COVID-19 Survey Data Management Plan

Precision Agriculture for Development

April 2020

Table of contents

1 Introduction

The aim of this data management plan (DMP) is to document the process for creating, storing, and maintaining data associated with PAD's COVID-19 project. Specifically, this DMP outlines systems for:

- Ensuring data searchability, comprehension, quality and security
- Ensuring subjects' privacy, confidentiality and anonymity

This DMP is a live document and will be regularly updated following consultations with international and local IRBs and/or changes to the project's scope.

2 Guidelines

The Health Media Lab IRB has approved PAD to collect and secure data for the COVID-19 project as follows:

Description of Data Collection: Surveyors will receive a list of sampled subjects' phone numbers and names from a secure cloud server. Surveyors will collect data using the ODK Collect application for Android smartphones. No new PII will be collected. The data will be stored in the secure cloud server and will not be directly identifiable, but there will be a code that links to the identifiers (i.e. to the respondents' phone numbers).

Data Safety and Monitoring: Databases will be managed in password-protected computers and stored on an encrypted hosting service. Only PAD researchers, all of whom are certified in human subjects research, will have access to the databases. The code that links the data to the identifiers will be destroyed after 3 years.

Amendment 1: We will ask for farmers' consent to record the phone interviews for back-check purposes in some states. The recordings will not include personal identifiers and will be stored on an encrypted hosting service.

3 Data systems

Table 1: Data Instruments Produced

Data	**Format**	**Storage**	**Software used**	**Access**
Farmer survey instrument	Google Sheets, .xlsx, .xml	Google Drive, Dropbox	Uploaded via ODK Aggregate	Publicly available
Project farmer survey instrument	Google Sheets, .xlsx, .xml	Google Drive, Dropbox	Uploaded via ODK Aggregate	All PAD
Agro-dealer survey instrument	Google Sheets, .xlsx, .xml	Google Drive, Dropbox	Uploaded via ODK Aggregate	Publicly available

Project agro-dealer survey instrument	Google Sheets, .xlsx, .xml	Google Drive, Dropbox	Uploaded via ODK Aggregate	All PAD
Surveyor tracking sheet	Google Sheets, .xlsx	Google Drive, Dropbox	Google Sheets	All PAD
Protocol guide	Google Sheets, .xlsx	Google Drive, Dropbox	Dropbox, Google Docs	All PAD

Table 2: Data Created

Data	**Projects**	**Format**	**Storage**	**Software used**	**Access**
Samples	*All*	.csv, .xlsx	Dropbox, ODK Aggregate, Smartphones	- Generated using Stata - Uploaded to ODK with form	Core team + project research team only
Identifiers dataset	*All*	.dta	Dropbox	- Generated using Stata	Core team + project research team only
Raw survey data	*All*	.csv, .dta	Dropbox, ODK Aggregate	- Downloaded via ODK Briefcase using encryption key	Core team + project research team only
Interview recordings	*India only*	.wav, .aac	Dropbox, WhatsApp	- Recorded via Android recording app - Stored in Dropbox - (possible) Shared via WhatsApp	Core team, project research team, field team (partial)
Cleaned, de-identified survey data	*All*	.xlsx, .dta	Dropbox	- Cleaned and analyzed using Stata	Core team + project research team only

Table 3: Existing Data Sources Used

Data	**Programme**	**Usage**
PAD's activation dataset	All	Generate weekly samples for farmer survey

4 Data and metadata protocols

I. Samples

ROAs will generate samples to be uploaded to ODK and shared with surveyors. The do-file for generating these samples should be stored at 02_do/01_sample. Please generate the sample in the manner requested by the data manager or person responsible for coding. The sample should be output as a .csv and .dta to 04_dta/01_sample/01_weekly_sample. Please remember to name your samples consistently each week.

After preparing the sample, ROA should update the surveyor's tracking sheets with their assigned subjects. Please only include the unique ID of the respondent and no PII. The template for the tracking sheet can be found here. Tracking sheets should be output to 04_dta/01_sample/02_tracking_sheet. If needed, the tracking sheets can be uploaded to Google Sheets to be shared with surveyors.

When the sample is ready, please alert the data manager or person responsible for uploading the form into ODK Aggregate. Once the form is updated, ask surveyors to get the latest version of the form in ODK Collect, and remove the previous sample csv file from their smartphones.

