Utilizing technology and data for optimal crop selection (Waterways)

An introduction to Waterways

Waterways is a fictional project in the country 'Waterways'. Its planning and technical implementation can be used to understand the processes and planning required for **not only the collection and management of data, but also its enrichment, analysis and use**. The project is run by the imaginary NGO 'SoilScience'.

Waterways is a country situated in a plain crossed by a system of flowing rivers carrying the discharge from distant mountains. The climate, dry and hot for most of the year, is interrupted by seasonal heavy rains for two or three months between July and September. October is a key month for the agricultural cycle, as this is when most water is available for crops.

The area is still largely traditionally farmed. The farmers of Waterways have small plots of around one to 15 hectares where they grow two crops a year (although if enough water is available, they can exceptionally grow three crops) alongside raising a few cattle. The crops include rice, wheat, corn, potatoes, bitter gourd, okra, rai (a leafy vegetable), and black and green beans. To irrigate their crops, farmers pay for access to bore wells. This presents a financial limit to the number of crops they can grow and cattle they can support. Almost all villages in Waterways are now electrified, and mobile phone usage has massively grown recently as a result.

The Waterways Ministry of Agriculture (MOA) runs a research site, the Waterways Research Observatory. This operates across a 5x10km area (serving 24 or so farming villages) using a number of meteorological stations to monitor the area for rainfall, groundwater level, crop water stress, and temperature. The observatory records how boreholes are having to be dug deeper and deeper to reach the water as the groundwater level steadily falls due to insufficient recharge from rains. The traditional ways farmers used to hold water from rains, such as in ponds holding tons of water for use in the dry season, have lapsed. The holding of rainwater is now only sporadically followed.

Waterways is a project run by SoilScience that aims to mitigate the negative impacts of limited access to water sources by not only facilitating data collection with the research observatory, but also building comprehensive analyses to understand optimal choices farmers can make. An important facet of the project is facilitating access to the research for the farmers: too often, impactful research is done but never seen by those who need it most.

The following chapters are written from the perspective of a grantee as they plan the FAIR technical implementation of Waterways. 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 Waterways

Context: Further information

This section provides further information on the project that will contextualize the intervention planned in this document.

The country Waterways has a number of water sources coupled with a very hot and dry climate. As a result, there is high potential for farming in Waterways if, and only if, farmers have secure and reliable access to water for their crops. However, the farming areas of the country are slowly drying up as a result of hotter temperatures in the dry seasons and unpredictable rains in the wet season. Crops are suffering and the reliance on depleting groundwater sources is not sustainable.

As an illustrative example of the dangers of climate change for farming, agriculture in Waterways has been the topic of much academic research. However, very little of the information from this research reaches the farmers in the country: either data is stored in a server inaccessible to the farmers, or, more pressingly, the insights from analysis of the data are written in an academic paper that requires a fee for access that is greater than the farmer earns from his crops over a whole year.

As a result, farmers are limited in their capabilities to make informed decisions. Without reliable external advice, they can be swayed easily by the information provided to them by local fertilizer outlets with a vested interest, paying exorbitant prices for fertilizers that do not fit their needs at all, and in fact have been found to give skin rashes in some cases.

The farmers themselves recognise the situation they are in, and are willing to cooperate with researchers to better their lot. However, it is up to us to provide them with research they can actually action.

We have designed Waterways to ask and answer the following questions:

1. How can farmers in Waterways best optimize their crop yields?
2. How can satellite data show the impact of water availability?
3. How does GIS software enable the analysis and presentation of data?
4. How can multiple datasets be layered to disclose underlying principles?
5. How does satellite and field data match community perspectives and experience?

Context: Data actors

In Step 2.1 Identify personas and their value exchanges, we identified the key actors that will each have roles in Waterways 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 project. We believe that we are most likely to work with the European Space Agency [1] for our satellite data needs.
- **Waterways Research Observatory (WRO)**: Run by the Waterways Ministry of Agriculture (MOA), the WRO provides meteorological and agricultural data for a large region of Waterways.
- **Local farmers (LFs)**: Interviews with local farmers about their experiences, and how they have changed over time, will provide valuable enrichment to the data we collect. Interviews will be conducted by our partners in the MOA.

