysis				✓		✓

Data licenses

Preliminary analysis on the SIS' coverage of Dataland as well as the identification of general trends will be done with Python or R based Notebooks. Given these are third-party tools, the licenses for TPP datasets must be consulted to understand their restrictions for use (as given in the suggested approach in 6.6 Data licenses).

For Visual Crossing data, analysis with third-party software is allowed in its terms of use [3]. Proprietary commercial satellite data, if used, will likely have different restrictions.

Privacy and security

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

5. Data products

The SIS will be released as a repository accessible to researchers, Dataland's MOA, and farmers themselves in the highlands. Here, lots of technical considerations will need to be made.

	APIs	CDMs	Data catalogs	Data licenses	Data standards	Privacy and security
5. Data products	✓		✓	✓	✓	

Data catalogs

The metadata will be audited to ensure accuracy and usability for potential users of SIS. This also involves checking that metadata is easy to discover and read on the platform 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.

Data standards

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

Data licenses

We will need to investigate how to license the SIS data to researchers beyond Dataland's MOA and our project partners, with whom we will have data sharing agreements. We could use an open license like Creative Commons 4.0 [7], in which case we will reserve attribution rights but otherwise the data is free to use for any sort of project, or we could use a non-commercial license (Creative Commons BY-NC 4.0 [8])., which may be preferable to fit the overall community aims of SoilScience, but user research and some further stakeholder interviews would be required to judge which license to choose.

APIs

Although the Project SIS data will be held on the SoilScience website, we are considering building an API that can be used by researchers to import data that fits their needs directly into the analysis tools they use.

To do so, we will explore APIs in the agronomy sector that fulfill similar roles to SIS. The NRCS web soil survey API [9] might help in this regard.

How do these themes bring a project closer to FAIR?

Data cataloging once again enhances **findability** and, with the ontology field dictated by data standards, also enhances **interoperability**. Meanwhile, the final license on SIS data will govern its **reusability**, dictating in what way other people can use our work. If we build an SIS API, its design should promote **accessibility**.

Reflection

Throughout the planned technical implementation of project SIS, the core themes given in Step 6 apply in numerous ways. A good example of this is the APIs theme, where we must not only consider their use with regard to data onboarding, but also for eventual product release. We are grateful for the resources provided to us, which we will continue to use in order to gather a comprehensive understanding of the technical facets required in this project.

References

[1] https://www.niubol.com/Handheld-test-instrument/7-in-1-Handheld-Portable-Soil-Tester.html

[2] https://www.visualcrossing.com/

[3] https://www.visualcrossing.com/weather-services-terms

[4] https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index

[5] https://www.visualcrossing.com/weather-api

[6] https://bigdata.cgiar.org/resources/agronomy-ontology/

[7] https://creativecommons.org/licenses/by/4.0/deed.en

[8] https://creativecommons.org/licenses/by-nc/4.0/deed.en

[9] https://websoilsurvey.nrcs.usda.gov/app/