II. Identifiers dataset

After the ROA generates each week's sample, they should set up a separate do-file under 02_do/0_identifiers that creates an identifiers dataset to be stored under 04_dta/02_identifiers. This do-file will append all the weekly samples, then remove all other information other than unique ID and mobile (or other key variables needed to merge the survey data with activation or profiling datasets). Please remember to keep this dataset up to date.

III. Raw survey data

Note: A more detailed guide to working with ODK data can be found here.

Raw survey data will be stored encrypted in PAD's private ODK Aggregate server. In order to access project survey data, the project ROA will need to pull, export and decrypt the data using the ODK Briefcase app. The data manager will provide, to each, ROA, a secure username and password via LastPass to access PAD's ODK Aggregate server. The data manager will also provide the ROA with their project's private encryption key via Mattermost. **Please save the encryption key in your *local* drive and do not share it with anyone.**

First, ROAs should pull the encrypted data from PAD's Aggregate server using ODK Briefcase and store it inside the ODK Briefcase Storage folder in their local hard disks. Next, they should export and save the decrypted survey data with PII, in csv format, in the designated folder (03_raw/01_survey_PII). Then, ROAs should run the import do-file generated with *odkmeta* to transform the data set from csv into Stata dta format.

After this, ROAs should run the cleaning do-file which strips the PII from the raw survey data. This is the dataset that should be used for all other purposes.

IV. Interview Recordings

Survey recordings should be moved at the end of each day from surveyors' phones to Dropbox (03_raw/02_survey_recordings_PII) and saved as the respondent's unique ID. Survey recordings should then be deleted from surveyors' phones. Once back checks have been conducted (see Section IV), ROAs must ensure that survey recordings are deleted immediately from Dropbox.

In the situation that a surveyor cannot access Dropbox or does not have sufficient bandwidth to share recordings, they should:

- At the end of the day, the project ROA should download survey data and randomly select 10% of the total surveys
- The ROA should share the selected respondent unique IDs with the back checker and surveyor(s)
- Surveyors should share the recordings for the selected respondent IDs *only* via WhatsApp private message with the back checker
- Surveyors should then delete all recordings off their phones
- Once the back checkers receive the recordings via WhatsApp, they should move them as soon as possible to the assigned folder (03_raw/02_survey_recordings_PII) and delete them off their phone/local storage.

V. Cleaned, de-identified dataset

The cleaning do-file is to be prepared by ROAs and stored in the project folder (02_do/05_cleaning/01_cleaning_survey.do). The cleaning do-file should do the following:

- Generate a non-PII survey dataset. The **non-PII dataset should not include respondents' names and phone numbers. It should include respondents' unique ID.**
- **Do not make any other changes to this dataset.**

VI. Analysis dataset

After generating a "raw" de-identified dataset as per point V above, we will then move on to other cleaning and re-organization of the data in order to prepare the data for analysis. The steps of the cleaning file include:

- Perform any cleaning of text responses
- Consolidate new and old variations of the same variables
- Reshape or organize the data so that there is only one valid attempt per respondent (this should be the last or completed attempt)
- Generate columns for each previous attempt and reason for incompletion
- Save out the do-file for this:02_do/05_cleaning/02_survey_analysis.do
- Save out this analysis dataset: 04_dta/05_analysis/01_survey_analysis.dta.

VI. Back checks

a) Timelines

Please aim to run back checks on 10-20% of a surveyor's completed surveys and 10 - 20% of incomplete surveys. During the first two weeks of implementation, aim to administer back checks within two days so that any errors can be swiftly addressed.

b) Generating back check sample and dataset

Details of the back checks may vary by project and method of data collection. The steps for back checks include:

- At the end of each day, the ROA will generate a random sample of 10% of incomplete surveys and 10% of completed surveys *with recordings* (if recordings are available) using the non-PII survey data. Back checks of both types should be stratified by the enumerator. The do-file will be stored under 02_do/06_back_checks.
- The ROA will then output the back check dataset (which contains cleaned, de-identified survey data for only the back check sample) into another folder (04_dta/04_back_checks/01_data/01_survey).
- The ROA will also move recordings for the selected back checks into another folder (04_dta/04_back_checks/01_data/02_recordings). In some cases, the ROA may need to share the issue log described below with the back checker before moving recordings in case the surveyors were not able to share the recordings on Dropbox yet.
- The ROA will then prepare an issue log (see template here) with the UID of the selected surveys to share with the back checker on Dropbox (or Google Sheets if needed - but this should be updated to Dropbox and deleted as soon as possible). The issue log should be stored under 04_dta/04_back_checks/02_issue_log).
- For incomplete survey backcheck samples, some calls may require the back checker to call the respondent back. For these select respondents, the phone number and name of the respondent should be exported to the issue log as well (please see the example). Extra care should be taken to ensure that the issue log is only stored within Dropbox.

c) Conducting back checks

Once the back checker has the necessary items to begin the back checking process, they can follow these guidelines:

- For completed surveys with recordings:
    - The back-checker will listen to the assigned recording, and ensure that the survey was administered carefully (paying special attention to key variables or difficult questions that include skip patterns).
    - Back checker will update the issue log with details of any issues encountered. They should note the question number, error type and add any additional detail of the error they notice.