Data processors

- **Researchers at the WRO in Waterways**: WRO researchers are those collecting and validating data that will be a part of Waterways.
- **Researchers at SoilScience**: SoilScience researchers will bring the data collected by WRO together with TPP data and then analyze the resulting dataset to find answers for optimal crop selections in Waterways.
- **Project Officer at SoilScience**: The grantee manages all aspects of Waterways. They will take the leadership role in processing and analysis of the dataset given their previous experience in academic research for agronomy.

Data consumers

- **Project partner (PPs)**: Partners in Waterways will be able to use results from the analysis to recommend farming methods to LFs as well as policy changes to the MOA.
- **Local farmers (LFs)**: Local farmers will be able to access the analysis results via their mobile devices in order to make informed decisions in their future crop selections. If they lack the technical ability to do so, they will still benefit from initiatives run by PPs and the wider MOA.
- **Researchers in academia**: The findings of Waterways will be published in an academic journal for the wider scientific community to use.

Context: Data inputs

The TPP Satellite Data we use will likely be:

- Surface temperature data (from a satellite measurement, this shows the temperature on crop canopies)
- Fractional vegetation data (vegetation cover)
- Normalized Differential Vegetation Index data (vegetation health)
- Temperature Condition Index data (vegetation resilience)

From the WRO, we will receive:

- Meteorological data (particularly regarding temperature and precipitation)
- Ground field data
    - The WRO collects data on ‘recharge’, which is a straightforward measure of the level of the water table in Waterways. It involves measuring groundwater and identifying how much it increases as a result of precipitation and how much it decreases as a result of irrigation and human use [2].

PPs will interview LFs to provide a way to enrich the data we collect with lived experience.

1. Data onboarding

The first part of the project will be bringing data from the input sources above together into one repository. Here, the following themes will be relevant:

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

CDMs

Before any data is actually onboarded, we will design a Common Data Model (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 the data we collect) and the publishing of an academic paper with analysis of the data, which should provide recommendations for crop selection in Waterways. As a result, the overall structure of the project will look a bit like Figure 1, planned using a similar model for topographic map projects presented in 6.3 Common Data Models (CDMs).