- For completed surveys without recordings:
    - The back-checker will call the assigned phone number and get the original respondent on the phone. The back-checker will ask if the survey was administered and a series of key questions which answers would be static between the original survey and the back-check survey. Back-checker will then submit the survey for review.
    - An RA will run the original data collected against the back-check data and check for inconsistencies. Inconsistencies will be logged and addressed by the Field Manager and original enumerator.

- For incomplete surveys:
    - If the reason code states the number was unreachable: try to reach the number, if you get through, *administer the survey if possible*. If unreachable after three attempts, then continue.

    - If the reason code states the farmer refused to give consent for any reason: ask the farmer if they were contacted and understood the consent statement. Ask them why they refused consent and if they asked questions to the surveyor. If after speaking, they agree to the survey, *administer the survey if possible.* If they still are not willing, continue.
    - If the reason code states the farmer was too busy or couldn't speak: ask the farmer if they were contacted more than once by a surveyor to attempt the survey. If they were not, flag this survey/surveyor. Ask the farmer if they would be willing to complete the survey at that time, if yes, *administer the survey if possible.*
    - If the reason code states that the survey is incomplete (in this case there may be a recording available), try to call back the farmer and attempt to complete the survey if possible.
    - Everything should be logged in the issue log.

d) Analyzing and incorporating back checks

On a regular basis or at least once a week, the ROA should review the issue log to identify any patterns in problems. The ROA should use their discretion to raise any larger issues with the RM.

Once a week, the team should also hold a feedback session with surveyors to address any findings from back checks or ask for additional feedback.

VII. Running HFCs

A base code for high frequency checks will be developed and stored in the "templates" folder. RMs and ROAs should work to adapt the code according to your survey adaptations, but try to make as few adaptations to the code as possible. HFCs may include:

- Duplicated completed surveys
- Completed surveys by [day/week] and [surveyor/supervisor]
- Completed surveys by [location]
- Attempts per farmer by surveyor
- Consent rates by surveyor
- Rejection rates
- Survey duration by surveyor
- Form version
- Logic skip checks [TBC] by surveyor
- Doesn't know/prefer not to answer rates by surveyor
- Outliers in input and crop prices and quantities by surveyor
- Specify, other question type

The HFC needs to be run by the ROA on a daily or weekly basis (as discussed by the team). If you notice any unusual skip patterns or patterns by surveyor, please report this immediately to the RM so you can flag this and follow up with the surveyor or flag this to be checked and discussed.

VIII. Data organization

All projects will store their survey data and associated files within the “COVID-19 Survey” folder in PAD’s Dropbox. Each project research team will have access to the templates folder and their own project folder. Aggregated data will be stored in a limited access folder.

Project teams should adhere to the following guidelines when working within project folders.

- 00_admin
    - 00_readme: contains Read-me’s and other project documentation such as codebooks
    - 01_survey_instruments: contains survey document (.docx)
- 01_xlsform
    - 01_forms: contains .xlsx and .xml survey instruments for uploading and any archived versions
    - 02_prefill: contains all weekly samples produced by RAs
- 02_do
    - 01_sample: contains do-file for generating weekly sample and back check sample
    - 02_identifiers: contains do-file for generating sample
    - 03_generate_import_do_file: contains do-files to run *odkmeta* Stata command to generate import do-file.
    - 04_import: contains do-files generated with *odkmeta* Stata command to import data from csv to dta using the metadata of the xlsform
    - 05_cleaning: contains do-file for cleaning survey data
    - 06_back_checks: contains do-file for generating back checking dataset and issue logs
    - 07_hfc: contains HFC do-file
    - 08_analysis: contains do-file for any weekly reporting or analysis
- 03_raw
    - 01_survey_PII
    - 02_survey_recordings_PII
- 04_dta
    - 01_sample: contains any .dta files associated with samples
        - 01_weekly_sample
        - 02_tracking_sheets
    - 02_identifiers: contains identifiers dataset
    - 03_cleaning: contains cleaned survey data
    - 04_back_checks: contains back checking data and tracking sheets
        - 01_data
            - 01_survey
            - 02_recordings
        - 02_issue_log
    - 05_analysis: contains any .dta files produced for analysis

  - temp: contains any temporary or intermediary datasets (non-essential and generated within do-files)
- 05_out
  - 01_log: contains any log files produced
  - 02_sample: contains final sample lists (.dta or .xlsx)
  - 03_analysis: contains any output from analysis including weekly reports, formatted tables or graphs etc.
  - 04_hfc: contains hfc output
- 06_pilot: contains all pilot related material
  - 01_do
    - 01_sample: contains do-file for generating weekly sample and back check sample
    - 02_identifiers: contains do-file for generating sample
    - 03_cleaning: contains do-file for cleaning survey data
  - 02_raw
      - 01_survey_PII
      - 02_survey_recordings_PII
  - 03_dta
    - 01_sample: contains any .dta files associated with samples
      - 01_weekly_sample
      - 02_tracking_sheets
    - 02_identifiers: contains identifiers dataset
    - 03_cleaning: contains cleaned survey data
    - temp: contains any temporary or intermediary datasets (non-essential and generated within do-files)

In general, the folder structure should follow conventional best practices, including: parallel structure in naming (data and output produced by do-files should have the same or similar naming so they can be identified) and archiving any older versions of datasets or documents.

IX. Version tracking

Survey instruments to be edited before, during, and after piloting. Versions to be saved with “_*v#*” and older versions that are no longer in use to be moved to the respective archive folder. Version tracking to be documented in metadata as needed.

Version tracking sheets for all projects can be found in 00_templates/00_admin.

X. Metadata

Maintaining metadata is important for internal documentation as well as for data sharing purposes. For this project, metadata should be generated for the core survey and for project-specific surveys, in the form of a Read-Me .docx. The core Read-Me should contain the following:

- Title

- Date created and date(s) modified
- Author(s)
- List of all project do-files, documents and datasets with storage location/URL and purpose
- Information on partners and funders
- Access information
- Codebook for survey instrument(s), which can be attached in a separate .xlsx or .csv file
    - Variable name, label, any other notes

The Read-Me for projects should contain the following information:

- Title
- Date created and date(s) modified
- Author(s)
- List of all project do-files, documents and datasets with storage location/URL and purpose
    - Codebook only required if any variables specific to the project or generated by the project for analysis
- Information on partners and funders
- Access information
- Language
- Any other relevant notes on methodology

5. Summary of Roles and Responsibilities

Table 4: Project Roles and Responsibilities

Role	**Responsibilities**	**Access**
Data Manager	● Code survey instruments ● Upload instruments to ODK ● Upload sample to ODK ● Generate and share encryption keys for uploading and downloading data ● Generate HFC template ● Maintain data security and management processes	*All data*
Project Research Managers	● Adapt survey instruments to local contexts ● Oversee survey implementation within and across teams ● Review data quality checks (back-check logs and HFCs) ● Ensure all research activities have obtained the required IRB and government approvals ● Quality-ensure internal reporting	*All project specific data*

Project Research Associates	● Produce weekly samples ● Clean and de-identify data regularly ● Ensure raw data with PII and cleaned data without PII are stored separately ● Generate a sub-sample of survey respondents for back-checking and review back-check logs ● Adapt and run HFCs weekly ● Produce internal reporting	*All project specific data*
Field supervisors/ survey supervisors	● Provide training to surveyors ● Pilot survey instrument and report issues ● Conduct/ coordinate regular back checks ● Relay feedback from back checks/ HFCs to surveyors	Project-specific: Back check survey data Back check recordings

6. Data security and access

Table 5: Datasets containing personally identifying information (PII) and storage, access and protection protocols

Data	**PII contained**	**Storage**	**Access**	**Protection**
Identifiers datasets	Mobile number	Stored in project folders	Limited to programme RMs, ROAs and core team	Limited access
Raw survey dataset	Mobile number, name	Stored in project folders	Limited to programme RMs, ROAs and core team	Can only be downloaded using encryption key only available to ROAs; stripped of PII immediately and only non-PII data used for other purposes
Interview recordings	Voice	Stored on surveyor mobile phones, stored in project folders, stored on supervisor phones (possible)	Project teams, core team	Recordings not selected for back checks deleted from phones immediately, only selected recordings stored on Dropbox and shared with back

				checker, all recordings deleted once back checks complete