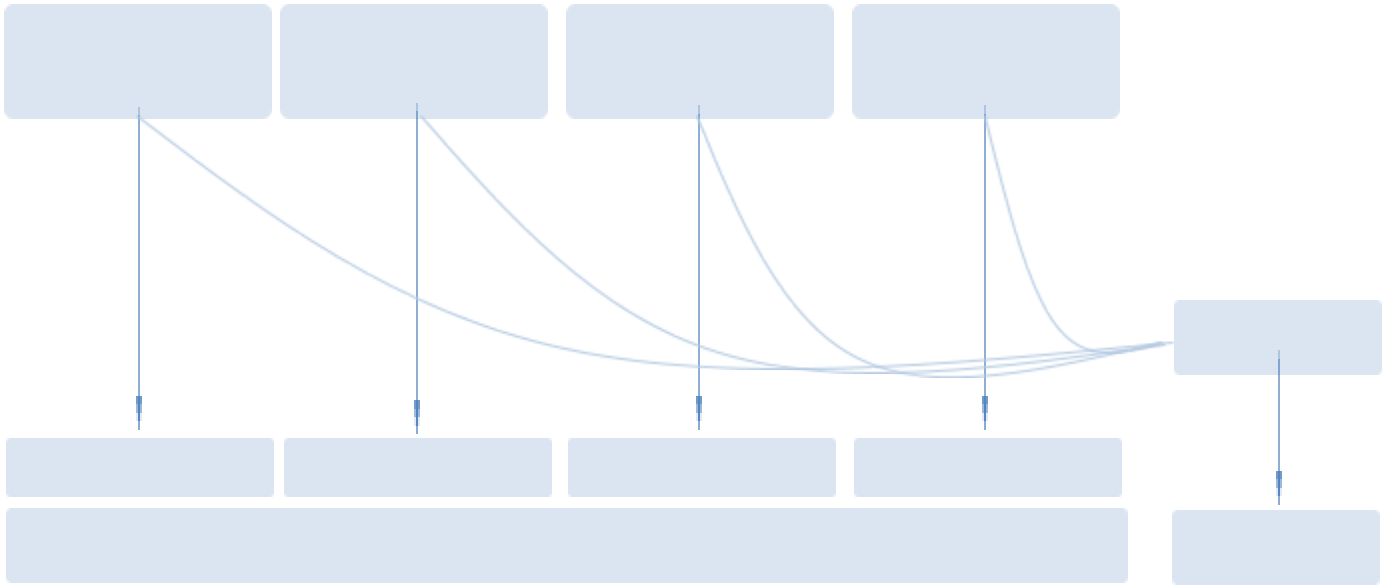

Figure 2 - The CDM for Waterways

Data licenses

Satellite data that we collect from TPPS must be used in a legal way as dictated by the dataset's licenses (6.6 Data licenses). Depending on the license, we may not have the authority to use data in certain ways. In the best case, we might only have to credit the original data authors (as is the case with CC BY 4.0 licenses [3]), whereas in the worst case, licenses may have 'No Derivatives' clauses that could prohibit the publishing of the topographical map as its own resource (as is the case with CC BY-NC-ND 4.0s [4]).

Similar to licensing, we must understand the means of authorization to get access to TPP Satellite data. Some datasets will require fees for access, which must be budgeted for. Thankfully, SoilScience has a professional network that may be able to help with fee waivers.

Finally, consent is required from LFs to collect their data via interviews.

APIs

APIs will be used to access the satellite data held by TPPs. For our queries, the APIs will respond with data on surface temperatures and vegetation health as either KML files, shapefiles, or basic CSVs.

To ensure the data we get back is limited only to our area of interest—i.e., the area in which WRO operates—we will build and apply a 'mask' as part of our API query.

Data catalogs

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.

Privacy and security

Our main concern with regards to privacy and security is ensuring that the data is only accessible to core people in the project. Though the data is not necessarily sensitive, we would like to keep a tight ship, especially if required by the satellite data licenses.

To do so, we are designing a login system with four levels, as suggested in 6.7 Data privacy and security).

- Uploading: WRO researchers can upload their collected ground field data and edit it if required. Similarly, PPs can upload transcripts from the interviews with LFs.
- 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 project officer at SoilScience will be the only person authorized 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.

How do these themes bring a project closer to FAIR?

The CDM works to build **accessibility** for the research team so that the overall project architecture is clear, which will also help in communication with funding organizations and our partners on the ground in Waterways. The login system works to build accessibility, where authorized people can securely work while data privacy constraints are well-respected and inclusive of stakeholders' privacy requirements.

2. Data processing

This stage involves bringing the datasets in the repository together as layers of a single topographic map, so that a user will be able to read the map and see how water availability and the overall climate differ across Waterways. This will provide a visual aid to farmers and the MOA as well as a starting point for our analysis. Note: here, only the first three layers of the map are being processed.

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

Data standards

Satellite data can come as KMLs [5], shapefiles [6], 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 the files so that they can be easily integrated into a topographical map via Geographic Information System (GIS) software.

For our current plan, we believe that converting all data into KMLs might be the best option, as they are particularly useful for presenting time-variable information on maps given their lightweight nature, unlike ESRI shapefiles, which are typically static. Time delineation would be nice to have on our final output (showing water table measurements and vegetation health data over the course of a year), therefore pushing us towards KMLs.

To do so, we will build a standardization script in Python to run on datasets coming into the project. We have previous experience with Python's Geopandas library [7] and are aware of the ability to enable drivers with environment variables to handle KMLs.

The script will also ensure latitude and longitude measurements are correctly formatted and lie within the bounds of the area of interest for this project. These latitude and longitude measurements can then be used to link incoming datasets together, which will be especially important for layer 3 data, which
is coming from three measurements (Fractional Vegetation data, Normalised Differential Vegetation Index, and Temperature Condition Index). As a result, the Python script will return
a standardized KML per layer of the topographical map.

Data licenses

Work with the satellite data will most likely use third-party tools. Besides the standardization script in Python, we will use GIS software for the map presentation (see below). Sometimes, data licenses can prohibit the use of third-party tools on the data, so our terms of use for the satellite data must be considered.

How do these themes bring a project closer to FAIR?

Settling on a chosen data standard—in this case, the use of KML—builds **interoperability** by making the integration of our data sources much easier (as a standardized method for all data sources).

Appendix: GIS tools to build topographies

The resultant KMLs need to be built into a map using GIS software, for which there are two options: ESRI [8] and Google Maps (note: Google Maps is technically not GIS software but is often used for geographical projects like ours).

ESRI is a subscription service that provides research-oriented tools specialized for the presentation and subsequent analysis of topographical maps. The interface is not very user-friendly, but nonetheless meets the needs of our research.

On the other hand, Google Maps is user-friendly and, more importantly, free. Although it lacks tools for comprehensive geographic analysis, it facilitates easy access for researchers and farmers alike. However, we have found the process for importing KMLs into Google Maps as layers of a map incredibly difficult, bordering on impossible. We have therefore decided to use ESRI.

3. Data enrichment

The lived experiences of farmers in Waterways are valuable resources that should be used for any meaningful research. Because some farmers have been working for over 40 years, with their families caring for their plots of land for potentially centuries beforehand, they can provide poignant perspectives on how agriculture has changed and how they are trying to adapt. It is hoped that we can gather insights that can enrich the data we will collect in order to build an understanding of what crops could work well for the farmers as the climate changes.

To do so, LFs will be interviewed by our PPs in Waterways. Their experiences and answers to our questions will be displayed on the topographical map and stored in the SoilScience repository for further analysis.

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

Data catalogs

Cataloging interviews, and attaching data to them, is extremely important because doing so provides 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 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.

Data standards

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, which shapefiles work well with. The points are simply the latitude and longitude of the center of each interviewed farmer's farm.

Privacy and security

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 Personally Identifiable Information (PII), like the farmers' names, could have similar effects, and must therefore be identified within the project (6.7 Data privacy and security).

Via a contract, we will have consent from interviewed farmers to publish this information, but far more important is education: 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.

How do these themes bring a project closer to FAIR?

Cataloging interviews enables **findability**, **interoperability**, and **reusability**, therefore enriching the interview dataset and the overall work that we do.

Appendix: How we build interview questions

After initial desk research and some discussions with LFs, PPs and the MOA, we think we will be able to produce some interview questions that will not only reveal data from lived experiences of farmers, but also hopefully prompt them to ask themselves questions and see the scientific side of agriculture. This can better enable future interventions, by changing the way farmers engage with external support like SoilScience.

Interviews will likely proceed along the same thread: 'How does your farming depend on the weather and water available to you, and how has that changed over time?'. Further prompts will be used to explore how farmers and their peers have approached problems, and what they think of their efforts to do so.

Informally before interviews, we will discuss with each farmer their understanding of the project, their comfort regarding how their data is used, and the potential ways that they could benefit from the outcome. This consideration is important and aligns with our strategy: putting the farmers first.

4. Data analysis

The resultant topographical map may be good for initial visual analysis, but the meteorological and agronomical measurements that underpin it will be most usable for actual in-depth work to identify optimal crop routines. To that end, the farmers' interviews will help to contextualize analysis and guide it into the right directions.

Our methodology will be as follows:

- As a team, we will read through the interview transcripts to first get a good idea of the experience of farming in Waterways. This should contextualize our research and inform our analysis. We consider it best practice to do so.
- Only after interviews have been read through will we start to analyze the topographical map. Our researchers have academic experience in geographical research, so we should be able to identify trends and patterns in the temperature, vegetation, and ground field data across the landscape. These trends and patterns will provoke research questions and initial suggestions for crops.
- We will conduct desk research (including a literature review) based on these questions and initial crop suggestions to identify the current research that can help our work.
- Working with the data, we will build empirical analyses (likely linear regressions, classifications, and principal component analyses) that can evidence patterns in the agricultural landscape of Waterways.
- These analyses will then be used to inform crop selections. Experts at SoilScience and in our professional network can take the results of our work and identify which crops could be best suited to the trends and patterns evidenced. These crop suggestions will consider both climate and economic factors—we are obligated to recommend to farmers the crops that can best enhance their standard of living and social mobility. We will also ensure that there will be many suggestions—it would be dangerous to only recommend one crop for the entirety of Waterways.

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

Data licenses

Data will be analyzed using RStudio or IPython notebooks. As these are both third-party tools, licenses for the Satellite data must be checked. This should be done in the 'Data Processing' stage, but it's worth considering here too.

Privacy and security

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

Appendix: How interview transcripts help

As mentioned earlier, we believe reviewing interview transcripts first is best practice. Researchers should have a comprehensive understanding of the lived experiences of stakeholders in order to do the work that aims to help them. In a previous (extremely small scale) pilot of this work, interviews really showed their strength.

Across interviews, farmers referenced a canal system that our project partners at the MOA were actually unaware of. Farmers would mention that in the north of Waterways, where canals are present, there is often enough water to even grow rice (a crop that requires a lot of water), whereas in the South, only wheat can really be grown as the rains don't deliver much water. For context, there is only around 10 kilometers between the 'north' and the 'south'.

Farmers in the north were not using the canal water for irrigation, instead sourcing water from bore wells with exactly the same architecture as those in the south. Ground field data found that the water recharge in the two areas was not too different either. So the question had to be asked: how was the water supply and therefore the overall farming conditions so much better in the north than the south?

On reading interviews, this question was at the forefront of our minds. We first set out to identify where the canals actually were with satellite photography, but we simply couldn't. But when we overlaid the surface temperature data, it was impossible not to see the almost straight lines of blue, meaning there were likely man-made structures that were reducing surface temperature. These had to be the canal systems, and what we were seeing was their true effect: they helped agricultural conditions in the north by drastically reducing surface temperatures, thereby ensuring water would not quickly evaporate, even in the hottest of summers. This information could be used to recommend better crops than rice for farmers to use in order to take advantage of the conditions in the north.

Meanwhile, these findings also contextualized the work that might be required in the south. What crops could cope with surface temperature problems causing water loss? With only the information from the pilot study, we cannot answer that question yet.

5. Data products

We want Waterways to provide value to all. To do so, we will release the topographical map to the public and promote it to farmers in Waterways. We also plan to release the results of our analyses to all in two ways: first, an academic paper that we will ensure is free to access, and second, a website with our findings in simple vocabulary and the local languages of Waterways for farmers to use to understand what is behind the crop suggestions that we (and the MOA) will give them.

These products will be presented to farmers in meetings organized by our project partners. We want these meetings to regularly continue for a few months so that we can monitor feedback and the overall adoption of our suggestions.

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

Data licenses

We will need to make sure we are allowed to publish the satellite data in the topographical map. If the data is licensed with a non-derivative component, that would be impossible. This should have been addressed in the data onboarding, so it is hoped that it would not be a concern by the time of product release.

Reflection

We are optimistic that Waterways will accomplish its aim to reform agricultural practices in Waterways in order to adapt farmers to the changing climate and water availability. Although the technical implementation plan is at first glance complicated, it does not require too much thinking and its components are quite flexible. Nonetheless, we are grateful for the resources provided to us, which we will continue to use to gather a comprehensive understanding of the technical facets required in this project.

References

[1] https://earth.esa.int/eogateway/catalog

[2] https://en.wikipedia.org/wiki/Groundwater_recharge

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

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

[5] https://en.wikipedia.org/wiki/Keyhole_Markup_Language

[6] https://en.wikipedia.org/wiki/Shapefile

[7] https://geopandas.org/en/stable/#

[8] https://www.esri.com/en-us/home